Universidad Zaragoza
1542

Trabajo Fin de Máster

Sistema Experto de Probabilidad y Severidad en Red

Autor

Gonzalo Ruiz Manzanares

Director y ponente

Director: David Íñiguez Dieste
Ponente: Sergio Ilarri Artigas

Escuela de Ingeniería y Arquitectura

2014

Sistema Experto de Probabilidad y Severidad en Red
RESUMEN

El suministro eléctrico es un elemento esencial en la vertebración y el crecimiento socioeconómico de la sociedad actual. Las exigencias que deben satisfacer las redes eléctricas y las distribuidoras en cuanto a la continuidad y seguridad del suministro son crecientes.

SEPS (Sistema Experto de Probabilidad y Severidad en Red) es un proyecto del programa INNPACTO del Ministerio de Economía y Competitividad liderado por la empresa Gas Natural Fenosa, en cuyo consorcio participan otras empresas y organismos de investigación. El objetivo fundamental del proyecto es desarrollar un sistema que a partir de la información de previsión meteorológica, de demanda y de generación, obtenga unos índices de probabilidad de ocurrencia de incidencias y restricciones en la red de distribución de energía; y a partir de información sobre el mercado afectado y eventos relevantes, sea capaz de estimar el impacto social de dichas incidencias; permitiendo mejorar así el servicio del suministro eléctrico.

Este trabajo fin de máster se centra en la vertiente social del problema, consistente en el análisis de datos de Internet para identificar la severidad reputacional y la repercusión social de los incidentes sobre la red de distribución de energía.

El trabajo se dividió en dos partes. Por un lado, se desarrolló un sistema para la captura de datos de diversas fuentes heterogéneas de Internet (sitios web, redes sociales, agendas online, etc.) sobre eventos de gran repercusión en los que un incidente eléctrico pudiera tener cierto impacto en la imagen de Gas Natural Fenosa. Para ello, se estudiaron y seleccionaron diferentes fuentes que contuvieran datos relevantes para el proyecto, se diseñaron una serie de robots o scrapers para la extracción automática de estos datos, y se definieron distintas variables para los eventos recogidos que facilitaron la integración de esta parte social en el modelo global de previsión de incidentes.

Por otro lado, una vez sucedido un incidente, se ha desarrollado un sistema que rastrea continuamente redes sociales, webs de noticias y otras herramientas, para medir el impacto que realmente ha tenido dicho incidente en la imagen de marca de la empresa en base a su gestión. Para ello, además del uso de técnicas de extracción de datos, se realizó un estudio del estado del arte en cuanto a las distintas técnicas de análisis de sentimiento existentes, implementando un sistema para la evaluación de la polaridad de textos.

Cláusula de confidencialidad

La presente memoria, correspondiente al Trabajo Fin de Máster titulado "Sistema Experto de Probabilidad y Severidad en Red", ha sido realizada con el consentimiento del Instituto de Biocomputación y Física de Sistemas Complejos (BIFI) y de Unión Fenosa Distribución.

Toda la información comunicada por una de las partes a la otra se considerará recibida con carácter de confidencialidad y será utilizada únicamente para los propósitos de este proyecto.

Este documento no podrá ser consultado ni duplicado parcial o totalmente para otro propósito que no sea su evaluación, salvo autorización expresa del BIFI, de Unión Fenosa Distribución, y del autor del mismo.

Esta cláusula ha sido incluida por voluntad expresa del organismo y la empresa citados como requisito para otorgar el consentimiento de la elaboración de la presente memoria.

Agradecimientos

A mi familia, que tanto me ha apoyado y animado durante todo este tiempo, en especial a mi abuela, que falleció recientemente.

A Desi, por aguantarme todos estos años.

A mis amigos y colegas, que han hecho más liviano todo este tiempo.

A Alfredo y Alejandro, por sus consejos y su apoyo durante el proyecto.

A David Íñiguez, por dirigirme y apoyarme en esta interesante aventura.

