Universidad
Rey Juan Carlos

Escuela Técnica Superior
de Ingeniería de Telecomunicación

Ingeniería de Telecomunicación y
Licenciatura en Admón. y Dirección de
Empresas

Proyecto Fin de Carrera

SOCIAL MEDIA EN EL MERCADO FARMACÉUTICO: SITUACIÓN ACTUAL, ESTRATEGIA Y DISEÑO E IMPLEMENTACIÓN DE PLATAFORMA ANALÍTICA DE APOYO

Autor: FRANCESC LUCENA JUVE
Tutor: GREGORIO ROBLES MARTINEZ
Cotutor: DAVID SIERRA MORO

Curso Académico 2012/2013

Proyecto Fin de Carrera

SOCIAL MEDIA EN EL MERCADO FARMACÉUTICO: SITUACIÓN ACTUAL, ESTRATEGIA Y DISEÑO E IMPLEMENTACIÓN DE PLATAFORMA ANALÍTICA DE APOYO

Autor

Francesc Lucena Juvé

Tutor

Gregorio Robles Martínez

Cotutor

David Sierra Moro

La defensa del presente Proyecto Fin de Carrera se realizó el día de junio de 2013, siendo evaluada por el siguiente tribunal:

Presidente:

Vocal:

Secretario:

y habiendo obtenido la siguiente Calificación:

Fuenlabrada, a de de 2013

A todos aquéllos que han hecho que esto sea posible

Agradecimientos

El presente proyecto cierra el ciclo más largo que he vivido hasta el momento: el de estudiante universitario. Para llegar hasta aquí ha sido necesario contar con la colaboración y apoyo de muchas personas. Algunas siguen muy presentes en mi día a día mientras otras parecen solo recuerdos, pero todas ellas imprescindibles para llegar a ser quién soy.

Quiero empezar por mi familia: mis padres, mi hermano, mis abuelos, mis tíos... todos ellos han estado siempre a mi lado. A veces con palabras de ánimo, otras veces con críticas y muchas otras simplemente sin decir nada me han dado la seguridad para seguir hacía adelante. Para sentir que esto era posible y para crecer hasta convertirme en una persona que puede moverse de manera independiente por éste mundo tan complejo en el que vivimos.

Debo dedicar unas líneas también a la persona que más tiempo pasa a mi lado últimamente. Puedo decir del cierto que sin ella, el fin de éste ciclo se habría alargado aún más. Con paciencia, esfuerzo y mucho corazón se pueden construir grandes cosas. Tengo que dar las gracias por compartir conmigo esas, y muchas otras, grandes cualidades.

No puedo olvidarme tampoco de un pequeño grupo de amigos muy particular, el de aquellos que alguna vez han vivido conmigo. Dicen que compartir piso no es algo sencillo, yo no podría estar más en desacuerdo. Gracias a todas aquellas personas con las que he convivido he sido feliz en mi día a día. Quiero mencionar también algunos amigos que se han convertido en parte de mi familia y que en todo momento me han apoyado para llegar hasta el final.

Finalmente quiero dar las gracias a mis tutores, que de forma desinteresada han dedicado tiempo y esfuerzo a este proyecto, aceptando en muchas ocasiones mis cambios de idea, siempre ayudándome con acertadas recomendaciones, y haciéndome sentir en todo momento que tenía quien desde el conocimiento me ofrecía soporte para convertir éste proyecto en un éxito. En este grupo me gustaría incluir también otros profesores que me he ido encontrando en la universidad, así cómo algunos profesionales que he conocido en mi lugar de trabajo.

Contenido

Resumen

El estallido de los medios sociales ha tenido un fuerte efecto sobre la sociedad. En el ámbito empresarial los medios sociales suponen una valiosa fuente de información a la vez que un nuevo canal para llegar a ciertos colectivos objetivo.

El mercado farmacéutico, por sus particularidades, requiere afrontar este reto desde una perspectiva especializada, hecho que motiva este proyecto.

A lo largo del presente proyecto se describe el mercado farmacéutico y el modo en el que se está viendo afectado por la irrupción de los medios sociales. Teniendo en cuenta lo anterior, se presentarán algunos principios que las compañías farmacéuticas deberían de tener en mente a la hora de diseñar su estrategia en medios sociales.

Presentaremos la importancia de disponer de una herramienta que permita monitorizar lo que sucede en las redes sociales introduciendo a su vez los principales conceptos de Business Intelligence.

Después de haber leído este proyecto el lector comprenderá la importancia de disponer analíticas que transmitan información relevante y que permitan a las personas encargadas de la toma de decisiones entender qué está sucediendo y cómo pueden actuar para que sus acciones tengan efecto.

Abreviaciones

BI	Business Intelligence
DW	DataWarehouse
FAQ	Frequently Asked Questions
IDE	Integrated Development Environment (Entorno de desarrollo integrado)
MVC	Modelo Vista Controlador
OTC	Over The Counter
PFC	Proyecto Fin de Carrera
PK	Primary Key
SaaS	Software as a Service

Capítulo 1

Introducción

El proyecto fin de carrera marca el final de una etapa y por lo tanto, el principio de otra nueva.

Algunos compañeros han elegido un proyecto ofrecido por un profesor, otros han propuesto el proyecto ellos mismos. Algunos están ahora mismo tratando de elegir o encontrar un tema sobre el que dedicar una gran cantidad de tiempo y esfuerzo. En cualquier caso, es una decisión difícil, que en mi caso, y por el modo en el que han transcurrido los últimos meses de mi vida, se ha complicado, y sobre todo alargado mucho más de lo que cabría esperar.

Finalmente, la decisión ha sido realizar una herramienta de "business intelligence" ad-hoc para el análisis y explotación de información pública recogida de forma automática en internet.

¿Por qué después de varios cambios, se acaba eligiendo una herramienta como ésta para el proyecto?¿Por qué una herramienta ad-hoc, cuando existe una gran variedad de soluciones de BI customizables? Éstas y otras preguntas son las que espero que queden resueltas después de conocer en profundidad este proyecto.

1.1. Motivación

En marzo de 2012, unos meses antes de la fecha esperada para terminar la carrera, busqué la oportunidad de combinar ese final con el principio de mi carrera profesional. Tomada la decisión, era el momento de pensar en qué dirección quería enfocar dicha carrera, (ya

sabiendo que para iniciar de nuevo una aventura emprendedora sería necesario adquirir algo de experiencia en un entorno empresarial auténtico). Siempre me ha gustado hacer un poco de todo, así que algo me decía que consultoría podía ser lo mío.

Me surgió la oportunidad de empezar una beca con analista en IMS Helth, importante multinacional que presta servicios principalmente de marketing a empresas del mercado farmacéutico.

Nunca me había planteado la posibilidad de trabajar en el mercado farmacéutico. Al fin y al cabo soy -casi- un ingeniero de telecomunicación y licenciado en dirección y administración de empresas.

Al conocer mejor la empresa, qué hacían y cómo lo hacían, me pareció que merecía la pena intentarlo. Y en ese entorno se ha gestado este proyecto.

1.2. Contextualización y conceptos básicos

El presente PFC trabaja principalmente en el marco de tres importantes conceptos: Social Media, Business Intelligence y la Industria Farmacéutica. Podemos considerar que el primero es la fuente de información que vamos a analizar. Utilizando técnicas de "Business Intelligence" vamos a analizar la información obtenida de los medios sociales. Finalmente el mercado farmacéutico es el sector o industria en el que todo el trabajo debe ser contextualizado.

Aunque los tres conceptos son en general muy conocidos, es imprescindible empezar con una definición de los mismos para poder entender mejor a que nos vamos a enfrentar a lo largo del trabajo:

1.2.1. Social Media

Según Kaplan y Haenlein [1], los medios sociales se definen como la unión entre la Web 2.0 y el contenido generado por el usuario. El término "web 2.0" fue utilizado por primera vez en 2004 para describir la nueva forma en la que los desarrolladores web y los usuarios finales empezaron a utilizar internet. Los contenidos y aplicaciones han pasado a ser publicados y constantemente modificados por el conjunto de usuarios y no por un único individuo.

Se hizo eco de ello en 2006 un estudio de Forrester Research llamado "Social Com-

puting: How Networks Erode Institutional Power and What to Do About It" [2], que identificaba esa nueva tendencia en la red por la cual la gente empezaba a interconectarse de diferentes maneras. Esa nueva forma de interacción suponía una amenaza para las empresas. Tradicionalmente, las empresas e instituciones se han construido sobre el control y los medios sociales lo quebrantan y debilitan.

Pero antes de entrar más en detalle en lo que "social media" supone para las corporaciones es importante acotar y entender la definición de dicho concepto. Para ello se estudiarán por separado los dos términos introducidos en la definición de Kaplan y Haenlein:

- Web 2.0: con el estallido de la burbuja de las punto-com muchos consideraron que la web estaba sobreestimada [3]. En una sesión de "brainstorming" entre Dale Dougherty y Tim O'Reilly creyeron que al contrario de esas creencias, la web tomaba un papel más relevante que nunca, con una gran cantidad de nuevos sitios apareciendo constantemente. Aún más, se dieron cuenta de que las compañías que habían sobrevivido al colapso tenían mucho en común. La conclusión fue que el estallido de la burbuja podía suponer un punto de inflexión para la web. Así acuñaron la denominación de web 2.0, que fue adoptada rápidamente.

  Para acotar el concepto de web 2.0 utilizamos la definición concisa dada por el mismo O'Reilly [4]: Web 2.0 es la red como plataforma, incluyendo cualquier dispositivo conectado a internet; las aplicaciones Web 2.0 son aquéllas que sacan el máximo partido de las ventajas intrínsecas de esa plataforma: entregar software como un servicio permanentemente actualizado que mejora cuanto mayor sea el número de gente que lo utiliza. Consume e integra información de múltiples fuentes, incluyendo usuarios individuales, permitiendo a la vez, que la información de los mismos pueda ser integrada y entregada a otros usuarios. Así se crean efectos de red a través de una "arquitectura de participación".

  Buscando una definición más reciente tenemos que es necesario considerar tanto aspectos de usabilidad como ciertos paradigmas tecnológicos, estrategias de negocio y tendencias sociales. La web 2.0 es más dinámica e interactiva que su predecesora, permitiendo a los usuarios tanto acceder al contenido de la web, como contribuir al mismo. También permite a los desarrolladores crear rápida y fácilmente nuevas aplicaciones web que utilicen datos, información o servicios disponibles en internet.

- Contenido generado por el usuario - UGC por sus siglas en inglés: contenido creado

por usuarios de internet que voluntariamente contribuyen con diferentes tipos de información que puede ser multimedia, y que normalmente se presenta de forma que sea útil a terceros usuarios. El uso de este tipo de contenido ha crecido rápidamente recientemente.

Así es sencillo clasificar algunos de los medios sociales mas utilizados según el modo en el que interactúan los usuarios [5] :

- Usuarios creando contenidos: blogs y podcasts. Espacios en los que los usuarios suben diferentes tipos de contenido como noticias, historias, canciones o vídeos. Información cuyo objetivo es, en general, el de ser compartido con otros internautas.

- Interconexión de usuarios: redes sociales y mundos virtuales. Los miembros de estos sitios mantienen un perfil y se conectan los unos con los otros. Facebook, Twitter y LinkedIn son los medios sociales de este tipo más populares, aunque existen una gran cantidad de ellas en muchos casos especializadas (veremos más adelante, por ejemplo, comunidades de pacientes).

- Colaboración entre usuarios: Wikis y Open Source. A diferencia de los casos anteriores, los usuarios de estos medios se organizan de forma explícita creando potentes herramientas.

- Usuarios reactivos: Foros y evaluaciones. Podríamos decir que los foros de discusión representan conversaciones a "cámara lenta", en las que un usuario lanza una pregunta, con la esperanza de que otros usuarios puedan ayudar a la resolución de la misma.

- Usuarios que organizan contenido: etiquetas "tags". La interacción a través del etiquetado es sutil. Los "tags" pueden utilizarse para organizar el contenido propio ignorando la componente social, pero supongamos que se etiqueta un artículo para denotar que es interesante y alguien llega al mismo gracias a esa etiqueta. Es posible que esa persona considere que sus intereses son similares a los de la persona que ha etiquetado la noticia y que decida navegar por las otras noticias etiquetadas por el mismo usuario.

- Medios para acelerar el acceso al contenido: RSS y Widgets. Son herramientas que llevan el contenido directamente al usuario. A pesar de que el RSS no es

una tecnología muy utilizada de forma directa, existen portales como iGoogle o MyYahoo, que permiten al usuario personalizar el contenido mostrado, y dicho contenido se obtiene a través de RSS.

1.2.2. Business intelligence

Se denomina "Business Intelligence" o BI al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos disponibles en una organización o empresa[6].

El concepto de BI es en realidad muy amplio. En esta breve aproximación nos centraremos en dos aspectos muy importantes a lo largo del presente proyecto. Por una lado se analizará el proceso que sigue una herramienta de BI desde la recogida de datos hasta la distribución de informes, y por el otro las etapas que hay que seguir para llevar a cabo un correcto desarrollo de una solución de BI. Los aspectos más técnicos de esta ciencia se analizarán en el capítulo 3.

Podemos referirnos a BI como el conjunto formado por el almacén de datos -al que nos referiremos por su denominación en inglés "data warehouse"-, la minería de datos -o data mining-, analíticas de datos, generación de informes y tecnologías de visualización [7]. Las arquitecturas de BI actuales están orientadas a la toma de decisiones estratégicas y/o operativas. Una herramienta de BI da soporte a todo un proceso realizado sobre datos de negocio. Este proceso consta de las cuatro etapas representadas en la figura 1.1:

1. Recogida de datos: una plataforma de BI se nutre de una gran cantidad de datos. La información con la que la herramienta trabajará es adquirida tanto de fuentes internas -sistemas operacionales, ERP, CRM, etc.-, como de fuentes externas -datos adquiridos a terceras empresas, estudios de mercado, etc.

2. Integración de datos: consiste en combinar datos que provienen de diversas fuentes proveyendo a los usuarios una visión unificada de dichos datos [8]. La complejidad de la integración vendrá dada por la complejidad de las fuentes de información, y por las características de los análisis que se quieran llevar a cabo posteriormente.

3. Análisis de datos: el análisis de datos es un proceso que incluye la inspección, filtrado, transformación y modelado de datos con el objetivo de destacar la información relevante y dar soporte a la toma de decisiones. Se ha citado con anterioridad el concepto de "data mining", muy utilizado en Business Intelligence; es una técnica

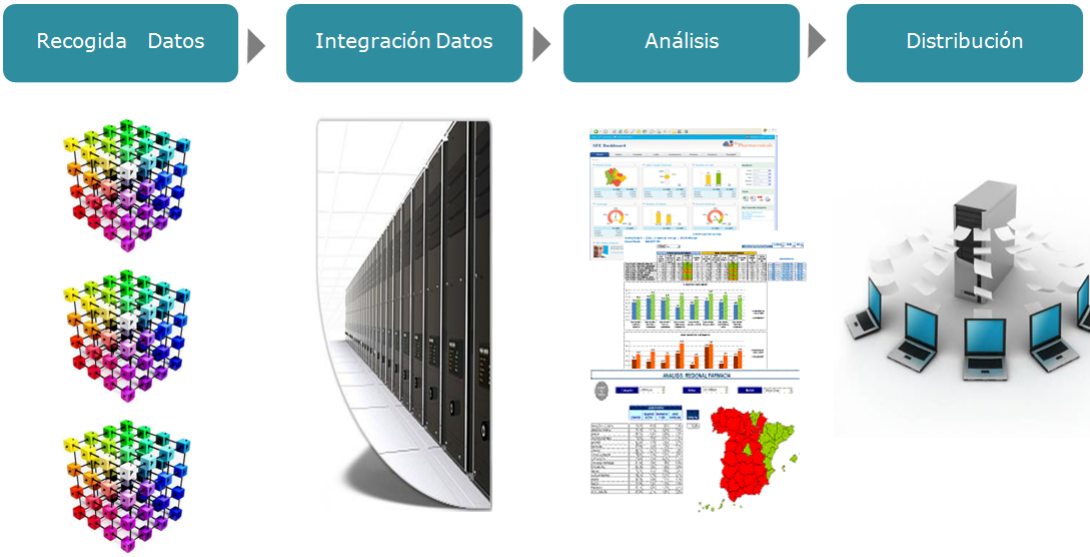


Figura 1.1. Business Intelligence. Proceso completo en una solución de BI. Fuente: IMS Health.

de análisis que consiste en el la búsqueda de modelos que permitan realizar análisis predictivos. El análisis de datos es la antesala a la representación de los mismos. Existe una gran cantidad de técnicas y herramientas de análisis y "reporting".

4. Distribución: Finalmente es necesario hacer accesible toda esa información a los encargados de la toma de decisiones -habitualmente referidos como *decision makers*-. La distribución de los datos depende también de muchos factores, como los requisitos de confidencialidad, los sistemas de información disponibles, etc.

1.2.3. Industria farmacéutica

La industria farmacéutica es el sector para el que se está desarrollando el proyecto. Forman parte de este mercado aquellas compañías dedicadas a la fabricación, preparación y comercialización de productos químicos medicinales para el tratamiento y también la prevención de las enfermedades.

El sector al que pertenecen los futuros clientes de la solución es en este caso especialmente relevante por las particularidades del mismo. A pesar de que los consumidores de los productos que las empresas farmacéuticas comercializan son los ciudadanos, éstos no son los que eligen o pagan [directamente] el producto. Este hecho condiciona en gran parte el sistema comercial adoptado en el mercado.

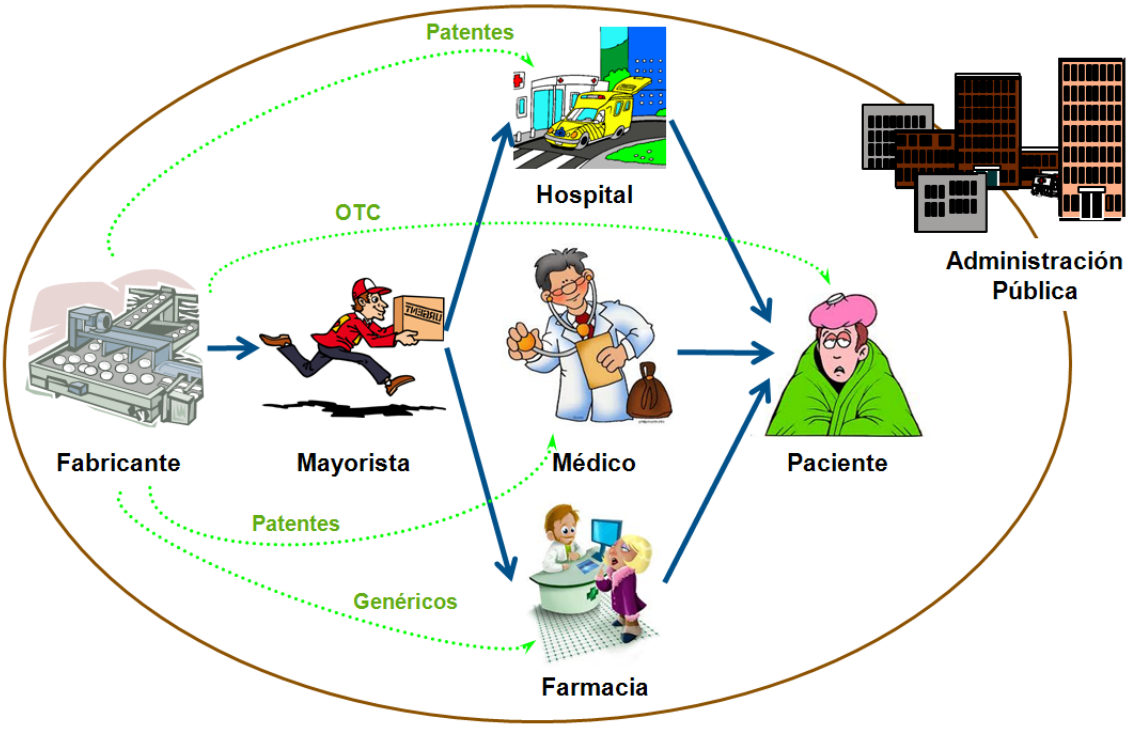


Figura 1.2. El mercado farmacéutico. Principales participantes y como interactúan. Fuente: IMS Health. Elaboración propia.

El mercado farmacéutico está altamente regulado, lo que hace que haya más participantes involucrados en el mismo que en otros mercados. Imaginemos, a modo de ejemplo, el mercado de los cereales. Un productor de cereales los puede lanzar al mercado sin tener que iniciar costosos trámites burocráticos que pueden alargarse durante meses, los puede anunciar libremente y comercializar en cualquier establecimiento, restaurante, etc. Puede llevar a cabo agresivas campañas de publicidad para que el consumidor potencial decida acercarse a su supermercado a adquirir dichos cereales. En ese caso el consumidor comprará los cereales en caso de considerar que su precio adecuado o terminaría eligiendo un producto sustitutivo[1].

El ejemplo de los cereales es poco es aplicable al mercado de los medicamentos. La aprobación de un medicamento lleva años y debe pasar por estrictos controles, demostrar su eficacia en ensayos clínicos y es necesario negociar el precio con las Administraciones Públicas. Una vez lanzado el producto, salvo en un conjunto de medicamentos -llamados

[1]En economía un bien se considera un bien sustitutivo o competitivo de otro, en tanto uno de ellos puede ser consumido o usado en lugar del otro en alguno de sus posibles usos.

Over The Counter (OTC)[2]- no es posible publicitarlo directamente al consumidor, al que llamaremos paciente. Además, el poder de decisión del paciente es muy limitado, siendo el médico el que prescribe, y por lo tanto decide, qué productos se van a consumir. Por otro lado, reformas legales recientes obligan ahora a los doctores a recetar por principio activo, y a la farmacia a suministrar el medicamento de menor precio dentro de su agrupación homogénea -medicamentos iguales, pero de diferente fabricante- y en caso de igualdad de precios, la farmacia estará obligada a suministrar el medicamento genérico [Artículo 4 Real Decreto-Ley 16/2012].

Así, como se muestra en la figura 1.2, el conjunto de "stakeholders" del mercado farmacéutico es muy amplio y el modo en el que interrelacionan es complejo. En el ámbito de este proyecto son relevantes:

- la fuerza de ventas de un laboratorio farmacéutico está formada principalmente por visitadores médicos, tambiñen llamados delegados. Los vistadores médicos son los encargados de mantener un contacto constante y cercano con los médicos informando a los mismos las últimas novedades del laboratorio farmacéutico a la que representen.

- A pesar de que se está reduciendo su poder de decisión, el médico sigue siendo el que mayor poder tiene a la hora de decidir que tratamiento sigue un paciente, y en consecuencia, qué medicamentos consume.

- Desde que se ha impuesto la obligación de recetar por principio activo para aquellos medicamentos a los que ha expirado la patente y el genérico está disponible, es la farmacia la que puede elegir de entre los diferentes fabricantes de genéricos, cuál suministran.

- Los pacientes son el última instancia los que consumirán los medicamentos y es básico conocer como perciben el tratamiento, si lo siguen hasta el final, etc. Además, en determinadas ocasiones pueden requerir al farmacéutico un medicamento de marca en lugar del genérico, aunque ello implique que acaben pagando directamente ellos el producto en el momento de la adquisición. Además, para el conjunto

[2]El medicamento de venta libre, también llamado Over The Counter (OTC, por sus siglas en inglés) o medicamento de venta directa o medicamento sin prescripción es aquel que no requiere una prescripción o receta médica para su adquisición. Se trata de una categoría de medicamentos producidos, distribuidos y vendidos a los consumidores/usuarios para que los utilicen por su propia iniciativa.

de medicamentos que pueden publicitarse -OTC- es el cliente el que elige que producto adquiere (p.e.: Frenadol o Bislogrip)

- Son sin embargo las aseguradoras, públicas o privadas, las que soportan la mayor parte del coste de los medicamentos. Eso las convierte en otro participante clave.

1.3. Objetivos

Este proyecto busca entender cómo afecta la irrupción de los medios sociales a la industria farmacéutica, y a través de este entendimiento, diseñar una herramienta analítica que permita extraer información relevante de los mensajes recogidos automáticamente en diferentes medios sociales.

Tras la realización del proyecto se espera encontrar hechos que faciliten la definición de una estrategia en medios sociales a través del conocimiento del sector, así como permitir su ejecución mediante el diseño y desarrollo de una herramienta de análisis que nos ayude a entender cómo estamos posicionados en internet.

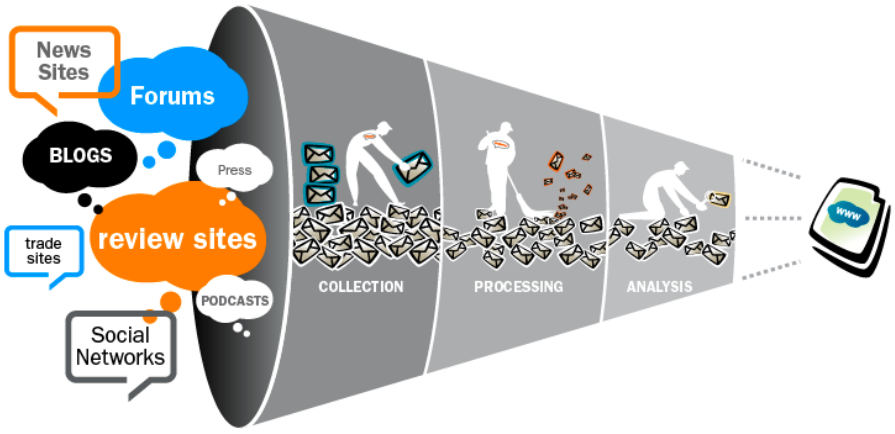


Figura 1.3. Infografía con una visión general del objetivo del proyecto: obtener conocimiento a través de la recogida, procesado y análisis de lo que se publica en medios sociales. Fuente: http://socialmediamagic.com.

1.3.1. Una Visión Estratégica del BI en Medios Sociales

El plan de marketing forma parte de la estrategia de una compañía, y cada vez más los planes de marketing incluyen una sección dedicada a la estrategia de la compañía en redes sociales.

Este proyecto consiste en el diseño de una herramienta analítica para entender mejor lo que se está diciendo en la red acerca de un producto farmacéutico, o dentro de un mercado del sector. Como destaca John Souza en una entrada del 30 de agosto de 2012 del blog de "Social Media Marketing University" [9], la irrupción de los medios sociales ha abierto todo un nuevo mundo de herramientas de Business Intelligence para el análisis de ese tipo de información. Entonces, ¿Por qué este proyecto?¿Dónde esta la innovación?

Este proyecto trata de entender cómo deben enfocar las compañías farmacéuticas su estrategia en los medios sociales, y en función de eso, diseñar una herramienta que aporte información relevante a sus necesidades.

Por otro lado, gran parte de las herramientas de recogida y análisis de datos están diseñados para funcionar en inglés, cuando en el mercado español, es fundamental poder trabajar con mensajes en español.