A Sergio Ilarri, por su atención, su disponibilidad, y por encauzarme en la dirección correcta.

Al BIFI en general, por ofrecerme la oportunidad de participar en proyectos como éste durante los últimos años.

Índice general

Índice de figuras

Índice de tablas

Parte I

Memoria

Capítulo 1

Introducción

1.1 Alcance del documento

Este documento recoge toda la información sobre el trabajo realizado en el TFM[1] titulado "Sistema Experto de Probabilidad y Severidad en Red", a partir de ahora SEPS, de la titulación de Máster en Ingeniería de Sistemas e Informática impartido en la Escuela de Ingeniería y Arquitectura de la Universidad de Zaragoza, desarrollado por el alumno Gonzalo Ruiz Manzanares durante el período comprendido entre los meses de julio del año 2013 y mayo del año 2014 para el BIFI[2] y la empresa Unión Fenosa Distribución.

Este TFM se engloba en el desarrollo de una plataforma para la previsión de riesgos y la estimación de su severidad dentro de una red eléctrica de distribución. Concretamente, se centra en la parte de la previsión y la estimación de la severidad de riesgos sociales. El trabajo realizado se divide en dos partes bien diferenciadas. Por un lado, lo que se pretende con este sistema a nivel social, es prevenir aquellos incidentes que puedan tener una mayor repercusión social en base a distintos parámetros, debido a que serán los que más afecten a la imagen de la empresa y a su servicio. Por otro, la empresa quiere conocer, de forma actualizada, el sentimiento de la sociedad hacia su marca en base a la gestión de dichos incidentes, tomando como referencia diferentes fuentes de información de Internet.

El autor de este TFM ha realizado todo el trabajo relativo a la vertiente social del proyecto, y en este documento se detallan las partes de las que ha constado dicho trabajo, las fases y problemas por las que ha pasado cada una de ellas, las decisiones y soluciones adoptadas con sus pertinentes argumentaciones, y la documentación en la que se ha apoyado.

[1] Trabajo Fin de Máster

[2] Instituto de Biocomputación y Física de Sistemas Complejos

1.2 Contexto de desarrollo

El proyecto SEPS perteneciente al programa INNPACTO de 2012 ha sido desarrollado por parte de un consorcio de empresas y de instituciones, concretamente las siguientes:

- Unión Fenosa Distribución, perteneciente al Grupo Gas Natural Fenosa (GNF), como líder del proyecto.
- Telvent, actualmente Schneider Electric, empresa experta en sistemas de información para el sector energético, transportes y meteorología.
- AIA (Aplicaciones en Informática Avanzada), empresa con experiencia en el modelado de redes eléctricas, así como en previsión de demanda y generación.
- SCIEN Analytics, compañía con conocimientos en modelos de previsión, sistemas de información geográfica y análisis espacial, así como desarrollo e integración de sistemas de información.
- Universidad de Girona, con experiencia en gestión de la información, sistemas expertos y descubrimiento de patrones.
- Universidad Carlos III, experta en investigación de modelos de sistemas eléctricos y en análisis de incidentes.
- BIFI (Instituto de Biocomputación y Física de Sistemas complejos) de la Universidad de Zaragoza, al que el autor de este TFM pertenece, que es un centro con conocimientos en análisis de redes complejas y extracción de conocimiento a partir de fuentes desestructuradas procedentes de Internet.

Cada empresa y cada institución desarrolla un papel distinto dentro del proyecto SEPS en los diferentes paquetes de trabajo en los que éste está dividido, pero este TFM se centra únicamente en el número cuatro, llamado "DATA MINING Y ESTIMACIÓN DE SEVERIDADES", cuyo responsable es el autor de este trabajo.

1.2.1 BIFI

El BIFI es un instituto de investigación de la Universidad de Zaragoza que promueve la interdisciplinariedad para afrontar los retos científicos y tecnológicos del presente.