Con la cantidad de nuevas herramientas que están apareciendo, el usuario de la información puede caer en la tentación de afrontar el posicionamiento en los medios sociales como algo táctico; como un proceso a través del que entendiendo lo que se dice en internet acerca de tu marca o producto se tomen decisiones. El objetivo último del proyecto es no sólo diseñar una herramienta, sino entender cómo ésta pude ayudar a elaborar una estrategia de posicionamiento en redes sociales que se traduzca, en última instancia, en un incremento de beneficio a medio y largo plazo.

Según Jay Baer en el artículo llamado "Social Media Strategy in 8 steps" [10], los 8 pasos para desarrollar una estrategia en medios sociales son:

1. Construye un "arca": los medios sociales influyen al completo a la organización, por lo que el equipo encargado de definir la estrategia en medios sociales debe ser multidisciplinar y estar integrado por componentes de diferentes departamentos. En la industria farmacéutica, por su complejidad organizativa, con más razón. Por ejemplo, podría ser interesante involucrar a gente de ventas, I+D, encargados de relaciones institucionales, etc.

2. Escucha y compara: es importante escuchar qué dicen en los medios sociales tus clientes, competidores, etc. Eso ayudará a entender por dónde empezar a interactuar. En este punto una plataforma analítica es fundamental, pero de nuevo, en el mercado farmacéutico existen complejidades añadidas. Como se ha contado, los clientes de las farmacéuticas no son, en general, los pacientes, ni los médicos, ni las

farmacias, ni las instituciones, sino que cada uno de esos colectivos juega su papel y sólo estando bien posicionado en todos ellos es posible que un producto tenga éxito. Además existen muchas limitaciones legales, tanto acerca de la publicidad que las compañías pueden realizar, como a la hora de dar acceso a datos de pacientes a éstas. Es por esa razón que este paso es especialmente complicado en el mercado farmacéutico.

3. ¿Qué quiero conseguir? A través de los medios sociales, se pueden conseguir muchos objetivos diferentes, pero a la hora de construir la estrategia es necesario fijar unas metas. Entender qué queremos conseguir nos ayudará a definir qué es lo que queremos hacer.

4. Elige métricas de éxito: ¿Cómo se va a valorar si la estrategia tiene éxito? Cuál es el retorno de la inversión cuando se invierte en poner en marcha una estrategia en medios sociales.

5. Analiza tu audiencia: con quién se interactuará en los medios sociales. ¿Cuáles son sus características demográficas y sociales? ¿Cuál es el impacto que se puede producir sobre los interlocutores? En línea con lo que se ha expuesto en el segundo punto, en farma es especialmente importante entender quién es el interlocutor, evitar no salir de las barreras legales establecidas, y sobretodo, entender cómo desde ahí, se puede influir a favor de una marca.

6. ¿Qué te hace especial? No importa quién eres, o qué vendes, el producto por sí solo no generará "simpatía". ¿Cómo lo vamos a hacer para llegar a las emociones de los interlocutores? En nuestro contexto, y para clarificar, podemos plantear preguntas como: ¿Un laboratorio que es más transparente en redes sociales despierta más simpatía entre los consumidores? ¿Podría variar eso los hábitos de consumo de los mismos?

7. ¿Cómo serás más humano? Los medios sociales se basan en las personas, no en productos o marcas. Una compañía se tiene que comportar, por lo menos hasta cierto punto, cómo una persona, y no solo como una entidad.

8. Crear un plan específico para cada canal: una vez se conozca como interactúan las personas dentro de los diferentes canales, es posible elaborar un plan de acción específico para cada uno de ellos. Por ejemplo, un laboratorio no se comportará igual en Twitter, donde debería de limitarse a dar información objetiva y genera-

lista, que en una comunidad de médicos, donde la información puede ser mucho más dedicada, específica y científica.

Si se percibe la necesidad de abordar la presencia en los medios sociales como un factor de éxito estratégico, se comprenderá el alcance de este proyecto y la capacidad de innovación a través de una posición privilegiada en el conocimiento del mercado farmacéutico español, así como de sus necesidades en materia de estrategia social.

1.4. Estructura de la memoria

La figura 1.4 representa el proceso a seguir para mantener siempre actualizada la estrategia de Business Intelligence Social. En este proyecto se desarrollarán cada uno de los bloques que aparecen en la figura.

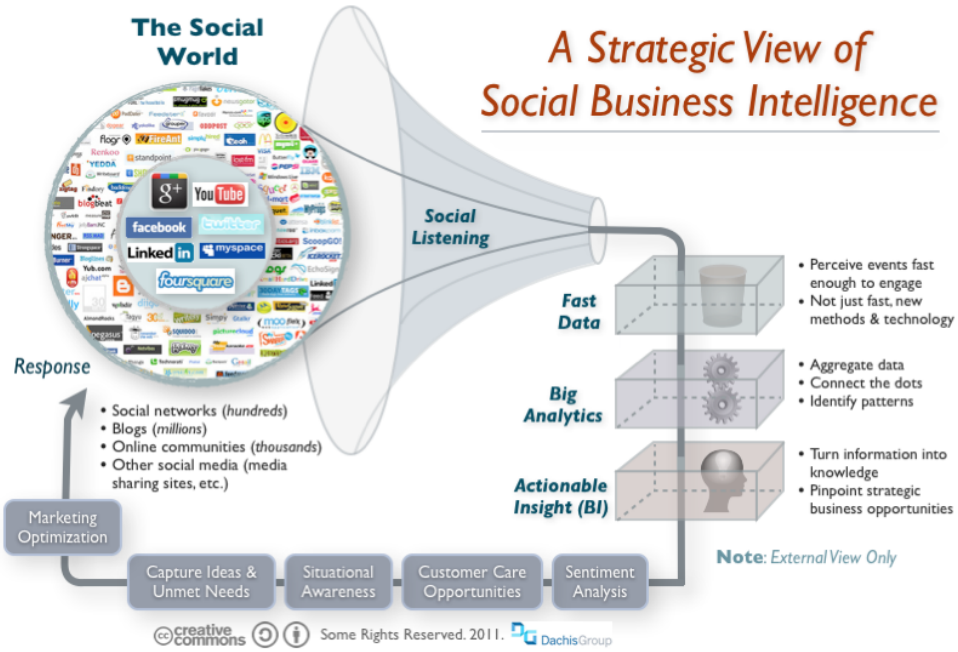


Figura 1.4. Modelo recogida y análisis de información para elaboración de la estrategia de marketing online. Fuente: http://socialmediamarketinguniversity.com/gain-business-intelligence-social-media/.

En las páginas que siguen a esta introducción, empezaremos por entender cómo están afectando los medios sociales a la industria farmacéutica, lo que nos aportará el marco necesario para poder entender qué sucede y cómo afrontarlo.

A continuación se analizará desde un punto de vista tecnológico el conjunto de tecnologías que, de un modo u otro, entran en juego en este proyecto. Esto nos servirá para fijar el marco en el que se ejecutará en desarrollo de la herramienta analítica.

Veremos que en nuestro caso las dos mayores complejidades con las que nos encontraremos son el gran volumen de datos con los que tendremos que trabajar, y el desarrollo de analíticas potentes para convertir esa gran cantidad de datos en información fácil de entender. El capítulo 4 incluye la definición de la lógica de negocio de la plataforma de análisis, afrontando así la segunda problemática.

A continuación se presenta la arquitectura de la herramienta, en la que se ha tenido en cuenta tanto el volumen de información, como las analíticas definidas. Ésta es posiblemente la sección en la que el detalle técnico es mayor.

Finalmente se expondrán las conclusiones y líneas de desarrollo futuro en el último capítulo.

Capítulo 2

Impacto de los medios sociales en la industria farmacéutica

Desde los años 90 la capacidad de influencia del paciente en el mercado farmacéutico ha sido sometida a estudio. No quedaba claro hasta que punto era positivo permitir al paciente jugar un papel activo en la gestión de sus tratamientos sanitarios -llamado Empoderamiento del Paciente, o *Patient Empowerment*- Un exceso de capacidad de decisión podía acabar confundiendo al paciente en lugar mejorar su satisfacción con el tratamiento; lo que puede justificar, en parte, por qué se ha mantenido muy reducido hasta hace poco. Sin embargo, con la aparición de los medios sociales el empoderamiento del paciente se ha convertido en una realidad [11].

2.1. Salud 2.0 y mHealth

El término "Salud 2.0" hace referencia a la transición hacia la gestión sanitaria participativa. La Salud 2.0 permite al ciudadano convertirse en un socio activo y responsable de su propia salud y abre nuevas posibilidades a diferentes vías de atención con el médico.

Un análisis de la publicación "Preventing Chronic Disease" [12] explica que los pacientes con enfermedades crónicas son los que están tomando un papel más activo en su gestión sanitaria a través de nuevas tecnologías. La publicación indica que dichos pacientes buscan en internet información para mejorar sus condiciones.

Aproximadamente el 75 % de todos los pacientes con enfermedades crónicas y que participan en la salud 2.0 afirman que información sanitaria buscada en internet ha

contribuido a la hora de decidir acerca del tratamiento a seguir. Un 69 % dice que la información consultada les hizo plantearse nuevas preguntas que trasladaron a su doctor.

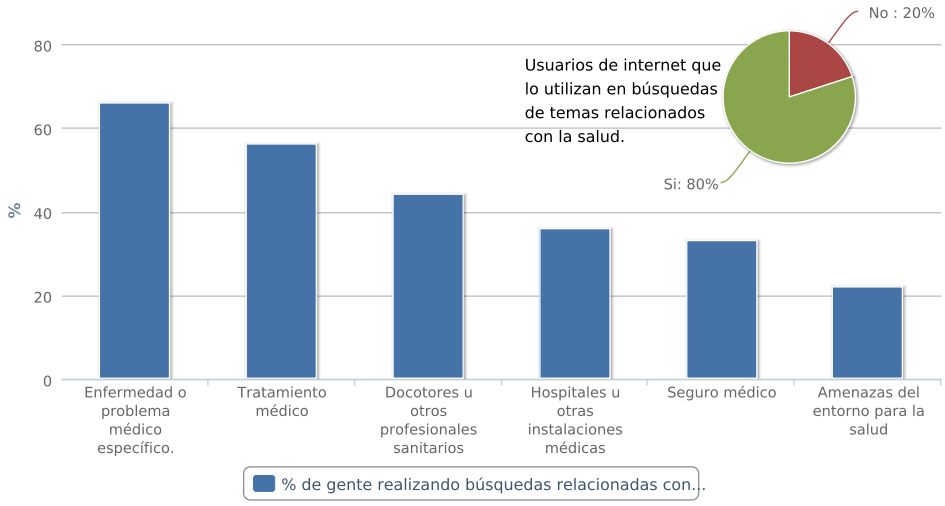


Figura 2.1. Uso de internet para realizar consultas relacionadas con la salud. Fuente: California HealthCare Foundation. Elaboración propia.

De un modo más general, la "California HealthCare Foundation" [13] publicó en febrero de 2011 los resultados de una encuesta que realizan periódicamente en Estados Unidos, de la que se desprende que el 80 % de los americanos que utilizan internet (59 % si se utiliza como referencia la población total de Estados Unidos), lo utilizan de forma más o menos regular para realizar consultas relacionadas con su salud [ver figura 2.1], situándose así como tercera actividad más popular dentro de las monitorizadas por la fundación. El porcentaje aumenta en 8 puntos cuando la muestra se centra en personas que han estado a cargo de forma no remunerada de algún enfermo.

Las consultas más frecuentes en internet son acerca de una enfermedad o problema médico concreto (66 % de los casos), principalmente tratando de encontrar consejos que ayuden a mejorar las condiciones de vida del enfermo. El 56 % de los internautas han realizado alguna búsqueda relacionada con algún tratamiento médico -o medicamento- concreto, dentro de esta categoría destacan los búsquedas por analgésicos, antidepresivos, medicación para la tensión, corticoides, diabetes, etc. También es muy frecuente -un 44 % de los casos- buscar en internet a médicos y otros profesionales sanitarios. En internet se realizan también consultas acerca de hospitales, aseguradoras médicas, etc.

Cabe mencionar también el término mHealth, o "salud móvil", que hace referencia al

acceso sanitario a través de dispositivos móviles. Se considera un segmento del eHealth. Las aplicaciones de mHealth incluyen el uso de dispositivos móviles para recoger información clínica de pacientes, enviar datos médicos a los doctores, investigadores y pacientes, monitorizar en tiempo real los signos vitales de los pacientes, etc. [14]. La "mHealth" ofrece por un lado una mayor cantidad de información de pacientes en la red, pero sobre todo, integra la salud con tecnologías de uso diario cómo el teléfono móvil. Así, el informe de la "California HealthCare Foundation" [13] indica que el 17 % de los usuarios de teléfonos móviles lo han utilizado alguna vez para realizar consultas relacionadas con la salud.

2.2. Impacto de los medios sociales en la industria farmacéutica

Cada vez es mayor el número de pacientes y familiares que interactúan online e intercambian sus experiencias con enfermedades y tratamientos. La gran cantidad de información acerca de la salud disponible en internet, supone una valiosa fuente de información para las compañías farmacéuticas. El análisis de esa información permite identificar fortalezas y debilidades de los medicamentos que comercializan. [15]

Como introduce el análisis "Social Media likes Healthcare" publicado por la consultora PricewaterhouseCoopers (PWC) en abril de 2012 [16], "los negocios sabios saben que deben ir allí en donde estén sus clientes". Mientras en mercados como el turismo, o el "retail" -productos de consumo- vieron rápidamente el potencial de los medios sociales, el sector sanitario se ha movido con mayor lentitud. Según PWC los diferentes stakeholders del sector farmacéutico [ver 1.2.3] pueden beneficiarse de esta nueva forma de comunicación interactiva. Rápidamente, cualquier persona puede publicar sus experiencias con médicos, medicamentos, dispositivos médicos, hospitales, etc. Siguiendo con datos de PWC, un tercio de los consumidores -datos de una encuesta realizada en EEUU- utilizan los medios sociales con fines relacionados con la salud. Es importante notar también que la actividad en redes sociales es diez veces mayor que en las páginas de las propias compañías farmacéuticas, para los mismos temas.

Del informe se desprende también que los "early adopters [1]" en la industria farmacéutica no sólo se preocupan de medir y analizar la actividad de los medios sociales, sino que también los están incorporando en su estrategia de negocio, participando activamente

[1] Primeros consumidores de una compañía, producto o tecnología. Es decir, una vez lanzado el producto, son aquellas personas propensas a realizar la adquisición antes que la mayoría.

en diferentes redes sociales, comunidades de pacientes, etc.

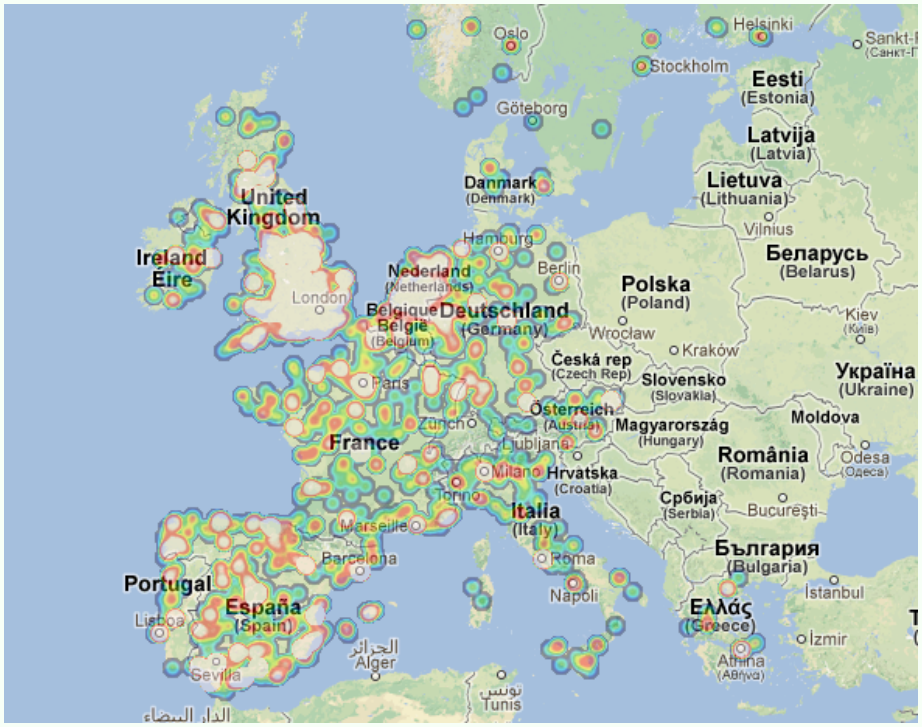
Con estas oportunidades aparecen también nuevos retos y amenazas, en forma de nuevos competidores. Cuando las farmacéuticas se muestran transparentes, las expectativas de los clientes aumentan, y a medida que aumenta la recogida de información que las compañías realizan sobre sus pacientes, mayor protección necesitan estos últimos para asegurar su privacidad y seguridad.

Los medios sociales ofrecen a los individuos nuevas maneras de gestionar su salud, ya sea consultando acerca de una enfermedad concreta, o uniéndose un grupo de soporte -o comunidad de pacientes- para compartir experiencias. El aspecto virtual de los medios sociales, potencia la comunicación ya que el usuario puede escribir desde donde se sienta cómodo (su casa, biblioteca, centro médico, etc.) y habitualmente de forma anónima.

El análisis de PWC destaca Facebook y Youtube como los canales preferidos a la hora de consultar información relacionada con la salud. Si tenemos en cuenta la edad, la gente de entre 18 y 24 años son el colectivo que muestra mayor predisposición a compartir sus experiencias, mientras que la generación del "baby boom" de entre 45 y 64 años son los que menos comparten en cualquier canal social. Así, más del 80 % de los americanos de entre 18 y 24 años estarían dispuestos a compartir su información sanitaria a través de los medios sociales, porcentaje que baja al 45 % cuando se mide en población de entre 45 y 64 años.

Quit Map: Dejar de fumar con el apoyo de los medios sociales

A finales de 2012 una iniciativa conjunta entre la "Irish Heart Foundation" y la compañía farmacéutica Pfizer puso en marcha una página web para ayudar a dejar de fumar. La campaña intenta que la gente ponga en común sus experiencias con el proceso de dejar el tabaco. Dicha página muestra un mapa de Europa en el que se muestra una "visión social del fenómeno de dejar de fumar". El mapa monitoriza en tiempo real el número de personas que hablan en Twitter de dejar de fumar, lo que ayuda a señalar aquella áreas donde más se está intentado dejar el tabaco.



`https://www.quitwithhelp.ie/quitmap`

El mapa refleja una tendencia creciente: el "fumador social" se está convirtiendo en el "dejador de fumar social", ya que los fumadores acuden a las redes sociales en busca de soporte en su propósito. Las utilizan tanto para anunciar su progreso, como en busca de consejo y ánimo por parte de amigos, familiares, e incluso extraños.

Para elaborar el mapa, el mecanismo es sencillo. Se cuentan las veces que aparecen un conjunto de palabras clave en Twitter, y con eso crean el mapa de calor que se muestra. En España, por ejemplo, las expresiones utilizadas son: "dejar de fumar", "dejar el cigarrillo", "quit smoking", "abandonar el tabaco" y "dejo el tabaco".

Cuando se trata de una decisión relacionada con la salud, los medios sociales pueden convertirse en una nueva fuente de información y vía de diálogo. Algunos pueden compartir sus objetivos para conseguir apoyo. También pueden ser utilizados para fidelizar a determinados grupos de enfermos mediante plataformas de pacientes.

En la figura 2.2 representamos resultados del informe de PWC [16]. Según éste, actualmente el 42 % de los consumidores -americanos- han utilizado las redes sociales para acceder a contenido relacionado con la salud publicado por otros consumidores en medios sociales.

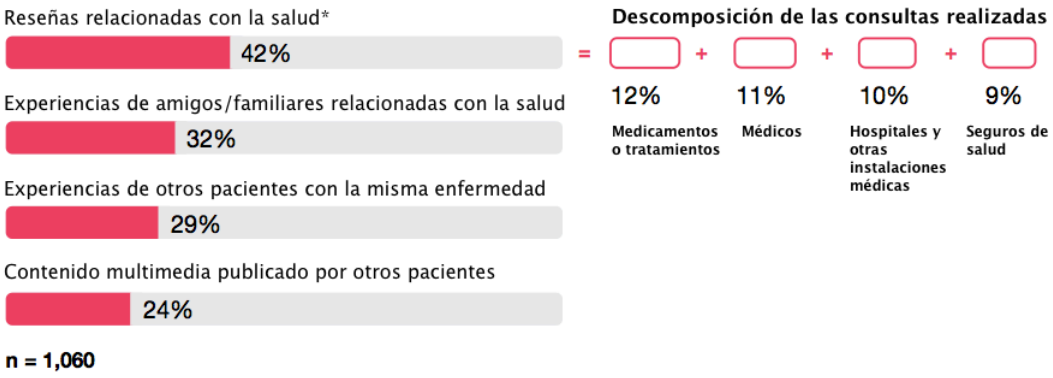


Figura 2.2. Porcentaje de consumidores consultando información sanitaria en medios sociales. *Reseñas acerca de medicamentos o tratamientos, hospitales o otras instalaciones médicas, doctores o seguros médicos. Fuente: PWC.

Además de acceder a información acerca de la salud, los medios sociales permiten a los consumidores publicar contenido. La figura 2.3 resume los hábitos en ese sentido, siendo el 100 %, el total de consumidores consultando contenidos sanitarios en redes sociales. La actividad más popular es la de ayudar a algún amigo o familiar en su causa, seguida de la publicación de experiencias de salud de terceros. Es muy interesante de cara a nuestro análisis el hecho de que más de el 15 % de la muestra haya realizado actividades como compartir síntomas, publicar opiniones de médicos, o escribir acerca de medicamentos o tratamientos. Todas estas publicaciones pueden ser de gran interés para cualquier empresa dedicada a comercializar productos sanitarios.

Además, es importante comparar estos números con los obtenidos al preguntar, por ejemplo, por reseñas de restaurantes. En ese caso, el 27 % del total de consumidores afirma haber publicado alguna. Esto parece indicar que la tendencia en el sector "healthcare"

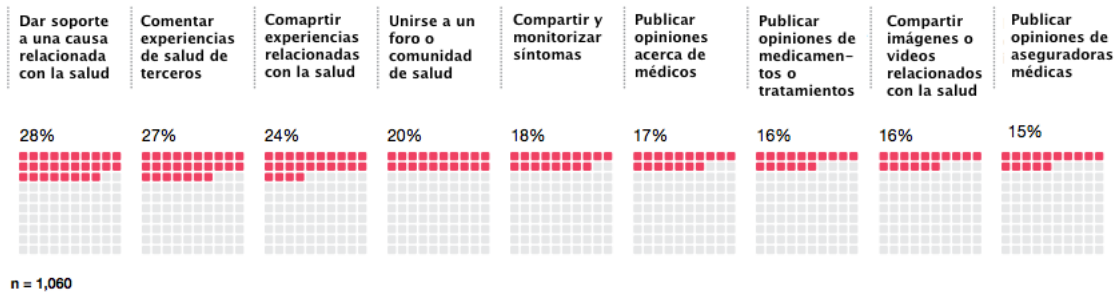


Figura 2.3. Porcentaje de consumidores utilizando medios sociales para actividades relacionadas con la salud. Fuente: PWC.

es a incrementar los porcentajes representados en la figura 2.3.

Es interesante analizar también las diferencias a la hora de compartir experiencias de salud, en función de si estas son positiva o negativas. Como se puede observar en la figura 2.4, las diferencias son reducidas, pero siempre a favor de las experiencias positivas. Ese dato es importante de cara al análisis. Imaginemos, por ejemplo, que para un medicamento encontramos una cantidad superior de experiencias negativas. En ese caso, podríamos asumir, sin arriesgarnos demasiado, que el número de consumidores descontentos es mayor al de los satisfechos, ya que a pesar de que los segundos tienden a publicar más sus experiencias, son los mensajes con experiencias negativas los que predominan.

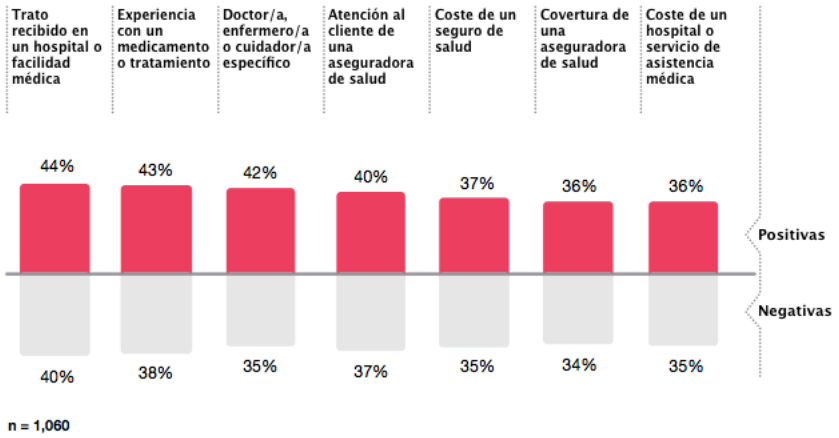


Figura 2.4. Propensión de compartir en la red una experiencia según si es positiva o negativa. Fuente: PWC.

Con todo, los consumidores utilizan cada vez más información de los medios sociales

en la toma de decisiones relacionadas con la salud. El 45 % afirma que afectaría su decisión de pedir un segunda opinión, y el 34 % modificaría la decisión de tomar un medicamento.

En cuestión de confianza los resultados, representados en la figura 2.5, del informe de PWC son también muy relevantes para nuestro propósito y es que los usuarios que mayor confianza generan son los doctores, por el contrario, son las compañías farmacéuticas las que menos confianza inspiran.

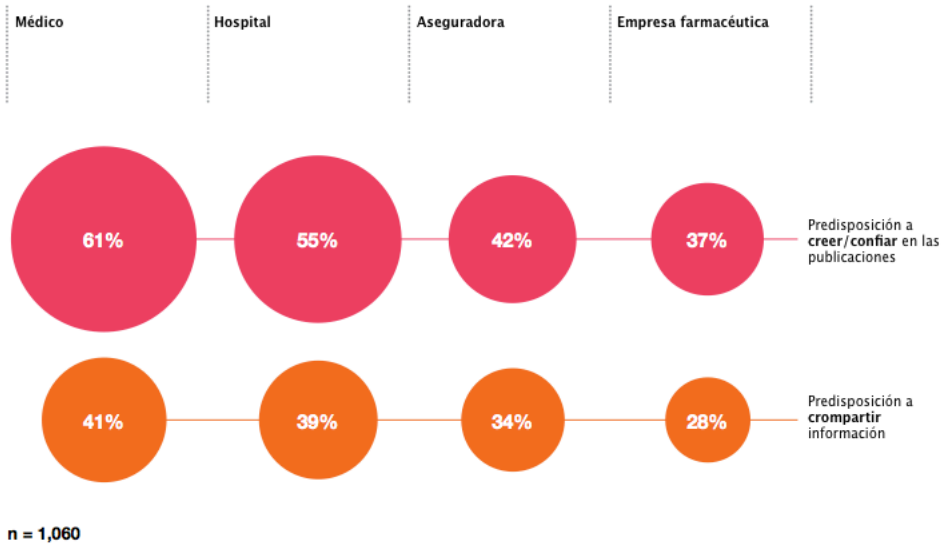


Figura 2.5. Predisposición de los consumidores a creer o compartir información con diferentes colectivos. Fuente: PWC.

Eso significa que los usuarios de redes sociales tienen tendencia a confiar más en la información publicada a través de cuentas de médicos u hospitales, y menos en la información que proviene de aseguradoras y farmacéuticas. Este fenómeno se puede explicar ya que las personas prefieren confiar en aquéllos que les prestan directamente, cuidan de ellos. Es más fácil confiar en una persona que en una organización. Médicos y hospitales tienen la posibilidad de establecer relaciones con los enfermos, lo que incrementa la confianza de estos, tanto a la hora de compartir información, como a la hora de confiar en una publicación.

2.2.1. Del marketing en medios sociales a la estrategia de negocio social

El marketing en redes sociales consiste en captar la atención del destinatario a través de redes sociales, y conseguir así un objetivo: dar a conocer la empresa o la marca, hacer

que el destinatario recuerde un producto, mejorar la imagen, etc.

Con los datos que hemos ido aportando con anterioridad, queda claro que no es suficiente con mantener estrategias de marketing en redes sociales que funcionen de forma unidireccional. Por el contrario es necesario que las empresas sean partícipes del cambio. Para ello es necesario tener en cuenta los siguientes factores [17]:

1. La comunicación está cada vez más controlada por los consumidores. Así, en lugar de invertir en publicidad electrónica tradicional (listas de correo, páginas web, etc.), es necesario mantener un diálogo abierto entre empresa y "stakeholders" de manera que puedan contribuir diferentes participantes.

2. Los pacientes, cada vez más, participan activamente en su propia salud. Los medios sociales representan nuevas oportunidades a la hora de gestionar la salud, ya sea buscando información acerca de una enfermedad concreta, o uniéndose a un grupo de soporte. Utilizan las redes sociales además para formarse o apoyar decisiones relacionadas con la salud.

3. Mejor acceso a la información significa que los pacientes piden mayor transparencia. Jamie Heywood, co-fundador y presidente de PatiensLikeMe, la red social de pacientes más importante del mundo afirma que "los medios sociales afectarán a todos los participantes en el sector de salud, aportando transparencia en el coste, valor y resultados de los procesos y productos de salud".

4. Incrementar la velocidad de comunicación y retroalimentación es beneficioso tanto para organizaciones como para consumidores. Las farmacéuticas pueden conocer muy rápidamente la percepción de un producto en el mercado. Por otro lado, los pacientes esperan respuestas casi instantáneas a sus consultas por parte de farmacéuticas, médicos, hospitales, etc. Por ejemplo, según el informe de PWC, el 49 % de los encuestados espera que de realizar una consulta a través de redes sociales, ésta sea contestada en unas horas (menos de un día).

5. La información de los medios sociales está impactando el modo en el que los pacientes seleccionan su tratamiento, así como su proveedor sanitario. Cada vez más, los pacientes consultan las redes sociales antes de realizar una decisión relacionada con su salud.

6. Los medios sociales llevan a la gente a incrementar su nivel de confianza hacia los gestores de su salud. Así, como mostrábamos en la figura 2.5, hasta el 61 %

de los encuestados se mostraba dispuesto a confiar en información publicada en redes sociales por médicos. La gente también muestra predisposición a compartir información personal con las instituciones y empresas.

7. Los medios sociales, que se han utilizado inicialmente como herramienta de marketing, deben formar parte de la estrategia de negocio. Los medios sociales deben ser un canal para extender la comunicación entre colectivos, responder dudas y prestar asistencia.

8. Los proveedores sanitarios deben utilizar los medios sociales para medir sus resultados. El marketing social debe centrarse en escuchar, más que en divulgar. Monitorizar el contenido de los medios sociales permite a las compañías adaptarse y tomar mayor relevancia para los consumidores [18]. En el mercado farmacéutico, en el que existen estrictas regulaciones legales acerca de la promoción de productos, cobra especial importancia la cita del Vicepresidente Ejecutivo de global marketing de Ford, James D. Farley, Jr. [19], "No lo digas, haz que se lo digan los unos a los otros".

2.3. Cómo monitorizar la actividad en medios sociales

Hemos visto cómo es necesario establecer una estrategia que incluya los medios sociales, y cómo éstos pueden influir de forma notable las decisiones de los pacientes. Se ha expuesto, además, que las redes sociales no pueden ser consideradas un canal más de difusión, sino que deben ser utilizadas como una canal de comunicación bidirecciónal en la que es importante escuchar primero lo que dicen los diferentes participantes, para poder luego interactuar del modo en el que se crea conveniente.

El grueso de este trabajo se basa en el diseño y construcción de una plataforma que permita a las empresas farmacéuticas "escuchar" eficazmente lo que se dice en los medios sociales acerca de diferentes temas que les resultan de especial interés.

Como punto de partida seguiremos el modelo propuesto por Carolin Kaiser y Freimut Bodendorf en su publicación "Mining Patient Experiences on Web 2.0" [15], en el que describen el modo de recoger y analizar reseñas de medicamentos publicadas en redes sociales. Objetivo que es muy similar al de este proyecto.

Cómo los autores señalan, dado que existen en los medios sociales una gran cantidad de publicaciones de pacientes acerca de asuntos médicos, un análisis manual sólo es

posible a una escala limitada. Es por eso que es necesario realizar una extracción y análisis de las opiniones de pacientes aplicando métodos automatizados.

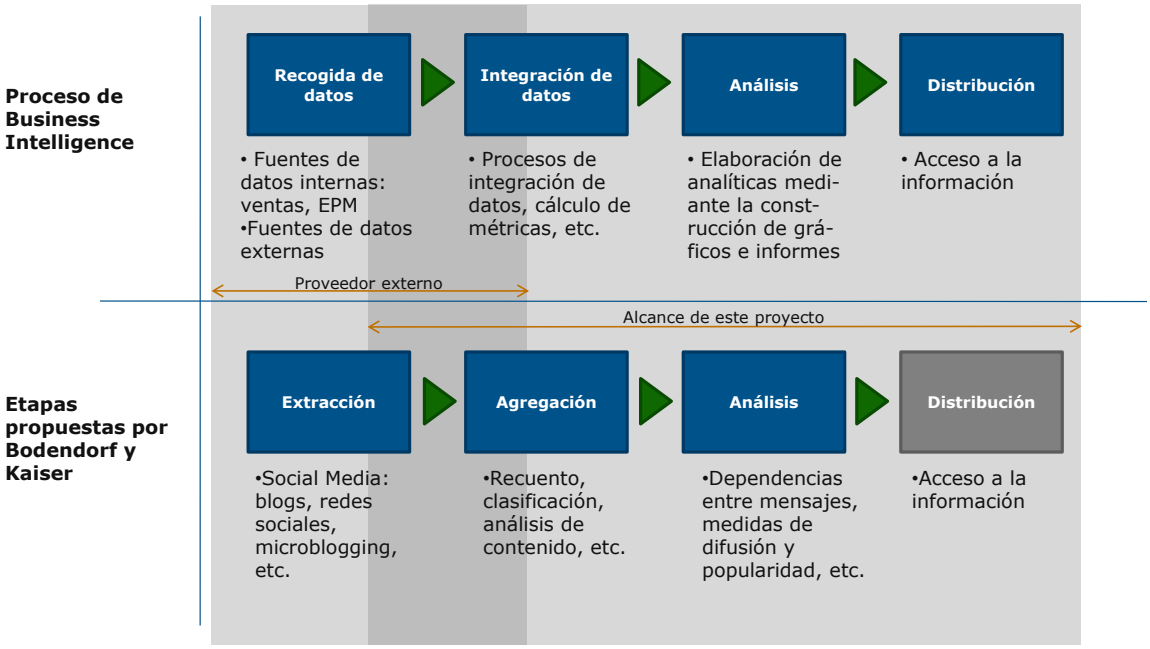


Figura 2.6. Business Intelligence con medios sociales. Aplicación del proceso de social media a la recogida de datos de medicamentos en medios sociales según Bodendorf y Kaiser.

Aunque existen muchas publicaciones que hacen referencia a técnicas de análisis de información de medios sociales, para productos de gran consumo, desafortunadamente, las que hacen referencia al mercado farmacéutico son muy limitadas. En el análisis de Bodendorf y Kaiser proponen un método que consta de tres etapas:

1. Utilización de métodos de análisis de texto para extraer, evaluar y clasificar de los medios sociales, publicaciones relacionadas con los temas que resultan de interés
2. Agregación de la información extraída y clasificada en la etapa anterior para realizar análisis y comparaciones entre productos que compiten en el mismo mercado.
3. Los mensajes son evaluados más profundamente utilizando métodos de "data mining".

La figura 2.6 muestra el paralelismo entre estas etapas y el proceso de Business Intelligence que se explicó en la sección 1.2.2 y representado en la figura 1.1. El primer

paso es la recogida de información, que es también la primera etapa del proceso de Bodendorf y Kaiser. A continuación tenemos la integración de datos y el análisis de los mismos, que coincide con la etapa de agregación y análisis. Luego los autores proponen una segunda evaluación, que quedaría aún integrada en el proceso de análisis. Para terminar, y aunque los autores no lo citan, es necesario dar acceso a los resultados.

2.3.1. Extracción de la información

La extracción de los mensajes queda fuera del alcance del proyecto, pero es importante describir de un modo general qué se realiza y cómo, para facilitar una visión general del proceso, desde la recogida de información hasta la entrega de analíticas o los clientes.

El objetivo de fase de extracción consiste en recoger mensajes de medios sociales, especializados en salud o no, y establecer una primera clasificación que ofrezca una "materia prima" sólida a la fase de integración. Para ello se utilizará el mercado como primer factor de clasificación. A pesar de que hemos hablado del mercado farmacéutico en general, debemos ahora ahondar en el tema para entender que dentro del mismo se pueden realizar muchas clasificaciones. Nosotros utilizamos una división de submercados por patologías, del mismo modo en el que muchas veces se dividen las unidades de negocio dentro de las empresas farmacéuticas. Así, no es lo mismo analizar lo que sucede en el mercado de la diabetes, que en el mercado de la artritis reumatoide, o que en el mercado de la gripe, y así con todos los demás.

Dentro de un mercado, debemos identificar características de los medicamentos. Según Bodendorf y Kaiser las características son el precio, los efectos secundarios, la efectividad y la manipulación de los medicamentos. Nosotros a esas características las llamamos dimensiones, y hemos ampliado el abanico hasta 13, aunque el objetivo es que el número sea flexible. Las dimensiones pueden considerarse como los descriptores de la temática de los mensajes.

Capítulo 3

Estado del Arte

Los servicios web proporcionan un entorno estándar de interoperabilidad entre diferentes aplicaciones de software que se ejecutan en diferentes plataformas y/o frameworks [20]. Esta definición, que ofrecía en 2004 el W3C, deja patente que a la hora de plantear el diseño de un servicio web entrarán en juego diferentes componentes que tendremos que analizar por separado antes de enacajarlas y ponerla en funcionamiento como una plataforma [Sección 5].

Parte del contenido que podría haberse incluido en esta sección se ha expuesto, por coherencia con el desarrollo de la memoria, en otras secciones anteriores (Social media, BI, etc.). En esta sección presentaremos, sin embargo, las herramientas tecnológicas, así como técnicas y metodologías que son la base de la realización de este proyecto.

3.1. Aplicaciones cliente-servidor

La arquitectura cliente-servidor es un modelo de aplicación distribuida en el que las tareas se reparten entre los proveedores de recursos o servicios, llamados servidores, y los demandantes, llamados clientes.

Un cliente realiza peticiones a otro programa, el servidor, quien le da respuesta. En esta arquitectura la capacidad de proceso está repartida entre los clientes y los servidores, aunque son más importantes las ventajas de tipo organizativo debidas a la centralización de la gestión de la información y la separación de responsabilidades, lo que facilita y clarifica el diseño del sistema [21].

3.2. Tecnologías Web

Desde el nacimiento de internet, éste no ha dejado de evolucionar hasta que prácticamente todo el mundo [en países desarrollados], tiene acceso al sistema. La evolución y crecimiento de la web ha venido acompañada e impulsada por la aparición de un conjunto de tecnologías que facilitan la implementación de aplicaciones, así como optimizan el uso de recursos, tanto locales como de la propia red de comunicaciones; éstas son las tecnologías web.

Aunque existen una gran cantidad de proyectos que podríamos considerar tecnologías web, veremos a continuación las que tendrán un impacto más fuerte sobre la plataforma analítica que se está desarrollando.

3.2.1. HTML4 - HTML5

HTML es el acrónimo de "HyperText Markup Language" y hace referencia al lenguaje de marcado predominante para la elaboración de páginas web que se utiliza para describir y traducir la estructura y la información en forma de texto así como para complementar el texto con objetos tales como imágenes [22].

El HTML da a los autores las herramientas para [23]:

- Publicar documentos en línea con encabezados, textos, tablas, listas, fotos, etc.
- Obtener información en línea a través de vínculos de hiperexto, haciendo clic con el botón de un ratón.
- Diseñar formularios para realizar transacciones con servicios remotos para buscar información, hacer reservas, pedir productos, etc.
- Incluir hojas de cálculo, videoclips, sonidos, y otras aplicaciones directamente en sus documentos

La versión 4.0 de HTML aparece en 1998 y la versión 4.01 que corrige errores e introduce algunos cambios en 1999. Esta versión es actualmente soportada por la mayoría de navegadores, y aunque no siempre se obtienen los mismos resultados con diferentes navegadores, se puede considerar que una página bien diseñada en HTML 4.0 funcionará correctamente.

HTML5 es la quinta revisión importante de HTML. Establece una serie de nuevos elementos y atributos que reflejan el uso típico de los sitios web modernos. Con la nueva versión aparecen nuevas funcionalidades a través de una interfaz estandarizada, mejoras en algunos elementos ya existentes, y la desaparición de algunos otros. HTML5 facilita el desarrollo de páginas web interactivas aunque muchos de los navegadores en uso actualmente [ver figura 3.2] no lo soportan total o parcialmente [ver figura 3.1]

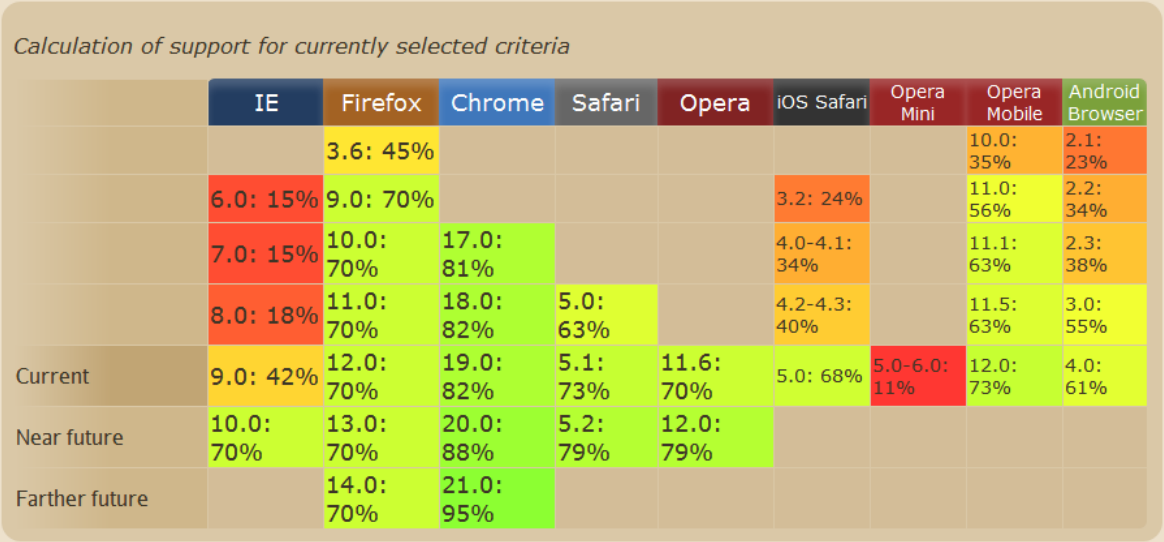


Figura 3.1. Compatibilidad de diferentes navegadores con HTML5. Principales navegadores y su compatibilidad con HTML5 según la versión. Fuente: http://caniuse.com.

Si se analizan conjuntamente las figuras 3.2 y 3.1 se llega a la conclusión de que para hacer una página web que pueda llegar al mayor número de usuarios, y en especial, a usuarios corporativos, con poca libertad para elegir qué hardware y software utilizan, es mejor utilizar por el momento HTML4 y tecnologías web compatibles.

A continuación se verán tecnologías web que unidas a HTML4.01 hacen posible alcanzar resultados muy similares a los que se alcanzarían con HTML5, pero compatibles con la mayoría de navegadores.

3.2.2. Javascript

JavaScript es un lenguaje de programación interpretado, orientado a objetos, basado en prototipos, imperativo, débilmente tipado y dinámico [24].

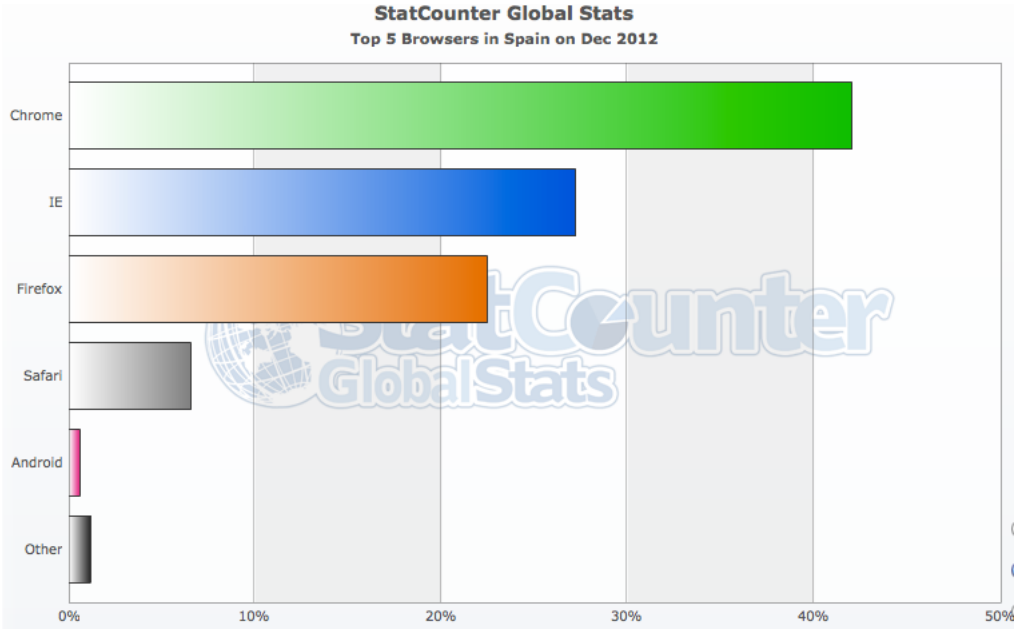


Figura 3.2. Cuota de mercado de los principales navegadores. Principales navegadores en diciembre de 2012 en España. Fuente: http://gs.statcounter.com/.

Se utiliza principalmente en su forma del lado del cliente, aunque también puede ser utilizado en el lado del servidor (i.e.: NodeJS[1]) También es utilizado en aplicaciones externas al mundo web, como en documentos PDF, para la generación de hiperenlaces internos, o en aplicaciones de escritorio como widgets.

Todos los navegadores modernos interpretan el código JavaScript integrado en las páginas web. Para interactuar con una página web se provee al lenguaje JavaScript de una implementación del Document Object Model (DOM) [24].

Javascript, además de poder acceder y modificar el DOM, permite interactuar con el servidor, de modo que una página puede servir contenido de forma dinámica a través del mismo, lo que se hace habitualmente utilizando AJAX.

AJAX es el acrónimo de Asynchronous JavaScript And XML (JavaScript asíncrono y XML). Consiste en una técnica de desarrollo web para crear aplicaciones interactivas de modo que es posible realizar cambios sobre las páginas sin necesidad de recargarlas.

Ajax es una tecnología asíncrona en el sentido de que los datos adicionales se solicitan al servidor y se cargan en segundo plano sin interferir con la visualización ni el comporta-

[1]Servidor orientado a eventos basado en este lenguaje, cuyo uso se está extendiendo en la actualidad como una de las alternativas para aplicaciones web de tiempo real.

miento de la página. El acceso a los datos se realiza mediante XMLHttpRequest, objeto disponible en los navegadores actuales. El contenido asíncrono es normalmente XML o JSON.

Ajax es una técnica válida para múltiples plataformas y utilizable en muchos sistemas operativos y navegadores dado que está basado en estándares abiertos como JavaScript y Document Object Model (DOM).

Otro concepto fundamental basado en JavaScript son la librerías disponibles en ese lenguaje, y de entre éstas, la más utilizada: jQuery.

jQuery como es definido en su página web[2] es una librería de Javascript rápida y ligera. Facilita la manipulación de elementos del DOM, así como el control de eventos. jQuery funciona además con todos los navegadores (modernos) y es extensible.

Esta última propiedad es muy importante, y es que existen muchas otras librerías de Javascript, que están basadas directamente en jQuery. Con todo, esta librería se ha convertido en una de las tecnologías web más utilizadas en el desarrollo de servicios en internet.

3.2.3. Highcharts

HighCharts es una librería escrita en Javascript que permite la creación de gráficas interactivas. La última versión disponible es la 3.0, descrita por sus creadores cómo "librería de gráficas escrita por completo en HTML5/Javascript, ofreciendo gráficos intuitivos e interactivos para una página web".

La librería funciona en las veriones recientes de los navegadores más populares, y los gráficos que se pueden crear son muy personalizables y ponen al alcance del desarrollador una gran cantidad de funciones interactivas cómo tooltips, gestión de eventos, reresco de datos mediante AJAX, etc.

A diferencia de muchas otras librerias diponibles, Highcharts funciona también en dispositivos móviles dado que no utiliza tecnologías como Java o Flash que no están soportadas por todos los navegadores y/o dispositivos.

[2] http://jquery.com

3.3. Business Intelligence

El concepto de Business Intelligence ya introducido en el capítulo 1.2.2 considera las herramientas analíticas, así como las demás técnicas y tecnologías que permiten convertir los datos "crudos" que la empresa tiene disponible en información valiosa para la toma de decisiones.

Un buen ejemplo de una solución de BI completa puede ser un Data Warehouse corporativo. Esto consiste en una base de datos unificada que contiene toda la información de negocio y la hace accesible a toda la compañía.

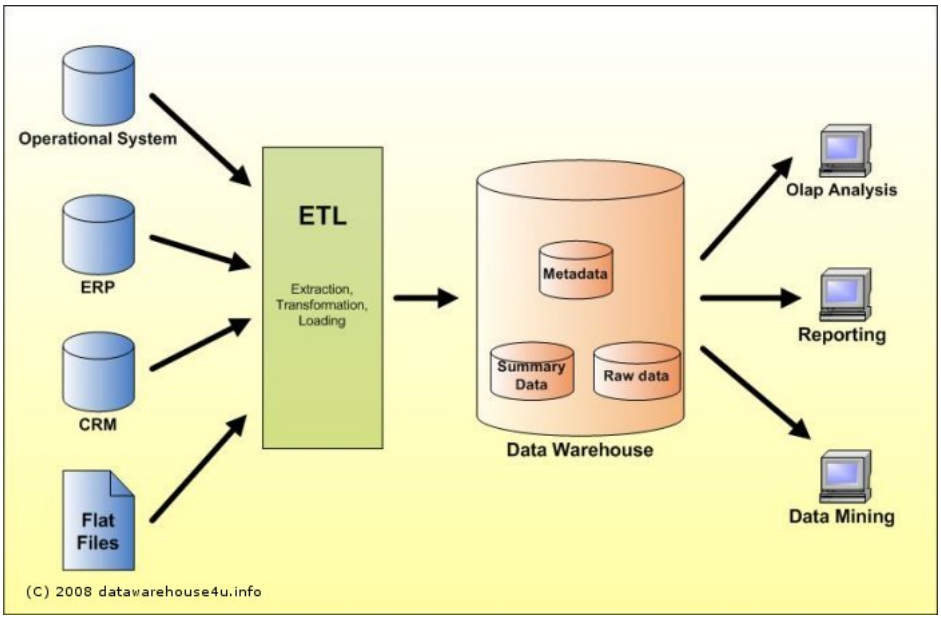


Figura 3.3. Arquitectura genérica de un datawarehouse corporativo. Principales elementos, e interrelación entre los mismos de un datawarehouse corporativo. Fuente: http://datawarehouse4u.info.

Aunque el alcance de este proyecto no consiste en construir un datawarehouse corporativo (más bien estamos trabajando con una de las fuentes de información que podrían alimentar un DW corporativo), hay muchos conceptos entran en juego en la construcción de un almacen de datos corporativo [ver figura 3.3] y que se utilizarán para fines más modestos en este proyecto.

3.3.1. ETL

Acrónimo de "Extract, Transform and Load", una ETL consiste en un proceso que permite mover datos desde una o varias fuentes, reformatearlos y limpiarlos, y cargarlos en

otra base de datos, data mart, o data warehouse.

Cómo el nombre indica, un proceso de ETL tiene tres importantes tareas:

1. Extraer los datos desde los sistemas de origen. Los formatos de las fuentes pueden ir desde bases de datos relacionales o ficheros planos hasta bases de datos no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.
2. Transformar: La fase de transformación aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Algunas fuentes de datos requerirán alguna pequeña manipulación de los datos.
3. Cargar: La fase de carga es el momento en el cual los datos de la fase anterior son cargados en el sistema de destino. En algunas bases de datos se sobrescribe la información antigua con nuevos datos, en otros sistemas se añaden los nuevos datos sin borrar los anteriores.

Existen muchas herramientas dedicadas a la creación y gestión de ETL, ya que en ocasiones estos procesos son de gran complejidad. Algunas herramientas open source disponibles son Kettle de Pentaho o Talend Open Studio. Entre las soluciones propietarias podemos destacar Integration Services, de Microsoft o Informática PowerCenter de Informática.

Como alternativa a las herramientas de creación y gestión de ETL, podemos construir nuestro propio proceso utilizando cualquier leguaje de programación que tenga la capacidad de acceder al origen de datos, transformarlo, y cargarlos en la base de datos de destino. Esta es la alternativa elegida en este proyecto, donde la ETL se ha desarrollado mediante varios scripts en Python.

3.3.2. Bases de Datos

El concepto de Base de Datos tiene un ámbito mayor que el de Business Intelligence. Por otro lado, en este proyecto la base de datos es un elemento más de un sistema de Business Intelligence, por lo que merece la desarrollar el concepto bajo esta sección, y manteniendo un enfoque orientado a BI.

En la definición de BI hemos dicho que su propósito es transformar datos en información. Esto hace patente la necesidad de disponer de un sistema de base de datos

adecuado para el propósito. Para que un sistema de BI funcione correctamente hay que tener en cuenta diferentes aspectos relacionados con Bases de Datos.