Está formado por investigadores de la Universidad de Zaragoza y de otras instituciones españolas y extranjeras. Su objetivo es desarrollar investigación competitiva en las áreas de computación aplicadas a la física de sistemas complejos y modelos biológicos.

Además de la investigación en ciencia básica, un punto fundamental del instituto es la transferencia de tecnología entre la universidad y el mundo empresarial.

Para llevar a cabo estos objetivos, el BIFI cuenta con el trabajo de investigadores de áreas diferentes cuya colaboración está llevando a sinergias muy significativas. En particular, expertos en supercomputación, físicos trabajando en ciencia de materiales, química cuántica o redes complejas, y biólogos trabajando en problemas estructurales como desarrollo de fármacos y plegamiento de proteínas.

1.2.2 Unión Fenosa Distribución

En el año 1999 se creó la empresa independiente Unión Fenosa Distribución S.A., que es la encargada de la actividad regulada de distribución eléctrica del Grupo Gas Natural Fenosa. La empresa atiende el suministro de energía eléctrica en cuatro comunidades autónomas y trece provincias de España, convirtiéndola en una de las principales distribuidoras eléctricas del país con importante presencia internacional del grupo.

Unión Fenosa Distribución siempre ha mostrado un gran interés en la participación en proyectos de innovación, de hecho, SEPS es en parte continuación de otras iniciativas suyas, pero con partes muy novedosas (como en la que se basa este TFM), y esto es debido a su constante afán por mejorar la calidad de su servicio hacia sus clientes y trabajadores. De hecho, una de las cosas que más ha llamado la atención al autor de este trabajo, es que la variable más importante para medir el éxito de la empresa es el número de días que llevan sus trabajadores sin sufrir ningún tipo de accidente.

1.3 Motivación del proyecto

El suministro eléctrico es un elemento esencial en la vertebración y el crecimiento socioeconómico de la sociedad actual. Las exigencias que deben satisfacer las redes eléctricas y las distribuidoras en cuanto a la continuidad y seguridad de suministro son crecientes, en un entorno en el que la necesidad de una mayor integración de las energías renovables en la red forma parte de la necesaria sostenibilidad medioambiental del sector.

En la consecución de este crucial objetivo existen importantes retos técnicos que los gestores de las redes de distribución deben afrontar para conseguir que la importante evolución tecnológica hacia redes eléctricas inteligentes que están experimentando las redes actuales vaya acompañada de crecientes estándares de calidad.

Dentro de dichos retos, sin duda, de los más complicados de gestionar están algunos factores extrínsecos que por su naturaleza no es posible corregir como es el caso de la variabilidad con la que la meteorología impacta sobre las redes eléctricas.

El objetivo fundamental del proyecto SEPS es desarrollar un sistema experto que a partir de la información de previsión meteorológica y previsión de demanda y de generación, genere unos índices de probabilidad de ocurrencia de incidencias y restricciones en la red de distribución; y a partir de información sobre el mercado afectado y eventos relevantes, el impacto social de dichas incidencias.

La importancia de disponer de indicadores fiables de previsión de incidentes de cara a la prevención de fallos en el suministro eléctrico a los usuarios es de vital importancia para las compañías de distribución eléctrica, ya que ven penalizada su retribución y su reputación, y también para la mejora de la calidad del servicio, la productividad de las empresas consumidoras y la satisfacción de los usuarios domésticos.

Para ello se elaborarán algoritmos de previsión de incidencias basados en la información histórica y las previsiones meteorológicas, que junto con otros, sirven para identificar la severidad y la repercusión social de los incidentes sobre la red a partir de aspectos intrínsecos (potencia y clientes afectados, tiempos de reposición, etc.) y extrínsecos (sensibilidad social, eventos relevantes, etc.).

Las redes eléctricas se pueden dividir en diferentes partes, y cada una de ellas está compuesta de distintos elementos como podemos ver en el esquema de la figura 1.1. Mediante este proyecto, se podrán establecer estrategias preventivas de explotación de las redes que permitan gestionar de forma anticipada condiciones meteorológicas extremas o excepcionales y minimizar su impacto sobre la continuidad y seguridad de suministro.