1. Tecnología de Base de Datos: Existen diferentes alternativas a la hora de elegir un motor de base de datos. Diferentes bases de datos ofrecen diferentes capacidades, y en la elección es necesario tener en cuenta aspectos como el tamaño que se espera que alcancen las tablas y/o bases de datos necesarias para el funcionamiento de la herramienta, la rapidez de respuesta esperada, así como el modelo de administración de datos que se va a utilizar (Relacional, OLAP, no relacional, etc.) y finalmente el modelo de datos, y es que un modelo de datos no se ajusta igual a las diferentes tecnologías de bases de datos.

2. Modelo de administración de datos: dentro de esta definición, consideramos un modelo de datos la descripción del contenedor de datos, así como de los métodos para almacenar y recuperar información de esos contenedores. Los modelos de datos no son cosas físicas: son abstracciones que permiten la implementación de un sistema eficiente de base de datos; por lo general se refieren a algoritmos, y conceptos matemáticos. Los modelos más importantes son [25]:

   - Bases de datos relacionales: cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para implementar bases de datos ya planificadas. Permiten establecer interconexiones (relaciones) entre los datos (que están guardados en tablas), y a través de dichas conexiones relacionar los datos de ambas tablas.
   - Bases de datos multidimensionales: se utilizan principalmente para crear aplicaciones OLAP y pueden verse como bases de datos de una sola tabla, su peculiaridad es que por cada dimensión tienen un campo (o columna), y otro campo por cada métrica o hecho.

     OLAP es el acrónimo de “On-Line Analytical Processing”. Es una solución utilizada BI cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Este tipo de estructura, basada en BBDD multidimensionales, junto a los editores de consultas existentes, permite recuperar información de la BBDD de una manera muy rápida y flexible, lo que agiliza la navegación por la misma así como la creación de informes
   - Bases de datos transaccionales: Son bases de datos cuyo único fin es el envío y recepción de datos a grandes velocidades. Estas bases son muy poco comunes

y están dirigidas por lo general al entorno de análisis de calidad, datos de producción e industrial. La redundancia y duplicación de información no es un problema como con las demás bases de datos.

- Bases de datos jerárquicas: almacenan la información en una estructura jerárquica que enlaza los registros en forma de estructura de árbol, en donde un nodo padre de información puede tener varios nodos hijo. A diferencia del modelo relacional, las relaciones entre datos se establecen siempre a nivel físico, es decir, mediante referencia a direcciones físicas del medio de almacenamiento, y no mediante relaciones lógicas (a índices o PK).

3. Modelo de datos: herramienta que permite expresar un subconjunto de información para facilitar su comprensión y comunicación. Determina la estructura de datos, normalmente de forma gráfica, considerando los diferentes elementos y la interrelación entre ellos. En el capítulo 5 veremos el modelo de datos de nuestra plataforma.

La base de datos utilizada en este proyecto es MySQL, que es una BBDD relacional, multihilo y multiusuario. MySQL AB (propiedad de Oracle) desarrolla MySQL como software libre en un esquema de licenciamiento dual.

Al contrario de proyectos como Apache, donde el software es desarrollado por una comunidad pública y los derechos de autor del código están en poder del autor individual, MySQL es patrocinado por una empresa privada, que posee el copyright del código fuente.

MySQL es una base de datos muy rápida en la lectura cuando utiliza el motor no transaccional MyISAM, pero puede provocar problemas de integridad en entornos de alta concurrencia en la modificación [26]. Nosotros modificaremos los datos de forma centralizada, mientras que los usuarios de la plataforma analítica accederán a los mismos sin poder modificarlos.

Las diferentes propiedades de MySQL, así cómo el hecho de que sea una base de datos que funciona bien con el framework Django, la hace adecuada como motor de BBDD parece una solución analítica ad-hoc siempre que el volumen de datos almacenado sea moderado: según algunas pruebas [27], si la tabla está bien indexada y se utilizan esos índices en las consultas y uniones, MySQL responde rápidamente con tablas de hasta un millón de registros, aunque oficialmente una tabla puede llegar a tener un tamaño de

hasta 2GB (en un SO con Linux 2.2-Intel 32-bit) [28].

3.3.3. Herramientas analíticas y de reporting

Una herramienta analítica es aquélla que permite realizar informes y analíticas a partir de la información contenida en una fuente de datos. Normalmente la información de carga a una fuente de datos mediante una ETL, y a partir de ésta, o de fuentes de datos intermedias, más acotadas (data marts) se diseñan un conjunto de analíticas que se desarrollan y publican utilizando una herramienta analítica.

Cada herramienta analítica tiene sus particularidades, pero en general la implementación de una analítica o dashboard consta de varias fases:

1. Creación de la conexión con la fuente de datos: Hay que conectar la herramienta analítica con la base de datos que contiene la información. Lo más habitual es que la conexión se haga sobre una base de datos relacional o multidimensional (cubo OLAP).

2. Diseño del conjunto de datos: en la fuente de datos habrá normalmente una gran cantidad de datos, muchos de los cuales no se representarán en la analítica. El diseño del conjunto de datos incluye la creación de las consultas a la fuente de datos, la inclusión de parámetros que permitan realizar consultas dinámicas, la creación de variables calculadas -no presentes en la fuente de datos, pero obtenidas mediante la transformación de datos extraídos de la misma-, etc.

3. Implementación del dashboard: creación del informe propiamente dicho, incluyendo tablas de datos, gráficos, filtros, etc.

Como en el caso de las ETLs, existen muchas herramientas analíticas disponibles, algunas open source, como BIRT o Pentaho Report Designer, y otras de código propietario como QlikView, Reporting Services de Microsoft, SportFire, etc. Por otro lado es posible utilizar soluciones ad-hoc, que permitirán una mayor flexibilidad en la construcción de las analíticas.

Esta sección se desarrolla con más detalle en el capítulo 4.

3.4. Framework

En el origen de la web las páginas servidas eran estáticas, pero con el paso del tiempo esto ha ido cambiando. Actualmente es un requisito para una web de calidad ser dinámica. El usuario debe poder interactuar con la página, adaptarla a sus necesidades, y eso no es posible si no contamos con una herramienta tecnológica que permita añadir lógica programáticamente a la página.

Con ese fin han aparecido diferentes lenguajes y Frameworks. Uno de los más populares, por ejemplo, es PHP, en el que el código se embebe sobre la hoja HTML, y es el servidor web el que lo interpreta, para servir la página de forma personalizada a cada cliente.

En este proyecto se utilizará el framework Django, que es un entorno de desarrollo web a alto nivel, escrito en Python, y diseñado para facilitar el rápido desarrollo de servicios web.

3.4.1. Servidor Web - lighttpd

Un servidor web o servidor HTTP es un programa informático que procesa una aplicación del lado del servidor [Ver sección 3.1]. El código recibido por el cliente suele ser interpretado por un navegador web.

lighttpd es un servidor web diseñado para ser rápido, seguro, flexible, y fiel a los estándares. Está optimizado para entornos donde la velocidad es muy importante, y por eso consume menos CPU y memoria RAM que otros servidores. Lighttpd es software libre y se distribuye bajo la licencia BSD. Funciona en GNU/Linux y UNIX de forma oficial.

3.4.2. Python

Python es un lenguaje de programación creado en Holanda a finales de los años 80 y distribuido bajo una licencia de código abierto compatible con GPL. Es un lenguaje de programación orientado a objetos e interpretado, lo que nos ofrece algunas ventajas, como un rápido desarrollo.

Python busca una mayor facilidad de lectura del código y del diseño de la aplicación. Además existe una gran cantidad de librerías disponibles para complementar el lenguaje,

lo que unido a la facilidad de crear y gestionar clases y objetos, la gestión automática de memoria, y la presencia de estructuras de datos complejas (listas, diccionarios, etc.) hacen del lenguaje una muy buena opción para el desarrollo de programas de tamaño pequeño o mediano.

Hay muchas librerias escritas en Python que facilitan el desarrollo de aplicaciones. Algunas de las que se han utilizado en este proyecto son:

- xlrd: librería para trabajar en modo lectura con hojas de Excel en Python. Necesaria para el desarrollo de los scripts de carga de datos [sección 5.3.2]. La librería es Open Source, y su versión más reciente es la 0.7.3, liberada en febrero de 2012. Se puede descargar desde `https://pypi.python.org/pypi/xlrd`
- pdfcrowd: solución que permite crear pdfs a partir de código HTML, CSS y Javascript. En este caso la API no es de Open Source. El funcionamiento de esta librería permite exportar el contenido de las páginas web servidas a documentos en formato pdf. La renderización se realiza en los servidores de la compañía que presta el servicio, cuyo coste varia en fnción del volumen de exportaciones necesarias. Aunque existen alternativas a este servicio, ninguna de las que se han encontrado ofrece unos resultados tan satisfactorios.

Python es además el lenguaje en el que está programado, y en el que se desarrolla en el Framework de desarrollo web Django, y por lo tanto el lenguaje con el que se ha programado la lógica de la plataforma.

3.4.3. Django

Django es un framework de desarrollo web de código abierto escrito en Python. Fue desarrollado en origen para gestionar varias páginas orientadas a noticias de la World Company de Lawrence, Kansas, y fue liberada al público bajo una licencia BSD en julio de 2005. En junio del 2008 fue anunciado que la recién formada Django Software Foundation se haría cargo de Django en el futuro.

El principal objetivo de Django es facilitar la creación de sitios web complejos. Django pone énfasis en la reutilización, la conectividad y extensibilidad de componentes.

A pesar de que el equipo que mantiene Django se desmarca del seguimiento estricto de cualquier modelo de arquitectura de software, generalmente se considera a Django

como un entorno modelo-vista-controlador.

Modelo Vista Controlador

El Modelo Vista Controlador (MVC) es un patrón de arquitectura de software que separa los datos y la lógica de negocio de una aplicación de la interfaz de usuario y el módulo encargado de gestionar los eventos y las comunicaciones.

Para ello MVC propone la construcción de tres componentes distintos[29]:

- El Modelo: Es la representación específica de la información con la cual el sistema opera. Por lo tanto gestiona todos los accesos a dicha información, tanto consultas como actualizaciones, implementando también los privilegios de acceso que se hayan descrito en las especificaciones de la aplicación (lógica de negocio). Envía a la vista aquella parte de la información que en cada momento se le solicita para que sea mostrada. Las peticiones de acceso o manipulación de información llegan al modelo a través del controlador.
- El Controlador: Responde a eventos (usualmente acciones del usuario) e invoca peticiones al modelo cuando se hace alguna solicitud sobre la información (por ejemplo, editar un documento o un registro en una base de datos). También puede enviar comandos a su vista asociada si se solicita un cambio en la forma en que se presenta de modelo.
- La Vista: Presenta el modelo (información y lógica de negocio) en un formato adecuado para interactuar (usualmente la interfaz de usuario) por tanto requiere de dicho modelo la información que debe representar como salida.

En referencia al seguimiento del paradigma MVC, desde las FAQ de Django plantean la siguiente cuestión [30]:

> *Django parece ser un framework MVC, pero llamáis al "controlador" la "vista", y a la "vista" la llamáis "template". Por que no utilizáis la terminología estándar?*
>
> Bien, los nombres estándares son discutibles. Según nuestra interpretación del MVC, la vista describe la información presentada al usuario. No es necesariamente cómo son esos datos, sino qué datos se presentan. La vista describe que datos se pueden ver, no cómo se pueden ver. Es una distinción sutil.

Así, en nuestro caso, una vista es la función de "callback" de Python para una URL en particular, ya que la función de "callback" describe que datos se van a presentar.

Es más, parece sensato separar el contenido de la presentación, que es donde los "templates" entran en juego. En Django, una vista describe que datos se presentan, aunque normalmente acaba llamando una "template" que es donde se describe cómo se presenta el contenido.

Entonces, ¿dónde tendríamos el controlador? En Django es posiblemente el framework en si mismo: la lógica que envía la petición a la vista apropiada, en función e la configuración de URLs

...

Con todo, Django separa la lógica de los datos, lo que facilitado el desarrollo de aplicaciones web medianamente complejas.

3.4.4. Celery

Celery es una aplicación que nos permite crear tareas de trabajo asíncronas gestionadas por un gestor de colas. Se focaliza en operaciones en tiempo real pero también permite planificar tareas de modo que se ejecuten en un momento determinado o de manera periódica.

Las unidades de ejecución, llamadas tareas, se ejecutan de manera concurrente en uno o más nodos de trabajo. El sistema de mensajería (broker) recomendado por Celery es RabbitMQ aunque se proponen también otras opciones cómo Redis, MongoDB o una base de datos relacional.

Celery surgió ligado a Django, y como éste, está programado en Python. Django-celery es la versión de Celery disponible para funcionar en Django.

Aunque Celery tiene muchas aplicaciones diferentes, como distribuir el procesado de tareas entre diferentes procesos y/o servidores para que el tiempo total de respuesta sea menor [31], en nuestro caso nos interesa para poder controlar la evolución del proceso de ETL.

En ese caso, al recibir la petición que inicia el proceso de ETL, Django pasa el control de la ejecución de la tarea a Celery, y devuelve al cliente un templete informando de que

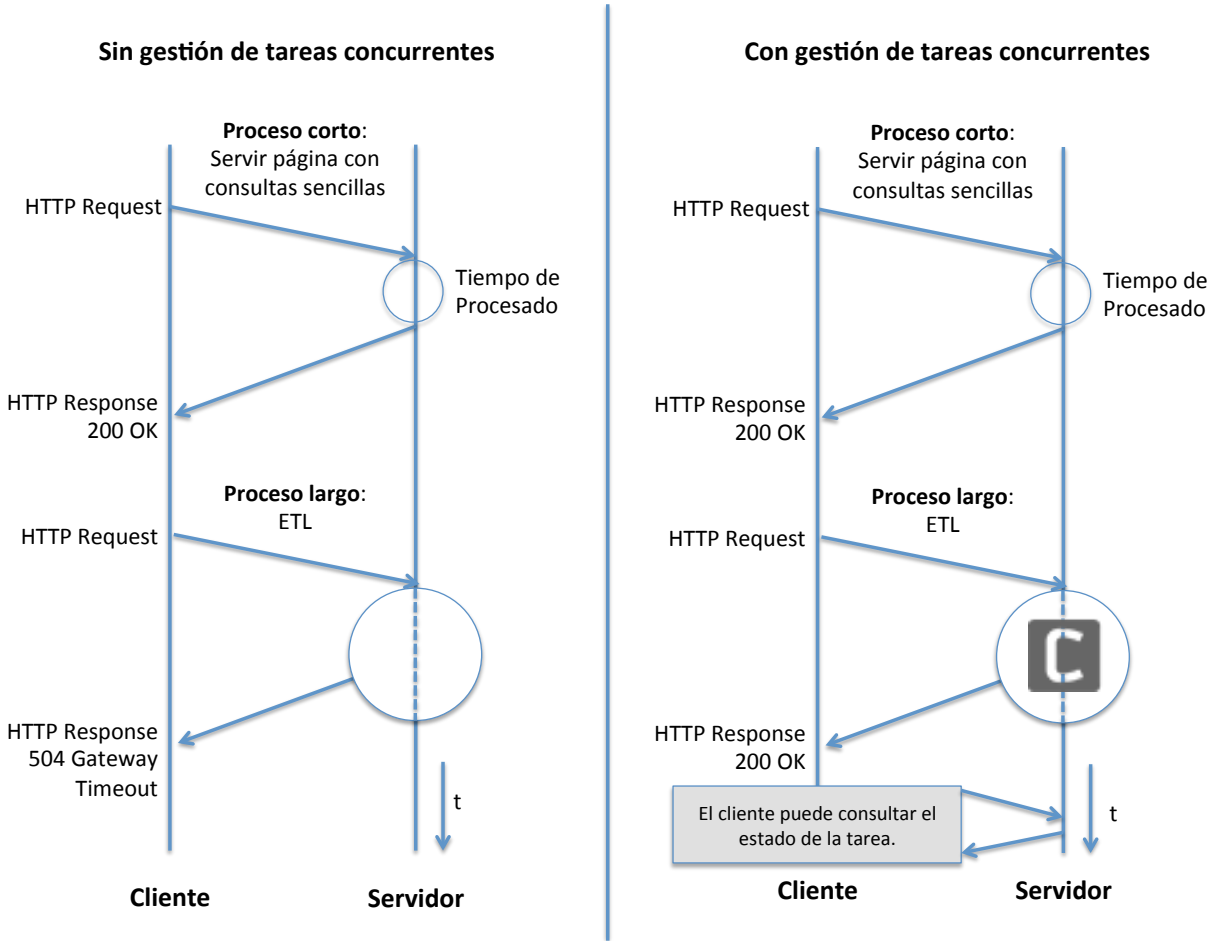


Figura 3.4. Gestión de tareas. Si después de varios segundos, el servidor web no recibe respuesta de Django, devolverá un mensaje 504 de error. No es posible comprobar el estado de la tarea después de haberla iniciado.

se ha iniciado el proceso. En el mismo hay funcionando un método en Javascript que pide periódicamente al broker el estado de la tarea hasta que esta ha terminado de forma satisfactoria o no, mostrando en todo momento los mensajes de estado, y el mensaje de error en caso de haber sucedido alguno.

Capítulo 4

Informes

Para analizar información recogida de los medios sociales es necesario entender bien primero cómo es esa información, y plantearnos que preguntas queremos responder con los análisis.

En un escenario ideal, cuando un laboratorio farmacéutico quiere entender qué se está diciendo en internet acerca de su producto, o de su mercado, lo primero que tendría que realizar el analista, es entender cuál es la pregunta que quiere responder, ya que ésta guiará el modo en el que se llevarán a cabo las siguientes etapas del proyecto. En general, una vez definida/s la/s preguntas/s, el analista se dedicaría a recoger "manualmente" la información, filtrarla, valorarla, clasificarla, representarla y analizarla.

El volumen de información que se genera diariamente en los medios sociales acerca de diferentes temas farmacéuticos, puede ser considerablemente elevado. Además, conseguir toda esa información es un proceso que requiere conocer muy bien las diferentes fuentes disponibles en internet, así como acceder y navegar por las mismas hasta encontrar información relevante. Todo esto hace difícil y muy costoso que el proceso descrito anteriormente pueda llevarse a cabo directamente por personas. Sin embargo, el precursor de las herramientas automáticas de recolección y análisis de datos es, sin lugar a dudas, ese proceso manual que tanto en IMS como en empresas competidoras, se ha realizado en diferentes ocasiones.

Por otro lado, conocer el modo en el que un proyecto "tipo" de análisis de posicionamiento en medios sociales, nos permitirá conocer que preguntas conducen el análisis, y que analíticas son necesarias para la resolución de dichas preguntas.

4.1. Recogida de la información de los medios sociales

Del mismo modo que en el diseño y construcción de una fábrica de sillas de madera, hay que tener en consideración de qué manera se recibirá la madera (cómo, cuándo, cuanta, etc.), en el desarrollo de este proyecto es imprescindible pararse a analizar todo lo relativo a los mensajes que van a nutrir las analíticas que se van a diseñar.

Siguiendo con el paralelismo de la fábrica de sillas, si no se dispone de madera de calidad, es complicado que las sillas fabricadas si lo sean. Así, por muy buenos que sean los métodos de procesado de la información, y por muy potente que sea la herramienta analítica, si no se dispone de suficiente información recogida, o si la información disponible es de mala calidad, será difícil que las analíticas aporten información valiosa.

4.1.1. Objetivos de la recogida

En esta sección se presentarán los objetivos que motivan la búsqueda de información en las redes sociales. Seguramente no son los mismos criterios a la hora de seleccionar información si queremos hacer un análisis del volumen de mensajes disponibles, que si queremos analizar por contenido.

4.1.2. Metodología de la recogida

Dado que la recogida de mensajes queda fuera del alcance de este proyecto, la metodología recorrerá de forma general cómo se ha diseñado el proceso desde el momento en el que se percibe la necesidad de empezar a monitorizar un mercado, hasta el momento en que se dispone de los primeros mensajes.

4.1.3. Principales retos

Como indica Roger Bretau [32], se han identificado un conjunto de factores que complican en general la monitorización de la información publicada en internet. En su análisis cita los siguientes factores:

1. No elegir bien las palabras clave: las palabras claves son un factor fundamental a la hora de conducir la búsqueda hacia los resultados deseados. Si no se eligen correctamente las palabras claves, puede que se recojan muchos resultados que no

tengan que ver con lo que se está buscando, o que no se recojan muchos mensajes que son relevantes.

2. Deficiencia de los sistemas automáticos de monitorización: el autor indica que "no hay ninguna herramienta que sirva para todo", y que el "el factor humano es clave" gracias a su capacidad crítica. Estas son principalmente las razones técnicas [1] por las que se ha tomado la decisión de externalizar la construcción de un sistema de recogida propio. Al trabajar con una empresa colaboradora para realizar el motor de búsqueda que recoge los mensajes tenemos como ventaja que el sistema de búsqueda estará desarrollado de forma específica para nuestro proyecto. Por otro lado, esto nos convierte en corresponsables del funcionamiento adecuado del mismo.

3. Tener una idea predeterminada y no querer cambiarla: este hecho podría provocar que los datos salgan de una determinada manera a propósito y que faltemos a la realidad/verdad. Este hecho está relacionado con la propia naturaleza del proceso de búsqueda de información subjetiva. Por ejemplo, si queremos probar, buscando en internet que la aspirina es buena, será una tarea sencilla, casi tanto como demostrar todo lo contrario.

4. Querer medir demasiadas cosas y no enfocarse realmente a los objetivos: a veces podemos caer en el error de entusiasmarnos con la medición de ciertos aspectos que en realidad no son interesantes para nuestros objetivos. Esto desvía la atención de lo que realmente es importante, y consume unos recursos que aportarían más valor de ser utilizados en otra cosa. Es importante, antes de realizar un análisis, entender el coste de oportunidad del mismo, y decidir luego si merece la pena seguir adelante con el mismo. Este punto será muy importante a la hora de realizar el diseño de los informes.

5. Monitorizar sólo cuando hay problemas: una actitud reactiva es una opción, que hasta el momento ha motivado la aparición de los proyectos en los que ha sido necesario realizar búsquedas ad-hoc, pero cuando tenemos un cierto volumen en redes sociales, es muy recomendable realizar una monitorización constante (mensual, semanal, etc.)

6. Centrarnos sólo en las menciones y obviar las no menciones: a veces es tan importante lo que se dice, cómo lo que no se dice, o como lo denomina el autor, las "no

[1] Existen también razones económicas y de organización

menciones". Muchas veces se utiliza para el análisis lo que se ha dicho sobre un producto, pero no se presta la atención necesaria a lo que no se ha dicho.

Con los factores anteriores en mente, y teniendo en cuenta las peculiaridades del mercado farmacéutico que se han presentando a lo largo de esta memoria, podemos extraer un conjunto de características que serán necesarias en nuestro sistema de recogida de mensajes.

La recogida de mensajes deberá de ser automática. Como ya se ha dicho, necesitamos recoger los mensajes en tiempo real y de forma que la información pueda comercializarse a un precio razonable, lo que hace necesario automatizar la recogida y un primera clasificación de esta información.

Es necesario acotar el ámbito de la búsqueda. No analizaremos todo el mercado farmacéutico de golpe. Considerar todo el contenido publicado acerca de un sector tan amplio como el farmacéutico llevaría a resultados sobre los que sería muy difícil extraer conclusiones. Además, las necesidades de información casi siempre se pueden acotar dentro de un mercado. Imaginemos que un antigripal quiere saber que se comenta en medios sociales que pueda tener efecto sobre su producto o marca. Los mensajes relacionados con la oncología, la osteoporosis, etc., no aportarán ningún valor a la hora de desvelar esas dudas.

Se podría haber optado por utilizar algún otro concepto para delimitar la búsqueda. Podríamos por ejemplo realizar búsquedas más acotadas, por marca o por producto, o búsquedas más transversales, utilizando la compañía como referencia, sin importar el mercado o el producto en cuestión. Se ha optado por este criterio dado que lo más habitual en la industria es analizar mercado a mercado, y dentro del mismo se quiere conocer la posición competitiva de los diferentes participantes (productos y/o marcas).

Lógicamente, para iniciar una búsqueda automática de contenido acerca de un mercado farmacéutico determinado, no es suficiente con conocer el nombre del mismo. Es necesario establecer un conjunto de criterios de búsqueda, que se utilizarán para controlar y orientar las búsquedas. Tener que establecer un conjunto de palabras clave implica el riesgo de polarizar la búsqueda a través de las mismas, por lo que será especialmente importante fijarlas bien, y no utilizar este método para influir de forma implícita e involuntaria los resultados.

También es importante considerar en que "lugar/es" se va a recoger información, y cómo se va a acceder a esos lugares. Por ejemplo, Google recorre la web de enlace en enlace para ir indexando las páginas que se va encontrando. Tratar de diseñar robots (procesos) que funcionasen de ese modo sería poco efectivo, ya que en general las noticias relacionadas con el mercado farmacéutico están concentradas en un conjunto de fuentes especializadas (blogs de salud, comunidades de pacientes, medios de comunicación especializados, etc.). Recorrer indiscriminadamente "toda la web" implicaría mayores costes, ya que requeriría servidores más potentes, bases de datos mayores, etc., y probablemente no resultaría en un incremento en el número ni la calidad de los resultados.

Para obtener los resultados deseados, se utilizarán dos clases de fuentes:

- Fuentes conocidas que se configurarán manualmente. Como se ha dicho, es posible conocer de antemano muchas de las fuentes que son susceptibles de contener información relacionada con el mercado farmacéutico. Es por eso que si queremos empezar a obtener resultados rápidamente al configurar una nueva búsqueda, será necesario introducir en que fuentes se va a realizar la búsqueda. De hecho, para explotar todas las posibilidades de información que estas fuentes ofrecen, el motor de búsqueda ejecuta procesos diseñados especialmente para cada una de esas fuentes, con lo que se puede aprovechar al máximo las herramientas que la misma web pone a disposición de los usuarios para explorar el contenido (buscadores, RSS, mapas web, etc.)
- Buscadores. Los parámetros que se utilicen para configurar la búsqueda, se utilizarán en buscadores siguiendo los algoritmos que el desarrollador del motor de búsqueda establezca para encontrar resultados relevantes. Se utilizarán diferentes tipos de buscadores para tratar de llegar al mayor número de fuentes posible.

El sistema de recogida de datos es diferente del sistema de análisis, y por eso será necesario también que este permita extraer fácilmente la información recogida, de modo que sea sencillo utilizar este "output" como "input" del sistema de análisis. Este aspecto se analizará en profundidad en el capítulo 5

La herramienta además de recoger noticias, las procesa para clasificar dos aspectos de las mismas:

- Contenido Objetivo: Dimensiones. Las noticias, se recogen en el ámbito de un mercado, pero eso no es suficiente para clasificarlas. Nos interesa conocer con más

detalle de qué hablan las noticias, y es por eso que aparece el concepto de dimensiones. Podríamos definir una dimensión *como el número mínimo de coordenadas necesarias para especificar cualquier punto de un espacio.* Para nosotros serían las temáticas de las que habla una noticia y que por lo tanto, podríamos decir que de algún modo la definen. Cabe comentar, que aunque consideremos las dimensiones como algo objetivo, como involucra interpretación del lenguaje, no puede considerarse totalmente como tal; pero si nos servirá para clasificar el contenido del mensaje sin entrar a valorar el posicionamiento de su autor.