Para ello, se va a trabajar a nivel de subestación de distribución, que son los elementos que se encuentran más cerca de los clientes finales, ya sean pequeños (viviendas) o grandes (hospitales, estadios, etc.). Un incidente meteorológico no previsto o un pico de demanda por un evento en el que no se ha dimensionado correctamente la potencia necesaria pueden producir consecuencias similares, dejando sin servicio desde un barrio de una ciudad a un hospital en el que se está tratando con vidas humanas, con todo lo que ello conlleva. Por ello, los beneficios fundamentales que proporcionará la utilización de SEPS pueden resumirse, por tanto, como:

- Mejora de la calidad de suministro
- Reducción del impacto social/mediático de los incidentes y restricciones de red
- Máxima integración de renovable en la red

Uno de los aspectos más innovadores de esta propuesta es la incorporación de un modelo de impacto sobre la reputación de la empresa considerando la posible repercusión social de un incidente. Para ello, se realizará una labor de recuperación automática de información de Internet acerca de acontecimientos sociales que puedan afectar a la repercusión mediática de una incidencia y a la previsión de demanda de energía.

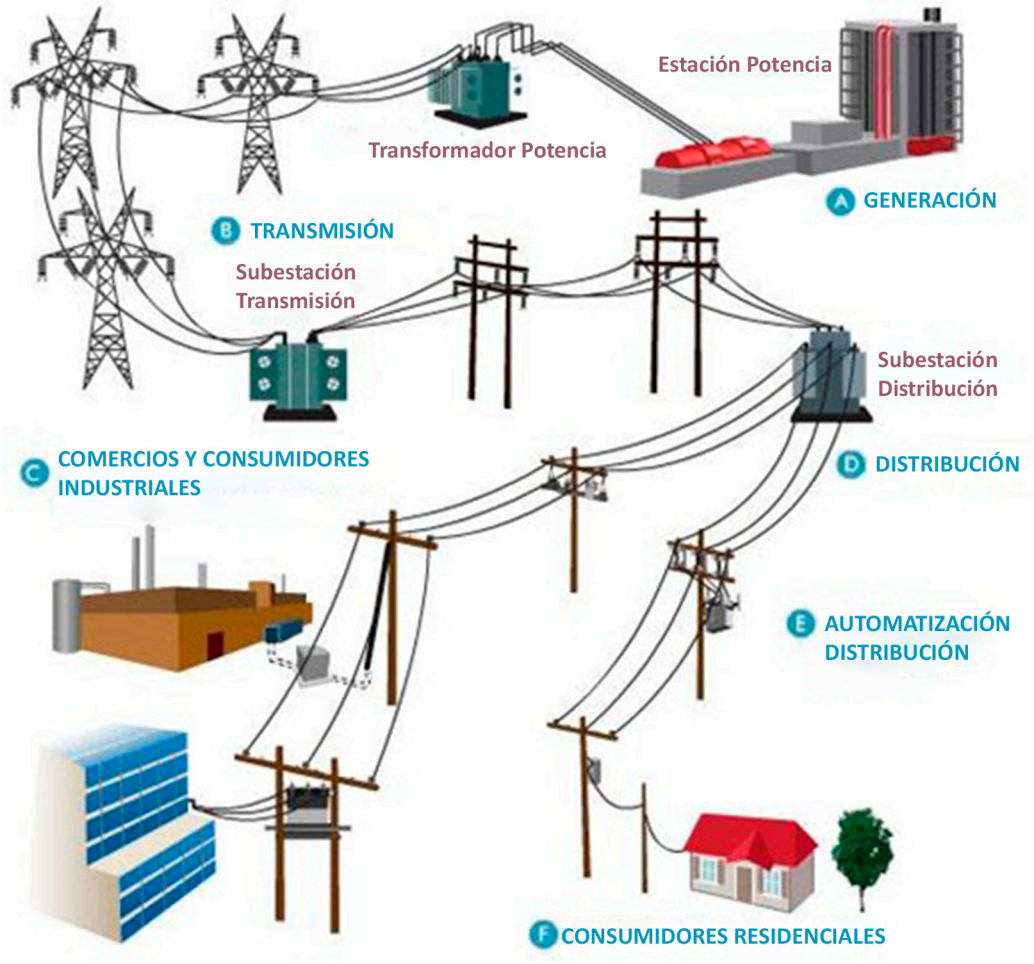

Figura 1.1: Estructura de una red eléctrica

Este TFM se centra únicamente en la parte social del proyecto que acabamos de mencionar, comprendiendo tanto la previsión de posibles eventos en los que un incidente pueda tener un mayor impacto social estimándolo a priori, así como el análisis a posteriori de la información distribuida en Internet sobre incidentes ya sucedidos para evaluar la repercusión de estos en la imagen de marca de Unión Fenosa Distribución.

1.4 Objetivos y alcance

Como acabamos de comentar, el objetivo principal del proyecto es el diseño y desarrollo de un sistema para la previsión de riesgos y la estimación de su severidad. Dicho sistema será capaz de mostrar a sus operadores las posibles incidencias que se puedan producir en la red eléctrica de Unión Fenosa Distribución dentro de la Comunidad autónoma de Galicia a 3 días vista en base a diferentes variables (clima, consumo, demanda, eventos, ...).

La interfaz del sistema será web, y permitirá visualizar en todo momento las incidencias previstas en un horizonte de 3 días. Las alarmas se encontrarán geoposicionadas con la mayor precisión posible, y proporcionarán un detalle de por qué se estima que en ese lugar puede haber un incidente dado un período, cuál es su severidad, y en base a qué motivo.