- Contenido Subjetivo: Sentimiento. Por el contrario, se valorará también la componente subjetiva de la noticia. Es decir, el posicionamiento del autor ante el tema sobre el que escribe.

Si valorar el contenido objetivo de una noticia de manera automática, a través de diferentes algoritmos, e complejo; evaluar el contenido subjetivo lo es mucho más, hasta el punto de que es hoy en día una tecnología en pleno proceso de crecimiento. La herramienta de recogida de mensajes clasifica las dimensiones a través de un conjunto de palabras asociadas a las mismas (se verá con más detalle en la sección 4.1.4), mientras que contenido subjetivo se evaluará utilizando software capaz de "aprender" a partir de una muestra de mensajes evaluados por personas e identificar patrones que sirvan para acabar clasificando las noticias en tres grupos: sentimiento positivo, sentimiento negativo y sentimiento neutro. De cualquier modo, no entra dentro del alcance de este proyecto profundizar en cómo se lleva a cabo la clasificación y preprocesado de mensajes, técnicas cuyo desarrollo podría ser objeto de un proyecto de final de carrera por sí solo.

Para acabar de matizar la diferencia entre la clasificación de la noticia según los dos criterios anteriores utilizaremos un ejemplo publicado en Twitter el 12 de septiembre de 2012, recogido posteriormente por la herramienta y que se muestra en la figura 4.1.

Del contenido objetivo del mensaje, podríamos decir que la usuaria de la red social Twitter comenta la desfinanciación, el precio de un medicamento o el hecho de que padece una enfermedad. Del contenido subjetivo extraemos que la autora está claramente enfadada y que, por lo tanto, el mensaje debería de ser clasificado dentro del grupo de sentimiento negativo.

Siguiendo con el ejemplo, podemos ver cómo clasificar el contenido objetivo sería más sencillo, ya que la noticia contiene la palabra "pagar", por lo que posiblemente habla del

precio del producto, la palabra "Venoruton", por lo que probablemente esté hablando de dicho producto, e incluye también la palabra "trombosis" por lo que cabe imaginar que el mensaje hace referencia de un modo u otro a dicha enfermedad.

Para analizar el contenido subjetivo, nos encontramos con mayores dificultad ya que igual que palabras como "puta" o "vergüenza" tienen connotaciones negativas, "GRACIAS" tiene connotaciones positivas. Esto hace que para evaluar un mensaje sea necesario interpretar el contenido, punto en el que el software a nuestro alcance aún no ha alcanzado del modo que nos gustaría.

Figura 4.1. Micronota publicada en Twitter. Noticia publicada el 1 de septiembre y recogida por la herramienta analítica dentro del mercado de antivaricosos.

4.1.4. Delimitar el mercado sobre el que se recogerá información

Con todo lo introducido en la sección anterior queda patente que antes de empezar la recogida de información será necesario configurar los parámetros de la búsqueda.

El primer paso es identificar un mercado de interés. La definición del mercado quedará implícita al configurar los parámetros de entrada para la búsqueda, pero es una buena práctica definir primero de forma cualitativa el mercado, ya que eso nos ayudará luego a encontrar los parámetros más adecuados para que la búsqueda devuelva un número elevado de resultados, y para que éstos sean realmente de interés para el análisis. A efectos de los análisis que queremos llevar a cabo, utilizaremos patologías como mercados. Así los mensajes que hagan referencia a productos, tratamientos para esa patología, formarán parte del mercado. Del mismo modo, noticias acerca de cómo prevenir o curar la enfermedad, efectos de la misma en la vida del paciente, etc. también se recogerán dentro del mercado.

Para definir un mercado, se han identificado tres grandes grupos de parámetros que tendremos que definir. La calidad y cantidad de noticias recogidas dependerá por un lado de la "calidad" de la herramienta de recogida, es decir, del modo en que está programada, la capacidad de los servidores en los que corre, etc. Por el otro lado, de los parámetros con la que nosotros configuremos la herramienta. Mientras que IMS no tiene control directo sobre el funcionamiento de la herramienta de recogida de noticias, sí que es nuestra responsabilidad el configurar adecuadamente la herramienta para cada uno de los mercados que se contraten.

Los parámetros de configuración se han agrupado en tres bloques [figura 4.2]:

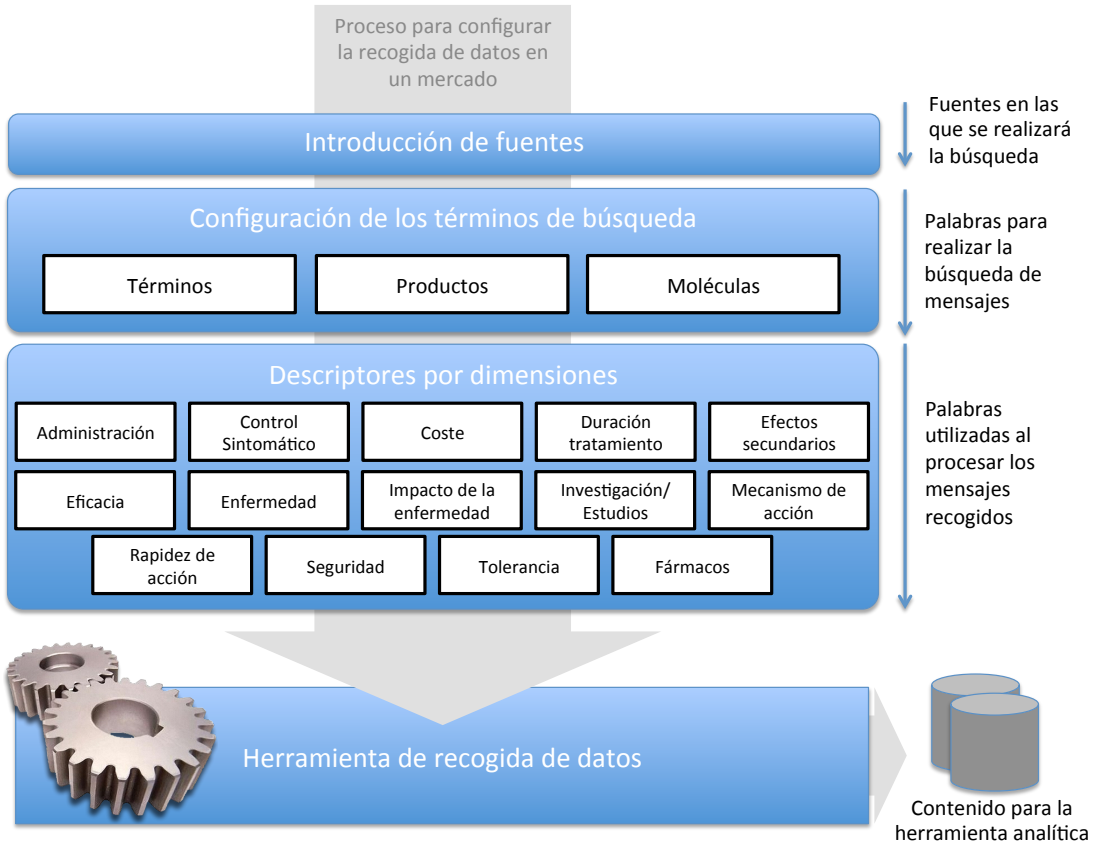


Figura 4.2. Proceso de configuración de un mercado. Información necesaria para definir un mercado y configurar la herramienta de recogida de datos.

- Introducción de fuentes. Hay un conjunto de fuentes a las que la herramienta acude independientemente del mercado que se esté buscando. Pero el hecho de que estemos trabajando con una potencia de computación limitada, y que los mensajes

que se quieren recoger pertenecen a un contexto determinado, se decidió configurar en función del mercado un conjunto de fuentes en las que se buscarán mensajes. Así tenemos una búsqueda mucho más dirigida y, por lo tanto, más eficientes en términos de tiempo y recursos consumidos para encontrar un mensaje relevante. Tener la capacidad de especificar un conjunto de fuentes permite además estudiarlas individualmente y desarrollar algoritmos optimizados para explotar toda la información disponible en la fuente.

- Configuración de los términos de búsqueda. Los términos de búsqueda son aquellas palabras o expresiones que se utilizarán como parámetros de búsqueda. Para conseguir mejores resultados, se han dividido las palabras clave en tres grupos, cada uno de los cuales afecta a la búsqueda de una forma diferente. En la figura 4.3 tenemos un ejemplo de cómo configurar las palabras clave para el mercado de hipogonadismo:

Términos de Búsqueda		
Términos	Productos	Moléculas
Hipogonadismo	Reandron	TESTOSTERONA
síndrome de Klinefelter	Testogel	
Gonadotropina coriónica hur	Itnogen	
Hormona liberadora de gona	Testim	
Testosterona	Testopatch	
Hipogonadotrópico	Testex Prolongatum	
Hipergonadotrópico	Proviron	
deficiencia de 5-alfa-reducta	Testex	
Azoospermia		
Oligospermia		

Figura 4.3. Palabras clave para configurar el mercado de Hipogonadismo. Palabras clave agrupadas por categoría para conducir la recolección de noticias acerca de "Hipogonadismo".

  - Productos. En este grupo tenemos todos los productos que se quieran monitorizar dentro del mercado que se esté configurando
  - Moléculas. La molécula, o principio activo, es el componente de un producto con las propiedades terapéuticas (un medicamento contiene típicamente la molécula y excipientes, estos últimos sin propiedades terapéuticas). En muchos casos se habla directamente de la molécula en lugar de hablar de la marca (i.e. he tomado un Paracetamol/me he tomado un Gelocatil. Paracetamol es la molécula y Gelocatil uno de los nombres comerciales -marca- bajo la que se comercializa). Al poder distinguir entre productos y moléculas podemos establecer relaciones entre ellos.

— Términos. Los términos son aquellas palabras o expresiones que sabemos/intuimos que nos pueden llevar a mensajes relevantes. A diferencia del producto o principio activo, en los que se buscaba la palabra exacta, los términos se convierten a lexemas y la búsqueda se realiza por estos últimos. Por ejemplo, si queremos buscar "deficiencia de ...", la frase se convertiría a algo como "defic de ..." con lo que no solo se encontrarían las noticias que contengan la expresión con deficiencia, también las que contengan "deficiente", etc.

- Descriptores por dimensiones: Los descriptores por dimensiones no se utilizan en la búsqueda de las noticias, sino en su clasificación. Como se ha explicado previamente, una vez recogidos los mensajes se procesan para clasificarlos en dos ámbitos diferentes: sentimiento y dimensiones. Estos descriptores son los criterios según los que se determinará si un mensaje trata de un tema determinado. Por cada dimensión se listará un conjunto de palabras. Si el mensaje contiene alguna de esas palabras, se considerará que la dimensión en cuestión está presente en el mismo. Las dimensiones se han definido teniendo en cuenta aquellos aspectos que son del interés de médicos, pacientes y laboratorios [Ver figura 4.2].

4.2. Herramienta analítica

Una de las primeras decisiones a la hora de poner en marcha una herramienta analítica, es decidir si ésta se va a construir de cero, o si se utilizará una solución disponible en el mercado, es decir, lo que en inglés se suele llamar el "Buy or Build".

La decisión de utilizar una herramienta ya desarrollada o empezar construyendo una, depende principalmente de los requerimientos de las analíticas. Típicamente la decisión se puede tomar en función de[33]:

- Número de informes: cuanto mayor sea el número de informes que hay que desarrollar más complicado es utilizar una herramienta ad-hoc. Las herramientas existentes, normalmente, facilitan la creación de informes además de incluir sistemas para la gestión de los mismos.

- Modo en el que se distribuirán los informes: si los informes se distribuyen siempre de la misma manera es más sencillo que podamos diseñar y construir nuestra herramienta personalizada, pero si los usuarios van a a acceder por diversos canales,

sería recomendable invertir en una herramienta que facilite dichos métodos de distribución de la información.

- Creación de informes ad-hoc: Si es necesario que los usuarios finales puedan crear sus propios informes y también será recomendable utilizar alguno de las soluciones analíticas disponibles. Al construir una solución ad-hoc, la construcción de informes se llevará a cabo normalmente, trabajando a bajo nivel, lo que puede requerir habilidad en el manejo de bases de datos y/o capacidad de programar. Por el contrario, en muchas de las soluciones analíticas existentes, se puede crear un análisis de forma muy similar a como se haría con una tabla dinámica de Excel.

Existen muchas soluciones analíticas en el mercado, tanto de código abierto como de código propietario. Siendo las segundas las más extendidas entre las empresas.

Posiblemente, el "bechmark" más aceptado internacionalmente de herramientas de BI es el que realiza anualmente la consultora especializada en tecnología Gartner. En ese análisis se toman en consideración 15 capacidades diferentes agrupadas en tres grandes categorías: integración, entrega de información y análisis [34]; para acabar clasificando las herramientas según dos características globales de las mismas: capacidad de ejecución, y la completitud de la visión, elaborando así el conocido como Çuadrante mágico de BI y plataformas analíticas"[Figura 4.5].

A la hora de plantearnos cómo es la herramienta analítica que necesitamos, debemos de tener en cuenta que estamos empezando un proyecto que a pesar de tener una gran trayectoria, tiene además un grado de incertidumbre muy elevado.

Sabemos que en un primer queremos empezar diseñando y desarrollando unos pocos informes que permitan dar respuesta a preguntas generales. Además, de momento no queremos dar al usuario (nuestro cliente) la posibilidad de elaborar sus propios informes, ya que según cómo se utilice la información disponible es sencillo que esta induzca a error. Finalmente sólo es necesario dar acceso a las analíticas por vía web.

Del análisis realizado por Gartner se pueden extraer las fortalezas y debilidades de las principales herramientas, aunque si tenemos en cuenta los factores expuestos al inicio de la sección y los requisitos que acabamos de comentar. Y si consideramos por otro lado el elevado coste de la puesta en marcha de la mayoría de las soluciones disponibles, se hace patente la necesidad de afrontar este proyecto con el desarrollo de una herramienta hecha a medida.

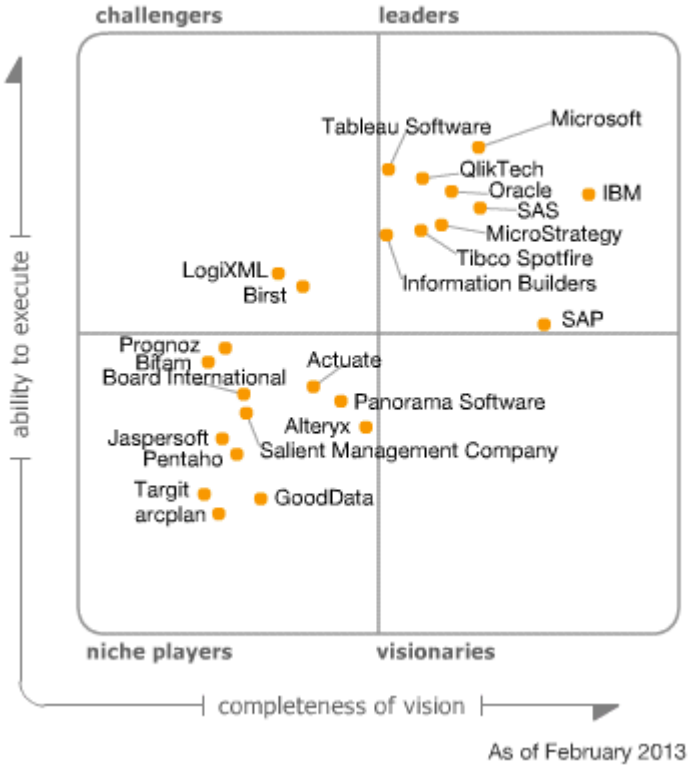


Figura 4.4. Cuadrante mágico de Business Intelligence y plataformas analíticas. Principales plataformas de analíticas según Gartner, y su posicionamiento según la capacidad de ejecución, y la completitud de la visión. Fuente: Gartner. Febrero 2013.

Utilizar una herramienta construida para la ocasión permitirá además llevar el experimento mucho más lejos. No podemos olvidar que el análisis de datos extraídos de medios sociales no es algo maduro y que, por lo tanto, nos conviene estar lo menos atados posible a las limitaciones que nos podamos encontrar si optamos por una tecnología determinada. Utilizando una herramienta ad-hoc tendremos tanta libertad cómo nos permitan las tecnologías web y de bases de datos que utilicemos, es decir, estaremos de inicio, prácticamente exentos de limitaciones.

4.3. Diseño de las analíticas

En el diseño de una herramienta analítica es básico acertar en el diseño del "Back-end", es decir, todo el motor para almacenar, recuperar y trabajar con datos; pero lo es, aún más, el diseño de las propias analíticas, que serán en esencia el "Front-end" de la herramienta.

Aunque dispongamos de muy buenas herramientas de recogida de datos, fracasaremos en el objetivo si no comunicamos lo números de manera efectiva [35]. Así, una fuerte inversión en la recogida de datos, y en el sistema que los almacena, no será efectiva si las analíticas no lo son, y es por eso que una parte importante de este proyecto se dedica al diseño de las analíticas.

Al final de esta sección trataremos de tener definidos un conjunto de informes. Cada uno de estos permitirá "jugar" con la información de determinadas maneras, y servirá para responder una o varias preguntas que nos ayudarán en nuestro objetivo de establecer y controlar la estrategia en medios sociales para un mercado/producto farmacéutico.

Cada una de las analíticas, y las funciones asociadas a la misma (selectores que permiten explorar la información, interactividad, etc.), la podemos llamar un "dashboard" o en español, cuadro de mandos. Una definición más formal de dashboard, la podemos encontrar en [36], en donde afirman: "Un dashboard en Business Intelligence es una visualización de datos que muestra el estado actual de las métricas y KPIs[2] de una compañía. Los Dashboards sirven para ordenar y consolidar cifras con un objetivo concreto".

4.3.1. Características generales de los informes

Para obtener un buen cuadro de mandos, es importante pararse en el diseño de la parte estética por un lado, y de la parte funcional por el otro.

Entre las características funcionales clave de un cuadro de mando podemos enumerar [37]:

- Fácilmente entendible y visualmente efectivo. La información debe poder entenderse rápidamente por sus usuarios.
- Se deben de mostrar KPIs críticos, que sean relevantes para la toma de decisiones efectiva.
- Toda la información presentada debe ser precisa para que el usuario pueda confiar en el cuadro de mandos. Toda la información se debe haber testado y validado previamente.

[2]Key Performance Indicators, o Indicadores Clave de Desempeño, miden el nivel del desempeño de un proceso, de forma que se pueda alcanzar el objetivo fijado.

- Es necesario establecer fronteras para la información representado (máximos, mínimos, valores deseables, etc.) de modo que cuando un valor quede fuera de dichas fronteras se activen alarmas: correos electrónicos, notificaciones, etc.

- Debe ser dinámico en el tiempo y mostrar la información lo más reciente posible de modo que facilite la toma de decisiones.

Otros aspectos deseables en un buen "Dashboard" de datos son:

- Debe ser interactivo y permitir al usuario navegar por la inforamción para obtener mayor detalle, las fuentes, etc.

- Debe permitir al usuario revisar la información histórica, así como las tendencias de un KPI determinado.

- La información debe ser personalizada, de modo que cada usuario tenga acceso solo a la información que le corresponda. Eso implica dotar a la herramienta analítica de permisos que permitan que cada usuario tenga acceso sólo a aquella información que le corresponda.

- Los informes permitirán a los usuarios realizar análisis guiados. El usuario debe poder navegar fácilmente por la información de manera que pueda plantearse hipótesis y extraer conclusiones.

- La información contenida en el cuadro de mandos será, a menudo, compartida entre diferentes personas, por lo que cualquier buena analítica debería de ser colaborativa y facilitar el difusión del contenido.

- El usuario debería de tener la capacidad de realizar seguimientos de la evolución de los diferentes indicadores, de modo que hay que posibilitar la trazabilidad de la información.

Por otro lado, un diseño visual es casi tan importante como el funcional, y es que los usuarios de negocio[3] en general, no están obligados a utilizar los cuadros de mando, y por lo tanto, lo harán solo en caso de que éste sea efectivo y les ayude a aumentar su

[3]Llamamos usuarios de negocio a los que serán usuarios del cuadro de mando y que, a diferencia de los creadores de los mismos o "arquitectos" no tienen que conocer necesariamente la tecnología que hay detrás del Dashboard, o el modo de hacer uno.

productividad. Así, si el informe no es intuitivo y razonablemente sencillo de entender, no será utilizado [38].

Para el diseño y desarrollo de nuestra plataforma vamos a utilizar un enfoque "top-down"[4], empezando por diseñar algunos informes. Cabe añadir, cómo se hará evidente a partir de ahora, que aunque lo primero que se hizo fue diseñar los informes, en esta memoria, las descripciones de los mismos se acompañarán de imágenes de los informes una vez terminados, de modo que podamos ir ilustrando todo lo que se va exponiendo.

Siguiendo las buenas prácticas de diseño para cuadros de mando que se han ido exponiendo, lo primero que debemos de hacer es definir que elementos contendrá un informe, y cómo estos se van a distribuir en la pantalla (layout). Buscamos un patrón para las diferentes analíticas. De no seguir un patrón común, se haría más complicado para el usuario saber cómo utilizarlos, ya que no sería posible reaprovechar conocimiento de un informe a otro. Cada informe sería totalmente nuevo para el usuario.

Los diferentes elementos que en todo momento serán visibles al consultar un análisis son [Figura 4.5]:

1. Cabecera: contiene el logo de la herramienta

    a) Información de sesión: En la parte derecha de la cabecera, se reservará un espacio para mostrar el nombre de usuario y las funciones referentes a la gestión de sesión (en un principio, enlace para logout). Hemos comentado previamente que "La información debe ser personalizada", lo que nos obliga a (entre otras cosas) habilitar mecanismos de autenticación, que además de evitar el acceso a la herramienta por personas no autorizadas, nos permitirá establecer permisos de acceso a la información en función del usuario.

2. Configuración general de las analíticas: justo debajo de la cabecera tendremos una serie de elementos que nos permitirán interactuar con los informes -elementos de formulario- y cuyo valor tendrá vigencia en todos los cuadros de mando a los que tenga acceso el usuario. Podríamos decir que los valores que se establezcan en ese área serán variables globales, mientras que los demás parámetros de configuración

[4]"El enfoque top-down enfatiza la planificación y conocimiento completo del sistema. Se entiende que la codificación no puede comenzar hasta que no se haya alcanzado un nivel de detalle suficiente, al menos en alguna parte del sistema. Esto retrasa las pruebas de las unidades funcionales del sistema hasta que gran parte del diseño se ha completado." Fuente: Wikipedia

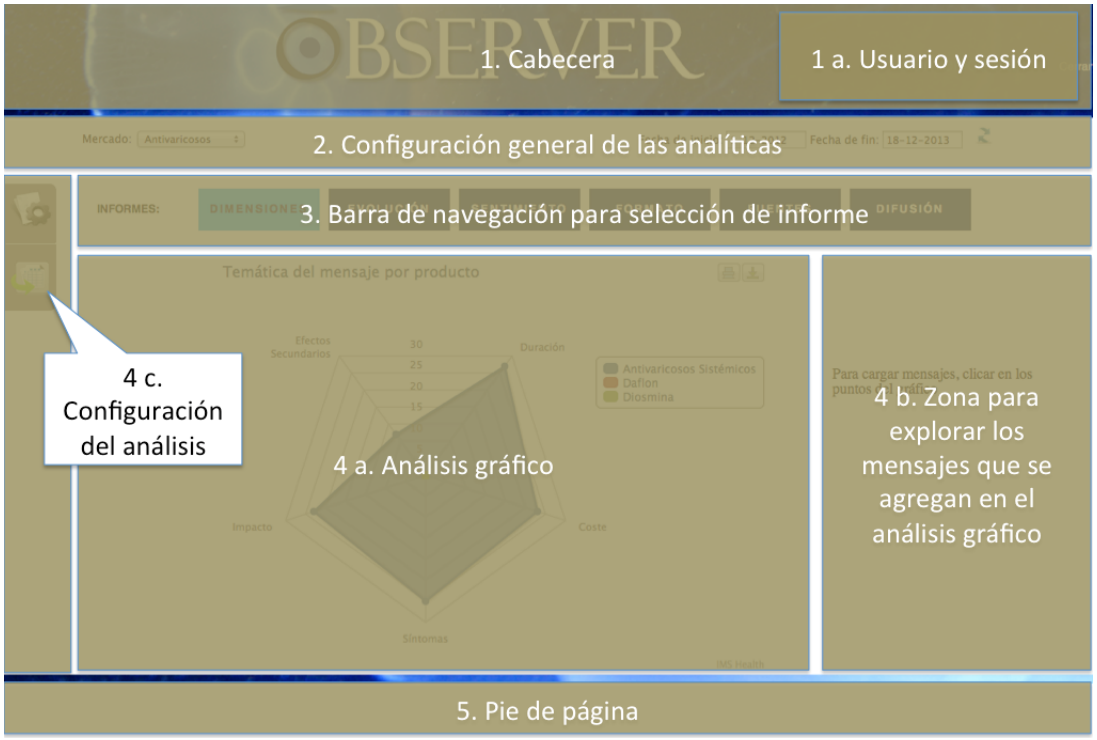


Figura 4.5. Layout del Dashboard. Principales elementos del cuadro de mandos y su disposición en el espacio..

del cuadro de mandos, serán variables locales al mismo. Una variable global es aquella que tiene efecto sobre los informes. De no disponer de variables globales se complicaría la comprensión del mensaje que el conjunto de informes quiere transmitir, será una variable global. Por el contrario, si la variable modifica algún parámetro concreto del cuadro de mandos al que haga referencia, no podrá ser considerada variable global.

Las variables globales que se han identificado inicialmente son:

a) El mercado al que hace referencia el análisis: como hemos dicho en la sección 4.1.3, la información, y también los análisis, se organizan por mercados. El mercado será el conjunto[5] sobre el que se realizarán los informes. Así, cuando un usuario modifique el mercado seleccionado, cualquier analítica que consulte a partir de ese momento hará referencia al último mercado seleccionado.

b) Fecha de inicio y fin del análisis: poder jugar con el horizonte temporal, además de ser una buena práctica en el diseño de cuadros de mandos, es un

[5] Conjunto entendido por su concepto matemático: agrupación de objetos considerada como un objeto en sí

requisito imprescindible cuando hablamos de análisis de social media. Las fechas de inicio y fin que se establezcan en esta sección, serán efectivas en todos informes, de modo que sólo se utilizarán en la elaboración de los mismos los menajes que se hayan generado entre dichas fechas (la fecha asociada a una noticia no es en la que se ha recogido, sino la fecha de la publicación del mismo).

3. Barra de navegación para selección de informe: en las secciones posteriores se diseñará un conjunto de informes. Es necesario poder navegar de un informe a otro. En Excel, por ejemplo, lo haríamos probablemente mediante las pestañas de selección de hojas. Como estamos diseñando una herramienta web, utilizaremos el concepto de barra de navegación, en la que tendremos enlaces a los diferentes informes disponibles. Esto permitirá que se pueda cambiar fácilmente de informe, no forzará al usuario a seguir un orden determinado para explorar la información, y facilitará la adición de nuevas analíticas, que una vez hechas solo necesitarán ser enlazadas desde esta barra de navegación.

4. Contenido del informe: ocupando toda la franja central del informe, tendremos el contenido. Mientras los demás elementos del "layout" deberían de ser idénticos en todos los dashboards, esta sección será en su mayoría variante según el informe de que se trate. A su vez, podemos distinguir dentro del contenido tres bloques diferentes:

   a) Análisis gráfico: es la gráfica elaborada a partir de la transformación (agrupaciones según diferentes criterios, ordenaciones, etc.) de los mensajes recogidos. El gráfico tratará, en cada informe, de transmitir un mensaje o responder a una pregunta determinada.

   b) Zona para explorar los mensajes que se agregan en el gráfico: también hemos visto que es muy importante que un dashboard permita realizar "drill down" de la información representada. En este caso, la forma más básica para acceder al detalle del dato que habilitaremos es la de poder acceder a cada uno de los mensajes que forman cada punto del gráfico pulsando en dicho punto.

   c) En la barra lateral izquierda se situarán botones que nos permitirán realizar transformaciones o cambios sólo sobre el informe que se está mostrando y que por lo tanto no afectarán a los demás cuadros de mandos del análisis. Inicialmente se habilitarán dos botones, el primero abrirá un "pop-up" que

contendrá diferentes elementos de formulario a través de los cuales se podrán modificar, de un modo u otro, las transformaciones realizadas con los mensajes; el segundo botón oculta el gráfico, mostrando en su lugar el dato tabulado.

5. Pie de página: espacio común en cualquier documento (digital o no), y en el que añadiremos la nota de "copyright".

Se han ordenado entonces las diferentes funcionalidades según su ámbito -exceptuando la cabecera y el pie de página- de modo que una metodología de trabajo lógica sería elegir primero el mercado con el que se trabajará, seleccionar luego el análisis, y acabar navegando por el mismo para entender mejor la información que éste transmite.

A continuación se definirán los diferentes cuadros de mandos que se desarrollarán inicialmente, o lo que es lo mismo; el contenido del bloque 4 del "layout" que acabamos de describir.

4.3.2. Informe de "Dimensiones"

En el desarrollo de esta sección y posteriores, será de gran utilidad ponernos en la piel de un "product manager"[6] de un laboratorio farmacéutico, o en el del responsable de la inteligencia de mercado, o de cualquier otra persona encargada de controlar la evolución de un producto, sus competidores, y en general, el mercado al que pertenece. Para que ésta se formule las preguntas correctas, así como para seguir un razonamiento adecuado a la hora de interpretar la información, es necesario que tenga en mente, entre otras cosas, todo lo que se ha expuesto en el capítulo 2.

Tomando todo eso en consideración, una de las primeras cosas que vamos a querer saber es: *¿De qué tratan los mensajes relevantes para mi mercado?¿De qué se habla en los medios sociales cuando se habla de mi producto?¿Y cuándo se habla de los productos de mis competidores?*

[6]"Product manager o jefe de producto es el máximo responsable de la gestión de producto de una organización, y forma parte de las actividades de marketing. Su implicación dura desde la concepción del mismo hasta su desaparición. Gestionará el producto a lo largo de todo su ciclo de vida definiendo en cada momento las estrategias comerciales y de marketing a seguir. También velará por la maximización de los beneficios producidos por el mismo mediante su re lanzamiento en fases de declive o la implementación de otras estrategias encaminadas a prolongar su existencia." Fuente: Wikipedia

En este análisis se clasificarán los mensajes según si hacen referencia a las diferentes dimensiones definidas en la sección 4.1.4. Puede haber mensajes en los que no se encuentren ninguna de las dimensiones definidas y mensajes en los que, por el contrario, se encuentren varias. Eso implica además, que si sumáramos todos los valores que aparecen en la gráfica, el resultado no sería igual al número de mensajes.

Queda claro entonces, que el gráfico no es un indicador del volumen de noticias publicadas. La analítica trata de informar de cuáles son los temas de los que más se habla y que es lo que más preocupa, para bien o para mal, a los diferentes grupos que forman parte del mercado farmacéutico.

Siguiendo el ejemplo encontrado en [15] para realizar un análisis similar, utilizaremos un "gráfico de araña" (ver figura 4.6) con lo que el área encerrada por la linea que representa cada serie es explicativa de para qué temáticas se han recogido más mensajes.

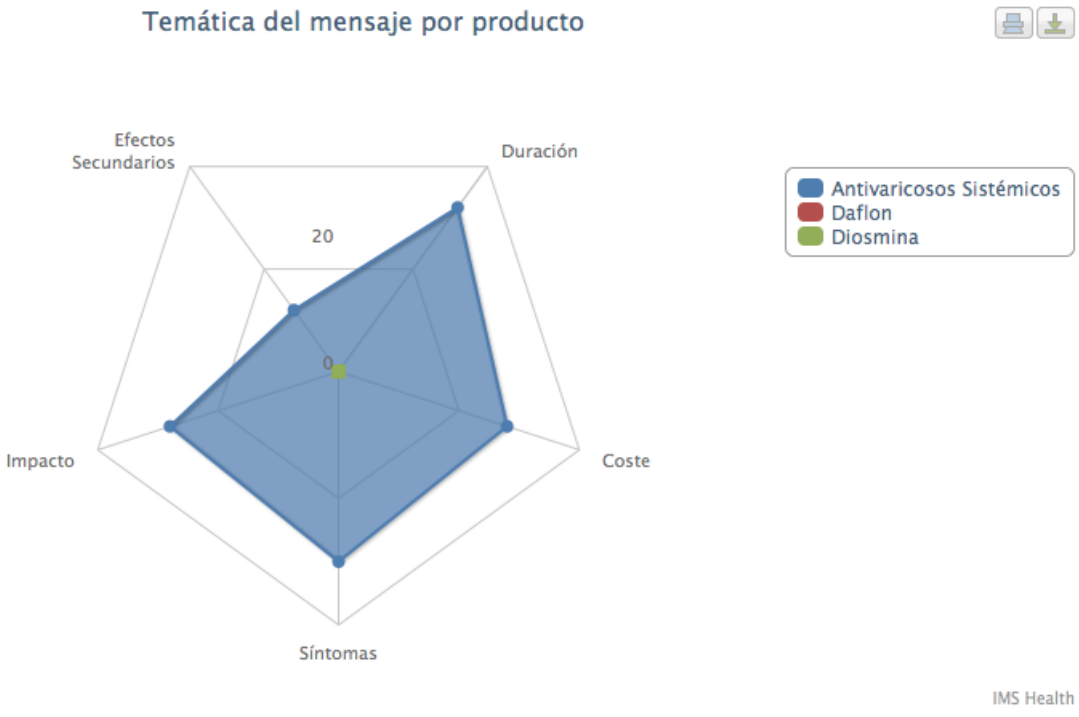


Figura 4.6. Gráfico de dimensiones. Muestra la temática de los mensajes. Cada producto es representado como una serie de datos diferente.

Este tipo de análisis nos permite entender mejor que temas preocupan a los colectivos que utilizan los medios sociales como herramienta de comunicación en el ámbito sanitario. A la hora de analizar los resultados será necesario ver por un lado un producto

determinado y luego compararlo con los demás, de modo que se pueda estudiar el interés que una determinada característica del producto despierta en los medios sociales. Imaginemos por ejemplo un mercado en el que hay un fármaco con efectos similares a sus competidores, pero que requiere una duración de tratamiento que es la mitad. Posiblemente, en ese mercado se generen varios mensajes acerca del tratamiento.

Además, el cuadro de mandos de dimensiones permite, a través de la opción de drill down, explorar las noticias que hablan de un tema en concreto. Pulsar en una de las aristas de los polígonos que forman el gráfico de araña hará que a la derecha del gráfico aparezcan filtradas las noticas para la serie y dimensiones a las que pertenece la arista pulsada. Los mensajes se filtrarán además de modo que sólo aparezcan los publicados entre la fecha de origen y fin (o los recogidos sin fecha, aunque éstos no se utilizan en la elaboración del gráfico), y lógicamente los que pertenezcan al mercado sobre el que se está realizando el análisis.

Además, como se ha explicado, cada informe dispondrá de varios parámetros de configuración adicionales que serán editables a través del botón correspondiente en la barra lateral izquierda de la pantalla. En este caso, al pulsar dicho botón aparecerá un desplegable como el de la figura 4.7

En el informe de de dimensiones, los parámetros de configuración son:

- Agrupar Keywords: para realizar una búsqueda completa de mensajes se puede llegar a utilizar una gran cantidad de palabras clave para conducir la búsqueda. En este informe cada palabra clave utilizada (en general serán el nombre del mercado, y los productos que forman parte del mismo) se representará como una serie diferente, es decir, cómo un polígono en el gráfico de araña. Si tenemos demasiadas palabras clave, será más complicado interpretar el gráfico. Todo esto hace que tengamos que limitar el número de palabras clave que pueden aparecer. Cómo solución de compromiso se han seleccionado un máximo de 10, quedando las demás agrupadas en “otras”. Esta primera opción permite modificar el número de palabras clave mostradas en el informe, desde 1 hasta 10. Así, si en el selector ”fijamos agrupar 3 principales y otras”, tomará considerando todos los demás parámetros del informe las 3 series (palabras claves) con más mensajes, y los mostrará junto a una cuarta, agrupando los mensajes restantes.

- Ver dimensiones como: hasta ahora hemos asumido que las palabras clave son

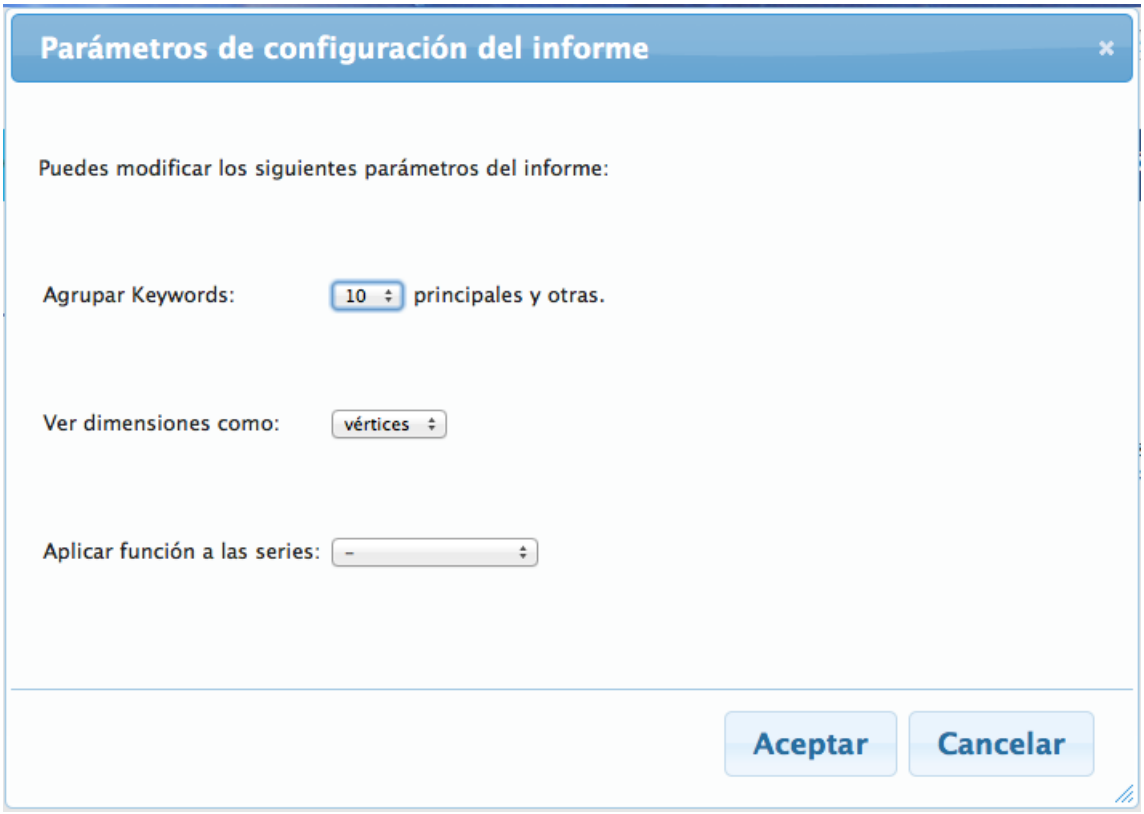
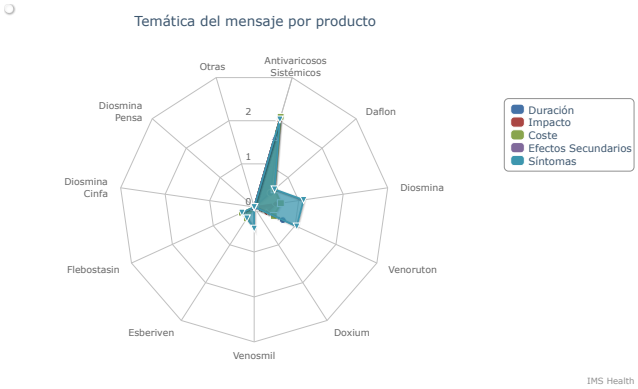


Figura 4.7. Parámetros locales al informe de dimensiones. Pop-up con los diferentes parámetros de configuración locales al informe de dimensiones. El mismo pop-up, con diferente contenido estará disponible para cada uno de los informes.

las series del gráfico, y que las dimensiones son los vértices, ya que éste es el razonamiento seguido en el diseño del informe. Pero al ver el gráfico representado nos planteamos también las preguntas del principio de la sección, pero al revés: *En una temática determinada ¿para que producto se publican más mensajes?* Por lo que nos pareció interesante invertir los ejes, y poner las dimensiones como series, situando las palabras clave en las aristas.

- Aplicar función a las series: habitualmente, al observar los resultados de la recogida de mensajes, podemos observar que algunas palabras clave conducen a muchos más resultados que otras. Es por eso que en ocasiones se hace muy complicado comparar dos series superpuestas en la misma gráfica. Esta problemática nos llevó a incluir una opción para aplicar una transformación al eje de la gráfica, es decir, aplicar una función a los resultados de la agrupación, de modo que el resultado obtenido



	Duración	Impacto	Coste	Efectos Secundarios	Síntomas
Antivaricosos Sistémicos	2.01	2	2.09	1.62	2.03
Daflon	0.48	0	0.6	0	0.6
Diosmina	0	0.48	0.6	0	1.11
Venoruton	0.7	0.3	0.48	0	1.04
Doxium	0	0	0	0	0
Venosmil	0	0	0	0	0.48
Esberiven	0	0	0.3	0	0.3
Flebostasin	0	0	0.3	0	0.3
Diosmina Cinfa	0	0	0	0	0
Diosmina Pensa	0	0	0	0	0
Otras	0	0	0	0	0

Figura 4.8. Gráfico de dimensiones en unidades logarítmicas y con los ejes cambiados. A la izquierda, un ejemplo del gráfico de dimensiones con los ejes cambiados, siendo las dimensiones las aristas, y aplicando la función logaritmo en base diez a las series, a la derecha la tabla de datos mostrando los valores que dan forma al gráfico.

fuese más sencillo de obtener. Aunque está función se prevé ampliar en el futuro, se han habilitado dos opciones:

1. Unidades naturales: el eje del gráfico muestra el número de mensajes que cumplen los criterios establecidos por los filtros del informe (mercado, fecha inicio y fin), dimensión y palabra clave correspondiente.
2. logaritmo base 10: se aplica el logaritmo en base 10 al resultado del apartado anterior más 1 para obtener resultados en el rango $[0, \infty]$. Por otro lado, por ahora, no se ha aplicado la escala logarítmica al eje, por lo que el resultado graficado es directamente el obtenido de aplicar el logaritmo a la función.

4.3.3. Informe de "Evolución"

El informe anterior agrupa los mensajes que quedan comprendidos entre la fecha de inicio y la de fin, y la única manera que tenemos de conocer cómo es la evolución temporal de los mensajes publicados es jugando con esos parámetros. Por otro lado, es muy habitual en cualquier solución analítica configurar informes que permitan conocer la evolución en el tiempo de una variable. Esto se hace normalmente mediante una gráfica de lineas o de barras, utilizando el tiempo como el eje de abscisas y la magnitud como el de ordenadas.

Los gráficos de líneas se suelen utilizar para mostrar tendencias o evoluciones más que para representar datos exactos. Es recomendable utilizarlos solo cuando el eje horizontal

corresponde al tiempo [39]. Si recordamos el modo en el que solemos representar las señales, utilizamos barras para representar señales discretas, y lineas para representar señales continuas. En este caso, estamos trabajando con una señal discreta, ya que el número de noticias recogidas será siempre natural, y la granularidad disponible por ahora es diaria. Por otro lado, el análisis no pretende ofrecer esa información, tanto como describir rápidamente la evolución del volumen de mensajes publicados para cada una de las palabras clave.

Entre las principales funciones de este informe tenemos las de monitorizar los resultados en la red ante una campaña determinada, entender la presencia que tiene un producto/mercado respecto a los demás en la red, etc.

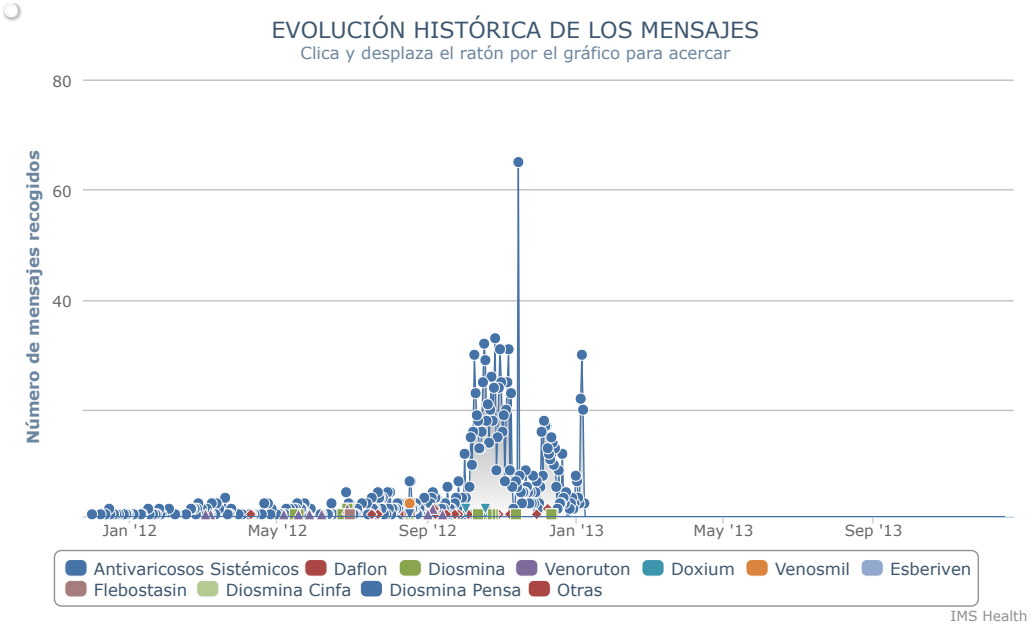


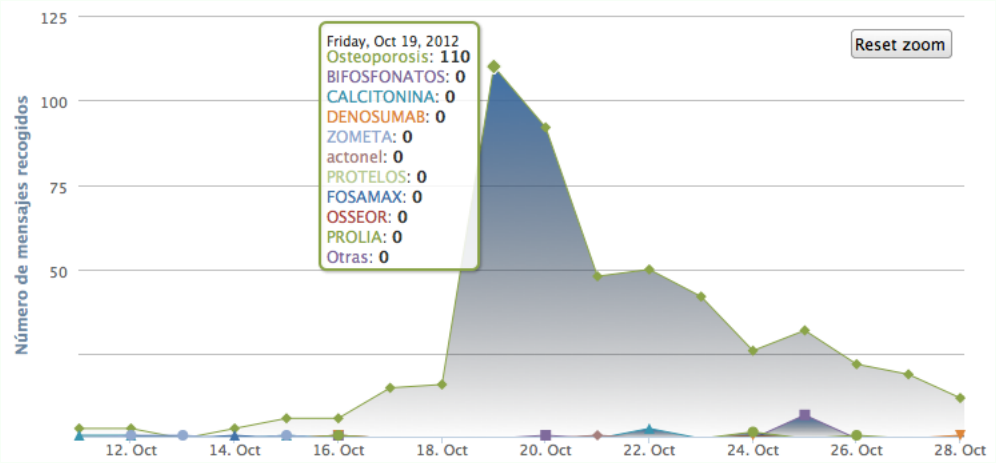
Figura 4.9. Informe de evolución. Ejemplo de un informe de evolución para el mercado de Antivaricos.

Caso de uso: seguimiento del día mundial de la osteoporosis

Una de las primeras pruebas llevadas a cabo con la herramienta analítica fue el seguimiento del día mundial de la osteoporosis, que es el 20 de octubre.

Como era de esperar, el volumen de datos recogidos acerca de la osteoporosis fue mucho mayor de lo normal el día dedicado a difundir la patología y concienciar acerca de la misma, aunque el pico de mensajes publicados fue el día anterior.

Otro efecto algo más sorprendente, es el modo en el que se mantiene un ritmo de publicación de mensajes superior al normal durante los días posteriores a la jornada mundial por la osteoporosis. Esto se puede deber a varios motivos; pero pone en evidencia que dedicar un día a la patología sirve para que realmente la gente hable más de la misma y, por lo tanto, puede ayudar a concienciar, justificando así el esfuerzo invertido



Evolución de los mensajes del mercado de osteoporosis los días anteriores y posteriores al día mundial de la osteoporosis

4.3.4. Informe de "Sentimiento"

Se ha explicado que los mensajes se clasifican según su contenido objetivo (dimensiones), o según su contenido subjetivo (sentimiento). Este informe representa gráficamente esa segunda clasificación.

En el análisis de los mensajes, se han establecido tres categorías:

- Positivo: el autor del mensaje/noticia está tratando de transmitir algo bueno: una experiencia positiva, un mensaje optimista acerca de algo, agradecimiento, etc.

  En general, los mensajes transmitirán buenas experiencias alrededor de la palabra

clave en la que se haya clasificado el mensaje y, por lo tanto, podemos identificar una proporción alta de mensajes positivos con una buena reputación de la palabra clave. Dado que el verde se utiliza habitualmente para representar buenas condiciones, se ha elegido para representar los mensajes positivos

- Negativo: el autor del mensaje/noticia está tratando de transmitir algo malo: una experiencia negativa, un mensaje pesimista acerca de algo, enfado, etc.

  Los mensajes clasificados con sentimiento negativo denotarán por el contrario mala reputación en la red. Se ha elegido el rojo para representar estos mensajes, ya es un color utilizado en ocasiones como antagonista al verde (por ejemplo en los semáforos)

- Neutro: algunos de los mensajes publicados, no transmiten ni un mensaje positivo ni negativo (i.e. estoy tomando Aspirina), sino que hacen pública una información. Para ese tipo de mensajes es necesario definir la categoría de neutra. Buscando un color que contraste con los dos anteriores, se ha optado por el azul

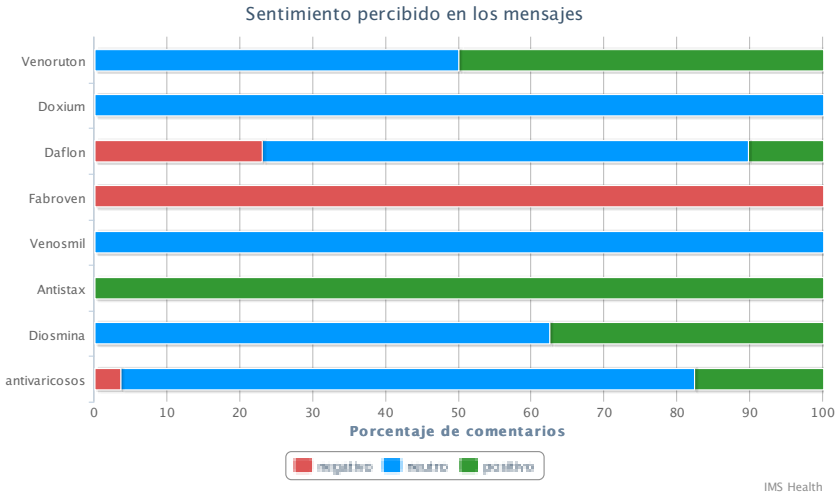


Figura 4.10. Informe de sentimiento. Muestra los mensajes clasificados según si el contenido expresa un sentimiento positivo, neutro o negativo.

Para representar el informe de sentimiento se ha optado por una gráfica de barras horizontales apiladas que representan la proporción de mensajes recogidos para cada una de las tres categorías anteriores [Figura 4.10].

Utilizamos este tipo de representación ya que en este análisis, no nos importa tanto el volumen de mensajes publicados, que ya se ha presentado en análisis anteriores, sino

la reputación de cada una de las palabras clave. Así, el grupo representando los mensajes positivos queda a la derecha en verde, el grupo representando los negativos a la izquierda en rojo, quedando en el centro los mensajes neutros en azul. Eso nos permite entender muy rápidamente cual es la tendencia para cada palabra clave.

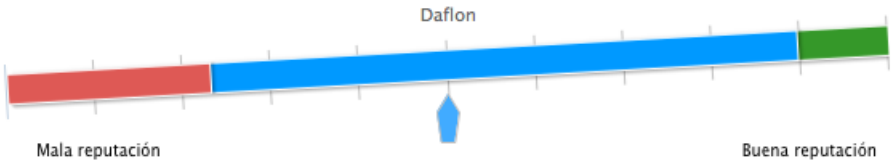


Figura 4.11. Interpretación del informe de sentimiento. Vista del informe de sentimiento cómo una balanza que indica la reputación de la palabra clave en la red.

Como muestra la figura 4.11, se puede asociar cada barra con una balanza. Si la barra roja es mayor que la verde, significa que hay un peso mayor de mensajes negativos, y por lo tanto la balanza se decanta hacia el lado de la mala reputación.

Si al monitorizar un producto propio vemos que la reputación en medios sociales no es buena, se requeriría establecer una estrategia para revertir la situación. Mediante este informe se podrían monitorizar los efectos de la campaña, cuyo objetivo debería de ser que el peso del grupo rojo fuese disminuyendo, a medida que el peso del grupo verde va en aumento.

4.3.5. Informe de "Formato"

Cuando hablamos de formato, nos referimos a un criterio de clasificación de las fuentes utilizadas en la recogida de mensajes.

Al presentar el concepto de medios sociales, vimos que estos hacían referencia a diferentes tipos de servicio web: foros, blogs, microbloging, etc. Cada tipo de servicio web será en este análisis un formato diferente.

Este análisis es necesario dado que cada formato tendrá un peso diferente en la interpretación de resultados. No es lo mismo un post publicado en un blog especializado de medicina seguido por cientos o miles de médicos, que un Tweet publicado por un usuario con poca influencia en el entorno médico.

Las fuentes se han clasificado en los siguientes formatos[7]:

- Microblogging: servicio que permite a sus usuarios enviar y publicar mensajes breves generalmente sólo de texto. Ej: Twitter, Identi.ca, Tumblelog, etc.
- Blog: servicio web periódicamente actualizado que recopila cronológicamente textos o artículos de uno o varios autores, donde el autor conserva siempre la libertad de dejar publicado lo que crea pertinente.
- Noticias: servicios en los que se publican textos informativos de forma profesionalizada. Ej: Periódicos y revistas on-line, publicaciones científicas, etc.
- Foro: aplicación web que da soporte a discusiones u opiniones en línea.
- Red Social: medio de comunicación social que se centra en encontrar gente para relacionarse en línea. Están formadas por personas que comparten alguna relación, principalmente de amistad, mantienen intereses y actividades en común, o están interesados en explorar los intereses y las actividades de otros. Ej: Facebook, Tuenti, redpacientes.com, etc.
- Web: Reservado para fuentes que no encajan en ninguna de las definiciones anteriores.

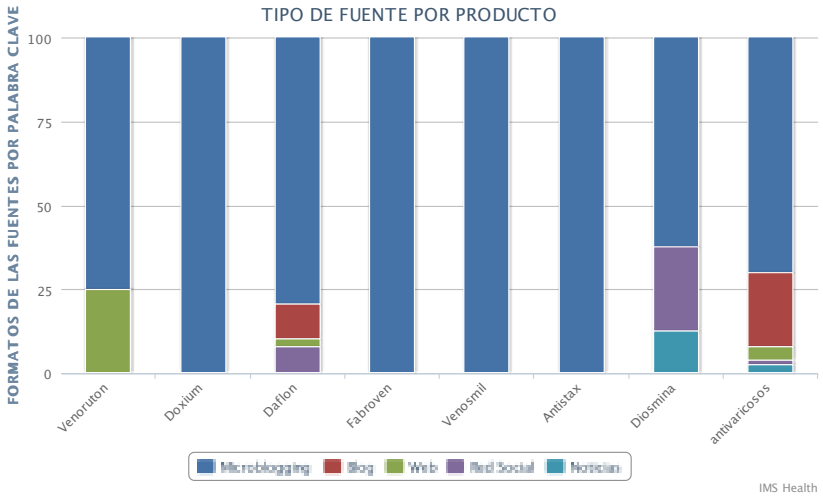


Figura 4.12. Informe de Formato. Muestra los mensajes clasificados según el tipo de fuente de la que han sido extraídos.

[7]Definiciones cortas adaptadas de Wikipedia

Para realizar el análisis se ha elegido un gráfico de columnas apiladas y como en el caso anterior, normalizadas al 100 %. Con este informe es posible identificar cuáles son los medios más utilizados en la divulgación de mensajes relacionados con un mercado.

Se ha elegido este tipo de gráfica ya que, de nuevo, no importa tanto la cantidad absoluta de mensajes recogidos como al proporción de éstos que se ha recogido en cada medio. Utilizar barras apiladas y normalizadas a 100 % permite comparar diferentes palabras claves para las que se pueden haber recogido una cantidad muy diferente de mensajes. La barras son esta vez horizontales, ya que en este caso no estamos tan interesados en entender los pesos, como en representar diferencias relativas entre palabras clave.

Es necesario, a la hora de afrontar una estrategia de posicionamiento en medios sociales decidir cómo se va a afrontar la presencia en cada tipo de medio. Por ejemplo, si de un producto se habla mucho en noticias, es probable que al laboratorio que lo produce le interese publicar algún artículo en ese formato. Si por el contrario, hay muchos usuarios publicando micronoticias, será útil entender que se dice en ese medio para reforzar que la visión predominante del producto se ajuste a las propiedades reales del mismo.

4.3.6. Informe de "Fuentes"

El informe de fuentes es similar en contenido y fisionomía al anterior, aunque en este caso el nivel de detalle es superior.

Mediante un gráfico de barras apiladas y normalizadas al 100 % muestra para cada palabra clave, las fuentes en las que se han recogido más mensajes.

Aunque algunas fuentes (en especial Twitter) aparecen constantemente cómo las que mayor contenido aportan, en ocasiones es necesario centrarse en el contenido de fuentes específicas, generalmente porque son más especializadas, o porque van dirigidas a un público específico que es de especial interés para la farmacéutica.

El informe de fuentes permite además, en el momento de planificar una campaña, entender que está sucediendo en una fuente antes de que realizar cualquier acción sobre la misma.

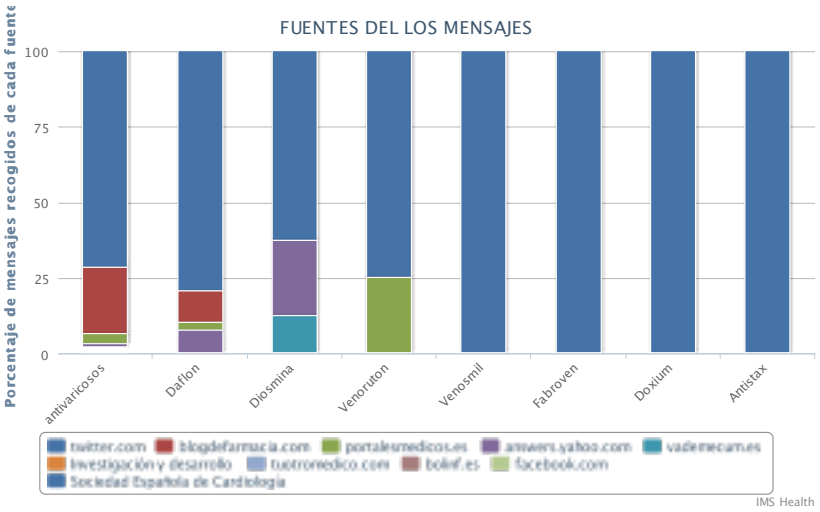


Figura 4.13. Informe de Fuentes. Muestra los mensajes clasificados la fuente de la que han sido extraídos.

4.3.7. Informe de "Difusión"

Los informes vistos hasta ahora se basaban en agregaciones directas de los mensajes según alguno de sus atributos. Estos análisis son muy útiles, sobre todo para entender que está pasando para cada una de las palabras clave.

En este caso se ha diseñado un informe orientado a comparar la presencia en medios sociales de las diferentes palabras clave dentro del mismo mercado.

En el informe de difusión se representarán conjuntamente tres variables calculadas sobre cada una de las palabras clave:

- Interés: los comentarios generados a partir de la publicación de una noticia, son un buen indicador del interés que suscita esa noticia y es que si se publican muchos mensajes, pero nadie los comenta, es que el tema no interesa demasiado. El interés de una palabra clave se calcula como[8] :

```
for k in keywords:
    if num_mensajes[k] == 0:
        interes[k] = 0
    else:
```

[8]Aunque aún no se ha presentado la plataforma que se utilizará ni el modelo de datos, utilizamos código python para explicar más fácilmente el modo en el que calculamos los diferentes indicadores

```
        interes[k] = \
            (num_mensajes_con_comentarios[k] / \
            float(num_mensajes[k])) * 100
```

Para que el volumen de mensajes recogidos no tenga efecto en el cálculo del interés, tenemos el número de mensajes en el demoninador, con lo que podríamos definir matemáticamente el interés cómo el porcentaje de mensajes recogidos para los que se ha publicado al menos un comentario.

- Alcance: la medida de alcance se ha definido de modo que permita medir la difusión de la noticia, sea cual sea el modo en el que esta se ha difundido. El cálculo del alcance sería:

```
for k in keywords:
    %contamos las veces que cualquier mensaje entre la fecha
    % de inicio yfin, se ha comentado, compartido en twitter,
    % retwiteado o compartido en facebook, siendo estas las diferentes
    % formas [monitorizadas] con las que medimos el alcance
    shared[k] = Message.objects.filter(content__date__range = dRange,
                message_keyword__keyword__name = k).aggregate(
                count_comments = Sum('social__comments'),
                count_twShared = Sum('social__twitterShared'),
                count_twRetweet = Sum('social__twitterRetweet'),
                count_fbShared = Sum('social__facebookShared'))
    % El alcance es la suma diferentes formas calculadas
    % anteriormente, dividido por el número de mensajes con
    % comentarios
    alcance[k] = (shared['count_comments'] +
                shared['count_twShared'] +
                shared['count_twRetweet'] +
                shared['count_fbShared']) / commentedMessages
```

En el algoritmo anterior dividimos por el número de mensajes con comentarios ya que eso queda recogido en la medida de interés.

- Mensajes publicados (Volumen): Cantidad de mensajes publicados entre la fecha

de inicio y fin para cada una de las palabras clave. Se podría representar matemáticamente como el sumario de la serie correspondiente en el informe de evolución.

Los gráficos que hemos visto hasta ahora, nos permiten representar para cada palabra clave un máximo de dos variables a la vez. (i.e. fecha y cantidad, formato de fuente y porcentaje de mensajes, etc.). En este caso queremos representar tres variables para cada palabra clave en un solo gráfico.

Los gráficos de burbujas nos permiten representar esas tres dimensiones de datos. En la posición x, representaremos el interés, en la posición y, el alcance, y el tamaño de la burbuja representará el número de mensajes recogidos.

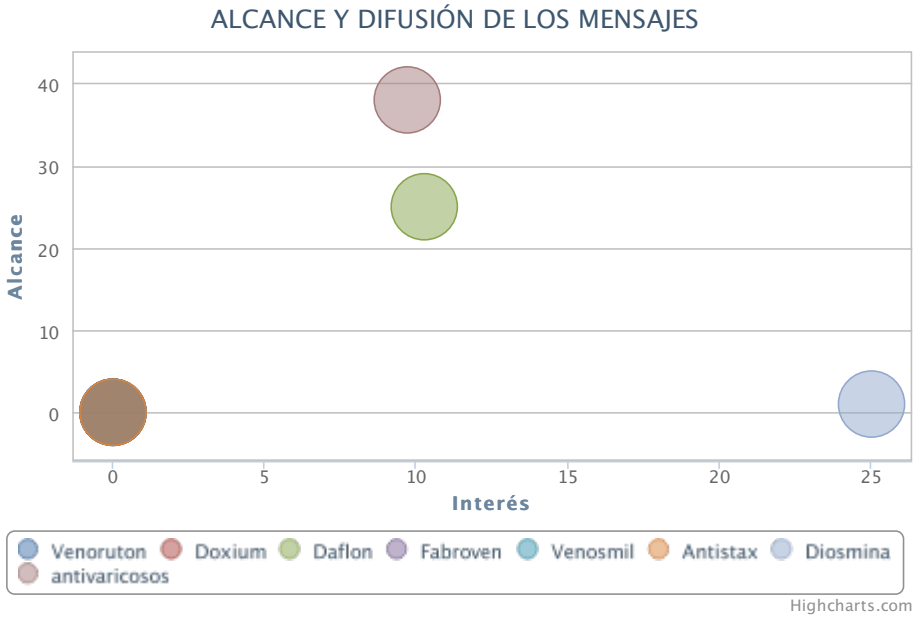


Figura 4.14. Informe de Difusion. Muestra la difusión de las diferentes palabras clave según tres criterios: número de mensajes, alcance e interés.

Capítulo 5

Diseño y Arquitectura del servicio

La arquitectura de un servicio web debería definir dicho servicio y el encaje del mismo dentro de sistemas más grandes. Provee un modelo conceptual y el contexto para entender el servicio web y las interrelaciones entre los componentes del mismo.

Por el contrario la arquitectura no es una descripción del modo en el que se implementará el servicio. Tampoco impone restricciones en el modo en el que los servicios web se podrán combinar.

5.1. Diseño del servidor

En el servidor se implementarán los procesos que sirven el servicio así como toda la lógica de negocio. Al recibir una petición, el servidor será el encargado de generar el análisis y los informes pertinentes.

Haciendo uso de las tecnologías presentadas en el capítulo 3 podemos representar esquemáticamente el servidor de la siguiente manera: el servidor web es el que recibe las peticiones HTTP. Si la petición consiste en un fichero estático, como una imagen o una hoja de estilo, ésta es servida directamente por Lighttpd. Si por el contrario, la petición requiere la ejecución de un script, se pasará el control Django, para que desde allí se llame a la función correspondiente (según lo configurado en el fichero urls.py dentro fichero del proyecto Django). Si se espera que la función requiera de un tiempo largo de proceso (del orden de minutos), se pasará el control de la misma a Celery, al que se ha configurado para utilizar la base de datos como broker de mensajes. Si el tiempo de

procesado de la petición es corto (del orden de milisegundos o pocos segundos), Django gestionará la ejecución del proceso, realizando si es necesario una o varias consultas a la base de datos, y devolverá la respuesta al servidor web, para que éste la haga llegar al cliente que inició la conexión [Ver figura 5.1].

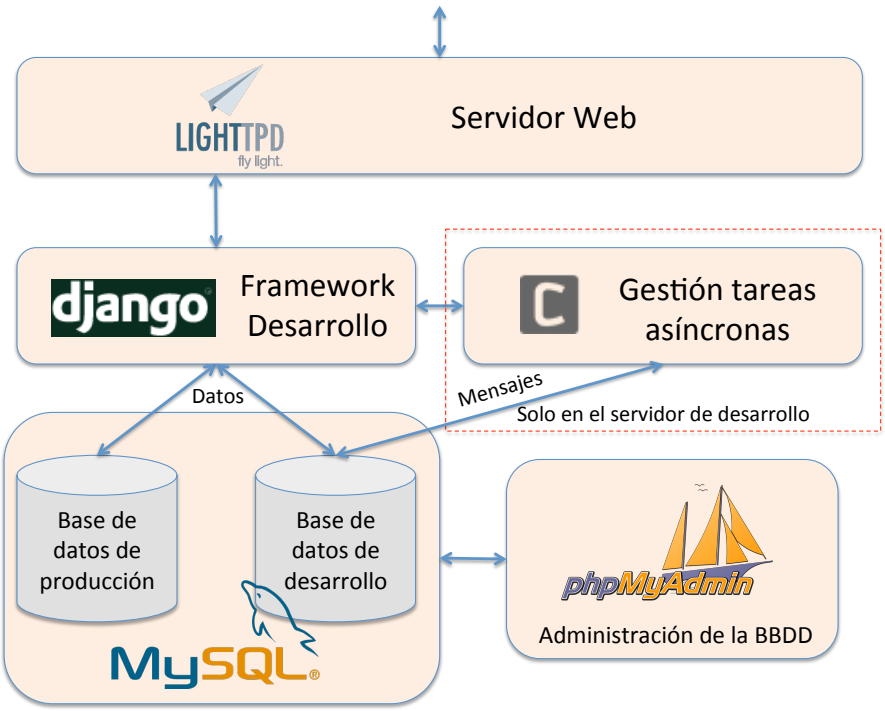


Figura 5.1. Esquema general del servidor. Diferentes elementos que forman parte del servidor e interrelación entre ellos.

Es una buena práctica en business intelligence, y en el desarrollo web en general, el disponer por el lado del servidor, de un mínimo de dos entornos independientes, o cuasi independientes [40]. Uno de ellos, el servidor de producción, es aquél al que acceden los usuarios finales. La plataforma servida desde ese entorno se presupone suficientemente testado y estable. El otro entorno, al que podemos llamar entorno de desarrollo, es por otro lado en donde se mantiene la versión de la plataforma que está en desarrollo, o que por cualquier razón no debe de ser accedido por el cliente directamente.

Los entornos de desarrollo y producción pueden funcionar en la misma o en diferentes máquinas y compartir o no la base de datos. En algunos casos, los requerimientos de recursos de uno y otro entorno pueden diferir. Por ejemplo, un servicio que tiene muchos usuarios ejecutando consultas a una base de datos, requerirá más recursos que uno en el que solo los desarrolladores están ejecutando consultas. Es por eso, que en ocasiones los

dos entornos se encuentran en máquinas diferentes, siendo una de ellas más potente que la otra.

En nuestro caso establecer un entorno de desarrollo era una necesidad imperiosa. No sólo para poder avanzar el desarrollo sin correr el riesgo de que los clientes se quedasen sin servicio, sino también para poder controlar, cada vez que se lleva a cabo una carga de datos, que la calidad de los mismos es adecuada, y que éstos pueden hacerse accesibles a los clientes.

La figura 5.2 representa la arquitectura que se ha implementado en nuestra plataforma teniendo en cuenta esa dualidad de entornos. Nosotros, por el momento, utilizamos los dos entornos en la misma máquina. Si unimos la figura 5.1 con la figura 5.2 tenemos que en nuestra configuración ambas aplicaciones (producción y desarrollo) comparten el mismo servidor web pero son dos aplicaciones Django separadas que en general utilizan Bases de Datos diferentes pero alojadas en el mismo servidor de BBDD.

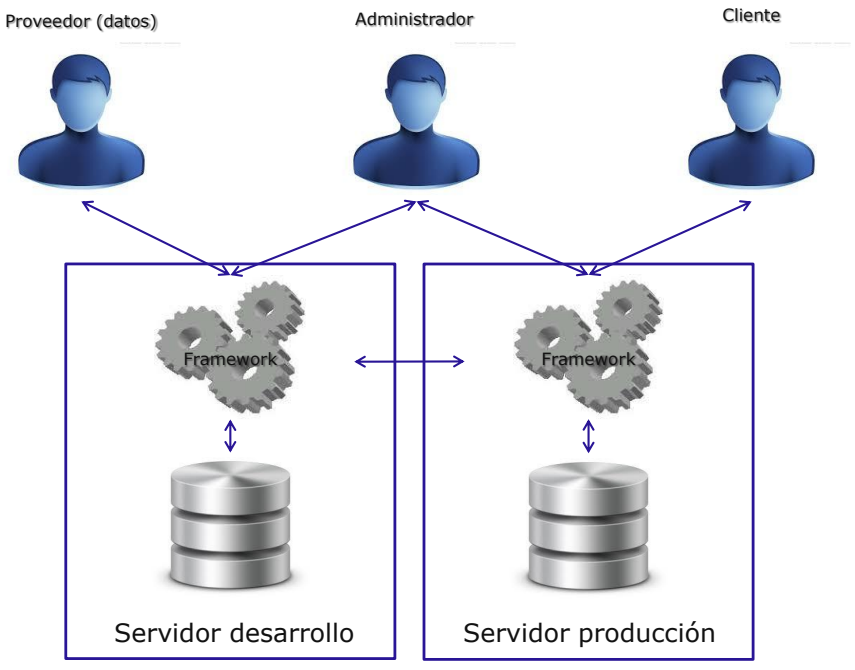


Figura 5.2. Estructura de la aplicación. Fuente: elaboración propia.

Otro aspecto importante en la definición del servidor es la identificación de diferentes perfiles de usuario, lo que también se ha reflejado en la figura 5.2. Estos tipos de usuario se han definido teniendo en cuenta las diferentes opciones a las que debe de tener acceso un usuario según la razón por la que se esté conectando a la plataforma.

- Administrador: los usuarios de tipo administrador tienen acceso a toda la funcionalidad de la plataforma. El admisistrador es el que velará por el correcto funcionamiento de la herraienta y realizará la gestión de datos contenidos en ésta.

- Proveedor: los proveedores de datos necesitan poder acceder a la herramienta para cargar periódicamente la información recogida. El proveedor puede subir los datos y actualizar el servidor de staging, aunque no puede pasar esos datos a producción, acción que sólo puede ser llevada a cabo por un administrador, asegurando así una segunda revisión de los datos.

- Cliente: usuarios que pueden solamente trabajar con las analíticas, limitadas además a aquellos mercados que hayan contratado.

5.1.1. Entorno de desarrollo

El entorno de desarrollo es la parte del servicio en la que invierte más tiempo el administrador. Las funciones principales de este entorno son:

- Ejecutar la ETL, para cargar la fuente de datos y transformarla de modo que encaje en la base de datos [sección 5.3.2].

- Analizar los datos cargados en un entorno igual al de producción antes de que los datos pasen sean accesibles al cliente. Es por eso que también llamamos a este entorno *staging area*.

- Probar nuevos desarrollos: nueva funcionalidad, informes adicionales, etc. Aunque lo ideal sería tener un tercer entorno para llevar a cabo este tipo de tareas, dado que nos encontramos en una fase inicial del despliegue del servicio, disponer sólo de dos aplicaciones facilitará la gestión de la herramienta.

5.1.2. Entorno de producción

Entorno de producción hace referencia al entorno desde el que se sirve el contenido que está listo para la producción, es decir, es el entorno abierto en internet que será accesible a los clientes [41].

En este proyecto estamos diseñando un sistema de análisis de datos recogidos en los medios sociales. El proceso hasta llegar a las analíticas es largo, siendo el servidor

de producción el que se ocupa de la parte más cercana al usuario: elaborar y servir las analíticas.

La base de datos que alimenta a este entorno no se modifica directamente, sino que se actualiza a partir de la base de datos del servidor de producción, donde los datos se transformarán y validarán. Todo este proceso se lleva a cabo desde el entorno de desarrollo.

El entorno de producción es el que debe de servir con rapidez las analíticas que permiten explorar la información disponible en la base de datos. Es por eso que se tratará de mantener la máxima estabilidad posible sobre este servidor. Los cambios, tanto de diseño como de contenido, deben ser accesibles en este servidor una vez se hayan probado bien, y se hayan validado.

Por las demás cuestiones, este entorno es una copia del entorno de desarrollo.

5.1.3. Conexión servidor desarrollo y servidor de producción

Hasta ahora, se han considerado ambas aplicaciones cómo servicios independientes, pero dado que una de las funciones encargadas al entorno de desarrollo es la ejecución de la ETL. Es necesario habilitar un modo de copiar la información de un entorno a otro. Para hacerlo posible, se ha dado acceso al proyecto de desarrollo a la base de datos de producción, además de a la suya propia.

Django permite que un proyecto utilice más de una base de datos configurándolo a través del fichero de *settings.py*, en el apartado DATABASES:

```
DATABASES = {
    'default': {
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'imsspain_stagingobserver',
        'USER': 'imsspain',
        'PASSWORD': '****',
        'HOST': '',
        'PORT': '',
    },
    'live': {
```

```
        'ENGINE': 'django.db.backends.mysql',
        'NAME': 'imsspain_observer',
        'USER': 'imsspain',
        'PASSWORD': '****',
        'HOST': '',
        'PORT': '',
    }
}
```

Es posible dar a cada base de datos el alias (siendo este la clave en el diccionario de configuración) que se desee, pero el alias *default* tiene una importancia significativa, ya que será el que se utilizará por defecto cuando no se especifica una de las bases de datos:

```
#Filtra en la base de datos default
Message.objects.filter(author = a).count()

#Filtra también en la base de datos default
Message.objects.using('default').filter(author = a).count()

#Filtra en la base de datos live
Message.objects.using('live').filter(author = a).count()
```

5.2. Diseño del cliente

El cliente es básicamente la página web que ve el cliente y la interfaz con la que éste interactúa.

Los requisitos de diseño del cliente pasan por la necesidad de desarrollar todas las analíticas y funcionalidades que se han ido exponiendo a lo largo del capítulo 4. Por otro lado hay que tener en cuenta un conjunto de principios válidos para el desarrollo de interfaces de usuario en general [42]:

- Rendimiento: procesado de una gran cantidad de datos. Hay que garantizar que el usuario no se ve sometido a largos periodos de espera mientras las analíticas se están generando
- Accesibilidad: la página, todo su contenido y funcionalidad deben ser fácilmente

accesibles para los usuarios.

- Compatibilidad: las analíticas deben de poder consultarse desde cualquier dispositivo mayoritario moderno.

El cumplimiento de todos los requisitos se ha procurado desde el diseño del modelo de datos, hasta la búsqueda de un servidor con capacidad suficiente para que la aplicación funcionase correctamente.

Centrándonos en la parte de diseño del cliente, para mejorar la experiencia de usuario, las gráficas se generan todas de forma asíncrona. A la petición de una analítica el servidor responde con el contenido del dashboard que no requiere procesado de datos: ficheros estáticos, HTML, etc. Una vez el navegador cliente ha cargado esa parte pide mediante AJAX los datos necesarios para crear el gráfico. Éstos son devueltos con JSON.

La ventaja de cargar las páginas de esta manera es que el usuario tiene la sensación de que el informe carga rápidamente. La primera petición realizada es servida por el servidor muy rápidamente ya que el proceso que lo hace no requiere consultas pesadas a la base de datos. Cuando eso sucede el usuario ve cómo el navegador carga el dashboard. Mientras se procesan los datos que dan forma al gráfico, la página queda sombreada mediante un pop-up modal. Al recibir el JSON con la fuente de los gráficos, una función de JavaScript actualiza el gráfico utilizando la librería Highcharts y cierra el pop-up modal de modo que el dashboard queda habilitado.

5.3. Gestión de Datos

El hecho de que estemos trabajando en la puesta en marcha de una herramienta de Business intelligence hace que el análisis de los datos que se representarán, así como el diseño del modelo de datos sea fundamental.

Un mal diseño de datos podría tener consecuencias muy negativas en el uso posterior de la herramienta como un tiempo de respuesta demasiado elevado o la incapacidad de escalar la plataforma (más mercados, noticias y/o más usuarios).

5.3.1. Modelo de datos

Modelado de datos en la ingeniería del software consiste en la creación de un modelo de datos para un sistema informático mediante la aplicación de técnicas formales de

modelado de datos.

El proceso incluye la definición y análisis los requerimientos de información necesarios para dar soporte a los procesos de negocio a los que el sistema informático dará soporte.

A lo largo del modelado, se construyen tres modelos de datos diferentes:

1. Modelo de datos conceptual (CDM): es el primer modelo que se construye, y también el más cercano a toda la lógica de negocio soportada por la herramienta informática que va a hacer uso del modelo de datos. Consiste en un mapa de conceptos interrelacionados representando el significado de los procesos a los que se dará soporte desde la base de datos. El modelo de datos conceptual es independiente de la tecnología de base de datos que se utilizará.

2. Modelo de datos lógico (LDM): documenta la estructura de datos que se puede implementar en la base de datos. Se basa en las estructuras definidas por el modelo de datos conceptual, y representa la estructura de la información de forma abstracta. La elaboración de un LDM en el proceso del modelado de datos ayuda a comprender los elementos de la información de negocio así como los requisitos. Además permite identificar y eliminar redundancias de datos así como evitar inconsistencias. Además, a la larga, ayuda a reducir los costes de mantenimiento.

3. Modelo de datos físico (PDM): finalmente es necesario convertir el modelo lógico en un modelo físico. El PDM es una representación del diseño de la base de datos teniendo en cuenta las prestaciones y restricciones de la tecnología concreta que se vaya a utilizar.

El modelo de datos conceptual y lógico

Para elaborar el modelo de datos lógico vamos a utilizar un diagrama entidad-relación (E-R) sencillo 5.3. Para ello es necesario identificar las diferentes entidades del modelo. Una entidad es objeto importante acerca del que se va a almacenar la información.

El diseño de las entidades se ha llevado a cabo de modo que fuese posible tener por separado los diferentes elementos que nos puede interesar analizar por separado.

- Entidad content: el contenido consiste en toda aquella información extraída de internet antes de haber sufrido cualquier tipo de manipulación y/o análisis. Es la

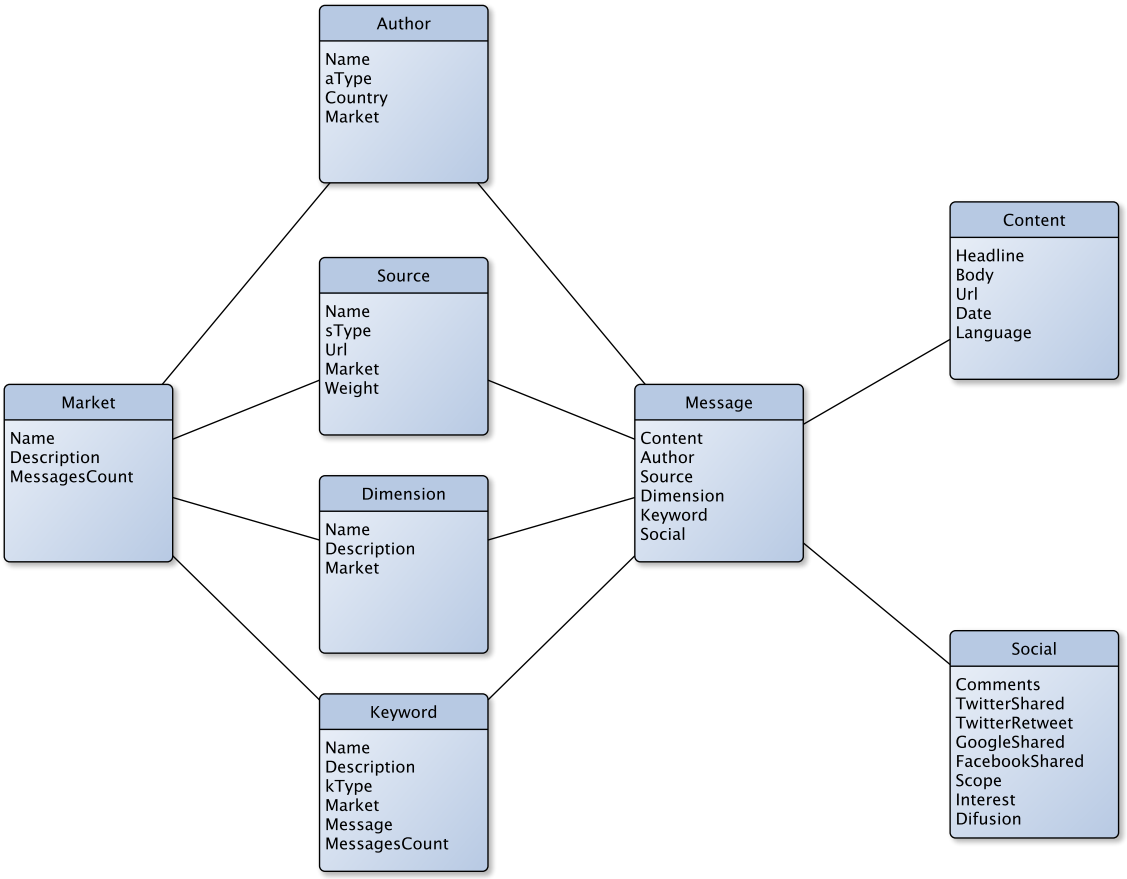


Figura 5.3. Representación del modelo lógico de datos. Representación de un diagrama E-R con las principales entidades en las que se ha dividido una noticia y sus relación.

información que alguien ha escrito y publicado en internet, y que posteriormente ha sido recogida por el motor de búsqueda de la plataforma. Es una entidad fuerte, ya que no necesita de ninguna otra entidad para ser definida inequívocamente. Los elementos recogidos de la noticia, y que serán atributos de esta entidad son:

- Headline: titular de la noticia
- Body: todo el contenido de la noticia
- Url: enlace a la noticia
- Date: fecha de publicación de la noticia
- Language: idioma en el que se ha escrito la noticia

- Entidad author: es el propietario de la página o perfil del que se ha recogido la noticia. Es una entidad fuerte cuyos atributos son:

  - Name: es el nombre, o pseudónimo de la persona o empresa que aparece como responsable de la publicación de la noticia.
  - aType: tipo de autor. No utilizado por ahora.
  - country: país en el que se ubica el autor. No se utiliza por ahora.

- Entidad Source: que describe concretamente la fuente de la noticia. El número de autores de una fuente puede variar sustancialmente, desde uno, como en un blog personal, hasta varios millones, como podrían ser facebook o twitter, en los que cada usuario, cuenta como un autor. También es posible que un autor publique en diferentes fuentes, pero dado que, por el momento, no podemos saber autores con el mismo nombre (name), publicando en fuentes distintas son en realidad el mismo autor, por ahora no se va a establecer ninguna relación entre el autor y la fuente, quedando esa posibilidad abierta para futuras ampliaciones del modelo de datos. Los atributos de esta entidad fuerte son:
  - Name: nombre de la fuente
  - sType: tipo de fuente. Entre los diferentes tipos tenemos: Web, Microblogging, Blog y Noticias.
  - Url: enlace a la página principal de la fuente

- Entidad Market: El contenido extraído de la red se analizará siempre por mercados. Así, nace la necesidad de considerar el mercado como una entidad. Es una entidad fuerte, siendo sus principales atributos:
  - Name: nombre del mercado
  - Description: breve texto utilizado para complementar la definición del mercado
  - MessagesCount: Contador con el número de mensajes disponibles para el mercado. No se utiliza por ahora.

- Entidad Keyword: al lanzar una búsqueda de contenido, es necesario definir una o varias palabras clave que servirán para localizar los mensajes. Esas palabras clave nos servirán luego en la clasificación y análisis de las noticias. Las palabras clave pueden repetirse entre mercados, pero en ese caso se considerarían palabras clave diferentes. Un mismo mensaje puede tener diferentes palabras clave, aunque en general todas serán del mismo mercado. De este modo, evitamos duplicar la entidad contenido cuando el mismo contenido aparezca lanzando buscadas mediante

diferentes palabras clave. Es una entidad débil, ya que no podemos caracterizar una Keyword sin conocer el mercado al que pertenece. Los atributos de la misma son:

  - Name: es la palabra clave en sí misma.
  - Description: texto que ayudará a comprender el significado de la palabra clave. No se utiliza.
  - kType: Tipo de palabra clave. Distinguimos por ahora entre producto, fabricante/laboratorio, enfermedad, etc. No se utiliza.
  - Market: mercado al que pertenece la palabra clave.

- Entidad Dimension: sirve para definir cada una de las dimensiones que pueden considerarse en un mercado. Por dimensión entendemos una temática concreta referente a un mercado. Por ejemplo, dentro de un mercado concreto, es posible hablar de la enfermedad, del precio de un medicamento, de los efectos secundarios de un tratamiento, etc. Cada una de esas temáticas, sería una dimensión diferente. Del mismo modo que la palabra clave, estamos ante una entidad débil cuyos atributos son:

  - Name: nombre de la dimensión.
  - Description: descripción que ayude a delimitar la temática a la que la dimensión hace referencia.
  - Market: Mercado en el que se sitúa la dimensión.

- Entidad Social: que consiste de toda aquella información que extraída directamente de la noticia, o calculada a partir de la misma, define la exposición que ha tenido la información en la red. Es una entidad débil, ya que toda la información hace referencia a una noticia concreta con la que estará relacionada. Los atributos son:

  - Comments: Número de comentarios que se han escrito referentes a la noticia o comentario principal.
  - TwitterShared: Número de veces publicado en Twitter.
  - TwitterRetwet: Número de veces que se ha retwiteado la micronota si se ha recogido de Twitter. Si no valdrá 0.
  - GoogleShared: Veces que se ha compartido la noticia en Google+
  - FacebookShared: Veces que se ha compartido la noticia en Facebook.

  - Scope: alcance. No utilizado.
  - Interest: métrica del interés suscitado por la noticia. No utilizado.
  - Diffusion: métrica para medir el nivel de difusión de la noticia. No utilizado

- Entidad message: es la entidad relacional que permite que integrar las demás entidades. A partir de esta entidad es posible conseguir toda la información disponible de una noticia. Los atributos son:
  - Content: contenido de la noticia. Relación con la entidad "content".
  - Source: información acerca de la fuente de la que se ha extraído la noticia. Relación con la entidad "Source".
  - Author: información acerca del autor de la notica. Relación con la entidad "Author".
  - Social: métricas del impacto que ha tenido la noticia en las redes sociales. Relación con la entidad "Social".

5.3.2. Extracción, Transformación y carga de datos

La carga de datos supone otro de los factores importantes en el diseño de la herramienta. Como se ha comentado anteriormente en este caso se ha optado por diseñar el proceso de ETL directamente mediante scripts de python.

La carga de datos para cada uno de los mercados se lleva a cabo mediante un Excel que contiene todas las noticias recogidas. Cada fila es una noticia, y cada columna corresponde a uno de los atributos de la noticia.

Al tener una sola fuente, el proceso de extracción de datos se simplifica. Por otro lado, el hecho de que el contenido de esa fuente se haya obtenido de medios sociales provoca el efecto contrario.

Si nos fijamos en la figura 5.4 vemos cómo el fichero llega a nuestras manos en forma de Excel. Lo primero que necesitamos (no está representado en la figura) es una interfaz de gestión de de ficheros para poder subir y/o descargar ficheros con datos.

Una vez el fichero está en el servidor, e indexado por la herramienta en la tabla *filestoload*, es posible llamar al proceso de extracción de datos, de modo que se recorren las diferentes filas del fichero Excel, y se carga toda la información con una estructura muy similar a la que tiene en el Excel en una tabla de nuestra base de datos. Este proceso

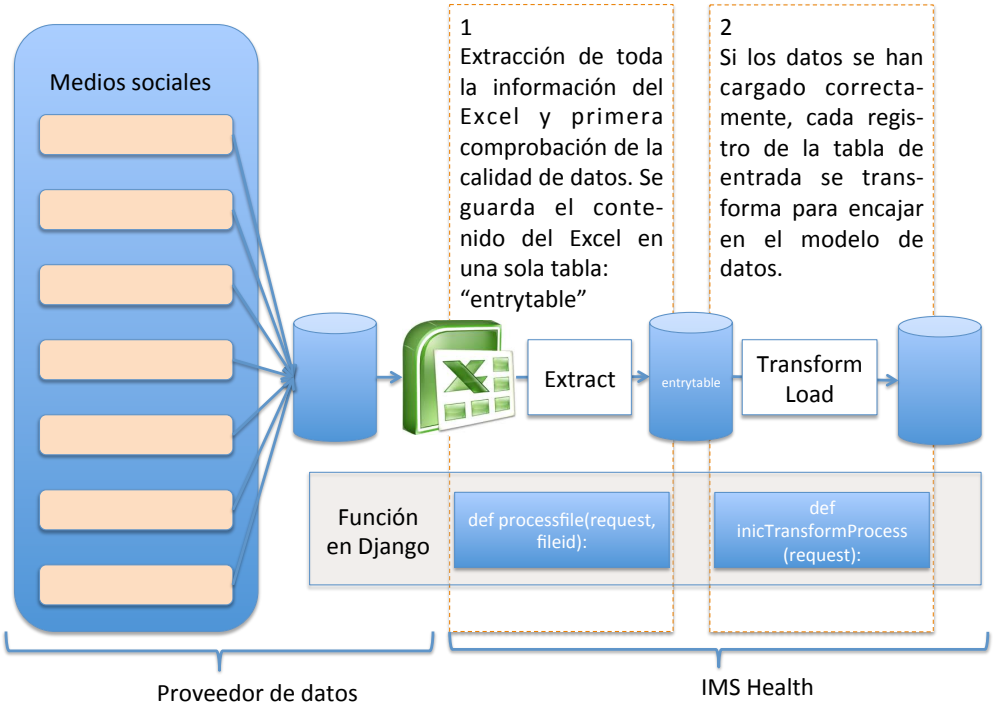


Figura 5.4. Carga de datos. Diagrama del proceso de carga de datos al servidor de desarrollo.

permite identificar incidencias como fallos de codificación, filas incompletas, formato de la fuente incorrecto, columnas con tipos de datos equivocados, etc.

Si hay algún problema en el proceso de extracción, éste se detendrá y Python levantará una excepción que nos permitirá identificar el origen del problema. Si en lugar de tener este primer proceso de carga, escribiésemos directamente sobre el modelo de datos, al fallar el proceso tendríamos la base de datos a medio cargar, y la herramienta en un estado no controlado, y posiblemente presentando resultados incorrectos.

Una vez el proceso de extracción ha finalizado correctamente, tendremos los datos en un formato más sencillo de manejar (ya están en nuestra base de datos y se habrán realizado varias validaciones sobre los mismos). Esto minimiza el riesgo de que falle el proceso de transformación de datos.

El proceso de transformación y carga de datos, recorre los registros de la tabla *entrytable*, y para cada uno busca si existen las entidades a las que pertenece (¿Existe la fuente? ¿Existe el autor? etc.) Si no existe, la crea, y si existe la guarda para acabar creando al final todas las relaciones necesarias.

Una vez terminado el proceso de transformación y carga de datos, se podrá acceder en el servidor de desarrollo, a las analíticas con los datos actualizados. Esto permitirá no solo validar que la carga se ha efectuado correctamente, también que la calidad del dato cargado es aceptable para ser publicada en el servidor de producción.

Cuando se considere que el cliente debe de tener acceso a la nueva información, se lanzará un nuevo proceso que actualizará el servidor de producción con los datos del servidor de desarrollo. El riesgo de este último proceso es muy bajo, ya que tanto tecnológicamente, como estructuralmente ambos servidores son iguales y el proceso lo único que hace es replicar la información de un entorno al otro.

Capítulo 6

Conclusiones y líneas futuras

En este capítulo se exponen los principales aprendizajes y hallazgos a los que se ha ido llegando a lo largo de este proyecto.

A medida que se avanza en la definición y desarrollo de una herramienta tecnológica se descubren también factores que de haberse conocido antes habrían permitido el desarrollo de una herramienta mejor. A partir de esos hallazgos se ha compuesto la sección de líneas futuras.

6.1. Conclusiones

A lo largo de este proyecto se ha entendido la importancia que están cobrando los medios sociales en parte de la estrategia empresarial. Tanto en el ámbito personal como en el profesional los medios sociales están ampliamente extendidos, eso hace que cualquier industria, y la farmacéutica no es una excepción, deba de establecer su propio planteamiento ante este fenómeno

Para poder elaborar una estrategia en torno a los medios sociales es de gran utilidad conocer las dinámicas con las que estos se rigen. En redes sociales, blogs, foros, etc., se genera constantemente una gran cantidad de información por muchos usuarios diferentes. Para tener conocimiento de eso es necesario disponer por un lado de una herramienta que recoja con el máximo grado de automatización posible el contenido que se va publicando, y por otro lado una herramienta de BI que permita procesar y analizar toda esa información para convertirla en conocimiento de negocio.

Más información no significa más conocimiento, y menos cuando ésta es tan heterogénea como lo es cuando proviene de medios sociales. Para transmitir de forma efectiva un mensaje que sea útil a la persona que toma las decisiones ha sido necesario diseñar con atención un conjunto de analíticas que transmitan el mensaje adecuado. Para ello hemos utilizando las capacidades en procesado de datos que nos ofrece Python trabajando con una base de datos MySQL sobre el framework Django, y la capacidad de generar gráficas mediante HTML y JavaScript ofrecida por la librería Highcharts.

Para este proyecto se ha decidido desarrollar la herramienta de análisis además de los análisis en sí mismos. En ese proceso también hemos ido descubriendo algunas de las herramientas de BI existentes y cómo funcionan.

También es importante destacar la importancia del modelo de datos a medida que el volumen y complejidad de estos va en aumento. El diseño de una estructura de datos adecuada es una de las partes más costosas en el diseño e implementación de una herramienta de BI. Equivocarse en ese proceso puede imposibilitar la inclusión de nuevos datos o hacer inviable la construcción de informes adicionales.

Finalmente cabe nombrar la importancia de ofrecer de manara holística el servicio a un cliente. Desde el conocimiento del mercado (farmacéutico, BI y social media), hasta una plataforma diseñada especialmente para dar respuesta a las necesidades de la industria.

6.2. Futuras Líneas de Trabajo

Los medios sociales evolucionan constantemente, las necesidades empresariales son cambiantes, el mercado farmacéutico se encuentra en una fase especialmente desafiante debido a la constante modificación del marco regulatorio, etc. Este proyecto es sólo una primera aproximación a las necesidades actuales de las empresas del sector farmacéutico en cuestión de social media. Se ha tenido en cuenta en todo momento el marco actual, pero sin dejar de pensar en que éste es muy dinámico. Para que el contenido de este proyecto mantenga la vigencia es necesario mantenerlo actualizado.

Durante el diseño y desarrollo de la herramienta analítica se han identificado aspectos de la misma que habrá que tener en cuenta en futuras revisiones:

- Algunas funciones de la herramienta no se han llegado a desarrollar en esta primera

fase, lo que hace que en ocasiones sea necesario ir directamente a la base de datos cuando es necesario configurar ciertos parámetros. Una interfaz de configuración de usuarios y mercados es posiblemente lo más urgente, ya que permitiría que un usuario pudiese añadir o quitar mercados y/o usuarios sin necesidad de acudir a la base de datos

- Las 6 analíticas que se han diseñado son sólo el principio. A medida que la herramienta se vaya poniendo a disposición de clientes irán apareciendo nuevas necesidades que se traducirán con mucha probabilidad en nuevos informes.

- Aunque los informe tienen varias funciones que permiten interactuar con la información hay muchas funciones, filtros, etc., que se pueden desarrollar para incrementar la interactividad de la información, y que aún no se han desarrollado

- Interfaz de análisis de las diferencias entre los datos contenidos en la base de datos de staging y la de producción. Se ha desarrollado una interfaz sencilla que controla el paso del contenido de la base de datos de staging a la base datos de producción. Esta interfaz muestra las diferencias en número de mensajes entre las dos bases de datos. Para entender mejor las diferencias entre los datos a disposición de los clientes y los que se van a cargar se ha detectado la necesidad de disponer de una interfaz que compare el contenido de ambas bases de datos y realice analíticas al respecto. Algunos aspectos que se podrían extraer son las fuentes añadidas, incremento o decremento de la media diaria mensajes, etc.

- El volumen de información publicado diariamente en los medios sociales es enorme. Dado que nos centramos en un sector muy estrecho dentro de todo lo que se publica, con sistemas relativamente poco sofisticados podemos gestionar por ahora toda esa información. Sin embargo, con el volumen de datos que hay actualmente en el sistema (del orden de varios miles de mensajes), el rendimiento a la hora de ejecutar ciertas consultas empieza a decrecer.

  Por ese motivo es necesario empezar a pensar en nuevas posibilidades a la hora de guardar y gestionar toda la información recibida. Para afrontar ésta problemática se podrían utilizar técnicas de big data. Este término surgido en 1997 [43] que ha cobrado mucha relevancia recientemente, hace referencia a sistemas que manipulan grandes conjuntos de datos. Para ello se utilizan herramientas en las que el procesado de datos suele ser distribuido.

  En cualquier caso, y sin adentrarnos más en un tema que daría para un (o varios)

PFC, Big Data es un mundo a explorar si queremos ser capaces de manejar de forma eficiente toda la información de medios sociales que es de interés para la industria farmacéutica.

Bibliografía

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business Horizons*, vol. 53, no. 1, pp. 59 – 68, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0007681309001232

[2] J. F. Chris Charron and C. Li, "Social computing," 2006.

[3] T. O'Reilly, "What is web 2.0: Design patterns and business models for the next generation of software," *Communications & Strategies*, no. 65, pp. 17 –37, 2007.

[4] (2012) O'reilly radar. [Online]. Available: http://tcc-web20.googlecode.com/svn/trunk/Pesquisa/O'Reilly%20Radar%20-%20Web%202.0%20Compact%20Definition.pdf

[5] C. Li and J. Bernoff, *Groundswell. Winning in a world transformed by social technologies*, 1st ed. Harvard Business Review Press, 2011.

[6] Wikipedia, "Inteligencia empresarial — wikipedia, la enciclopedia libre," 2012, [Internet; descargado 27-diciembre-2012]. [Online]. Available: http://es.wikipedia.org/w/index.php?title=Inteligencia_empresarial&oldid=62014726

[7] U. Dayal, M. Castellanos, A. Simitsis, and K. Wilkinson, "Data integration flows for business intelligence," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, ser. EDBT '09. New York, NY, USA: ACM, 2009, pp. 1–11. [Online]. Available: http://doi.acm.org/10.1145/1516360.1516362

[8] Wikipedia, "Data integration — wikipedia, the free encyclopedia," 2012, [Online; accessed 27-December-2012]. [Online]. Available: http://en.wikipedia.org/w/index.php?title=Data_integration&oldid=521418127

[9] (2012) How to gain business intelligence with social media. [Online]. Available: http://socialmediamarketinguniversity.com/gain-business-intelligence-social-media/

[10] Social media strategy in 8 steps. [Online]. Available: http://www.convinceandconvert.com/social-media-strategy/social-media-strategy-in-8-steps/

[11] D. C. S. G. Lodewijk Bos, Andy Marsh and M. Rees, "Patient 2.0 empowerment," *Proceedings of the 2008 International Conference on Semantic Web & Web Services SWWS08*, pp. 164 –167, 2008.

[12] M. S. Amanda K. Hall and J. M. Bernhardt, "Healthy aging 2.0: The potential of new media and technology," *Preventing Chronic Disease*, vol. 9, 2012.

[13] (2011) Health topics. [Online]. Available: http://pewinternet.org/Reports/2011/HealthTopics.aspx

[14] Wikipedia, "Mhealth — wikipedia, the free encyclopedia," 2012, [Online; accessed 1-January-2013]. [Online]. Available: http://en.wikipedia.org/w/index.php?title=MHealth&oldid=527383148

[15] C. Kaiser and F. Bodendorf, "Mining patient experiences on web 2.0 - a case study in the pharmaceutical industry," in *SRII Global Conference (SRII), 2012 Annual*, july 2012, pp. 139 –145.

[16] L. S. Karla Anderson and D. Garrett, "Social media likes healthcare. from marketing to social business," PriceWaterhouseCoopers, Tech. Rep., April 2012.

[17] (2012) 9 ways social media is impacting the business of healthcare | healthcare finance news. [Online]. Available: http://www.healthcarefinancenews.com/news/9-ways-social-media-impacting-business-healthcare

[18] (2012) Social media measurement should focus on outcomes, not output. [Online]. Available: http://veryofficialblog.com/2010/06/12/social-media-measurement-outcomes-not-output/

[19] (2012) Social media roi: Socialnomics - youtube. [Online]. Available: http://www.youtube.com/watch?v=ypmfs3z8esI&feature=player_embedded

[20] (2004) Web services architecture. [Online]. Available: http://www.w3.org/TR/ws-arch/#id2260892

[21] Cliente-servidor. [Online]. Available: http://es.wikipedia.org/wiki/Cliente-servidor

[22] Html. [Online]. Available: http://es.wikipedia.org/wiki/HTML

[23] introducción al html4. [Online]. Available: http://html.conclase.net/w3c/html401-es/intro/intro.html

[24] Javascript. [Online]. Available: http://es.wikipedia.org/wiki/JavaScript

[25] Base de datos. [Online]. Available: http://es.wikipedia.org/wiki/Base_de_datos

[26] Mysql. [Online]. Available: http://es.wikipedia.org/wiki/MySQL#cite_note-2

[27] Maximum table size for a mysql database. [Online]. Available: http://stackoverflow.com/questions/48633/maximum-table-size-for-a-mysql-database

[28] (1997) 1.4.4. dimensiones máximas de las tablas mysql. [Online]. Available: http://dev.mysql.com/doc/refman/5.0/es/table-size.html

[29] Modelo vista-controlador. [Online]. Available: http://es.wikipedia.org/wiki/Modelo_Vista_Controlador

[30] (2013) Django faq. [Online]. Available: https://docs.djangoproject.com/en/dev/faq/general/#django-appears-to-be-a-mvc-framework-but-you-call-the-controller-the-view-and-the-view-the-templa

[31] (2011) Apsl blog - introducción a celery. [Online]. Available: http://blog.apsl.net/weblog/2011/01/14/introduccion-a-celery/

[32] Factores que complican una monitorización online. [Online]. Available: http://www.slideshare.net/rogerbretau/factores-que-complican-una-monitorizacin-online

[33] Reporting tools selection in data warehousing. [Online]. Available: http://www.1keydata.com/datawarehousing/toolreporting.html

[34] (2013) Magic quadrant for business intelligence and analytics platforms. [Online]. Available: http://www.gartner.com/technology/reprints.do?id=1-1DZLPEP&ct=130207&st=sb

[35] Dashboard design. [Online]. Available: http://blogs.ischool.berkeley.edu/i247s12/files/2012/01/Dashboard-Design-Overview-Presentation.pdf

[36] (2010) What is business intelligence dashboard? [Online]. Available: http://searchbusinessanalytics.techtarget.com/definition/business-intelligence-dashboard

[37] S. Malik, *Enterprise Dashboards: Design and Best Practices for IT*. John Wiley & Sons Inc., 2005.

[38] (2011) How to design effective dashboard displays. [Online]. Available: http://www.dashboardinsight.com/articles/digital-dashboards/fundamentals/how-to-design-effective-dashboard-displays.aspx

[39] (2010) Bar charts vs. line charts. [Online]. Available: http://blog.axc.net/?p=971

[40] (2002) Best practices for site deployment using content management server site deployment manager. [Online]. Available: http://technet.microsoft.com/en-us/library/bb676198.aspx

[41] Web servers and workflow. [Online]. Available: http://webdesign.about.com/od/servers/qt/web-servers-and-workflow.htm

[42] (2013) Front-end design principles. [Online]. Available: http://clearleft.com/thinks/front-enddesignprinciples/

[43] (2013) A very short history of big data. [Online]. Available: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

[44] F.-Y. Wang, K. M. Carley, D. Zeng, and W. Mao, "Social computing: From social informatics to social intelligence," *Intelligent Systems, IEEE*, vol. 22, no. 2, pp. 79 –83, march-april 2007.

[45] S. Murugesan, "Understanding web 2.0," *IT Professional*, vol. 9, no. 4, pp. 34 –41, july-aug. 2007.

[46] J. Krumm, N. Davies, and C. Narayanaswami, "User-generated content," *Pervasive Computing, IEEE*, vol. 7, no. 4, pp. 10 –11, oct.-dec. 2008.

[47] World wide web. [Online]. Available: http://es.wikipedia.org/wiki/Web#Historia

[48] (2012) The entity design pattern. [Online]. Available: http://www.codeproject.com/Articles/4293/The-Entity-Design-Pattern

Índice de figuras