Todos los datos serán recogidos de diferentes fuentes, procesados y almacenados en base a un modelo de datos común para su utilización de forma automática por el sistema. Los modelos desarrollados en cada uno de los paquetes de trabajo los utilizarán como entradas para producir los resultados de previsiones y estimaciones.

Como hemos apuntado previamente, este TFM se centra única y exclusivamente en la vertiente social del problema, la cual se ha dividido en dos partes bien diferenciadas. Por un lado se intentarán prever incidentes en base a eventos de diversa índole (deportivos, culturales, lúdicos, informativos, etc.) y se tratará de estimar su severidad social en base a la opinión de los expertos o históricos. Por otro, se capturará de forma continua información de la red (webs de noticias, redes sociales, etc.) con el fin de tener un termómetro de cómo ve la sociedad la imagen de la empresa en base a estas incidencias de suministro y a la gestión de las mismas.

La captura de la información se realizará a través de scrapers (también denominados robots, crawlers o arañas) que accederán a los distintos recursos online, filtrando única y exclusivamente los datos que sean relevantes para el sistema. Estas arañas simulan el comportamiento de un usuario, pero de forma automatizada. Será importante estudiar bien qué información recoge cada fuente para elegirlas bien, así como su método de acceso y su política de límites para evitar abusos o sobrecargas.

Los datos de eventos se integrarán en un modelo jerárquico en el que habrá cuatro variables que compondrán la rama de severidad social. Estos modelos son diseñados en conjunto entre los distintos socios y los expertos de Unión Fenosa de forma que se van dividiendo las diferentes variables que afectan a la red en otras más concretas hasta llegar a las hojas. Una vez

definidas, se pondera su correlación, de manera que se obtiene una función que puede operar directamente con los valores finales de las hojas para obtener un valor único que indique la severidad global estimada.

A través de técnicas de análisis de sentimiento, se analizarán los datos capturados de redes sociales, webs de noticias, comentarios, etc. Dichos datos serán relevantes especialmente cuando contengan ciertas palabras directamente relacionadas con incidencias y la empresa, pero sobre todo, serán cruciales tras una incidencia.

1.5 Restricciones, suposiciones y dependencias

La principal restricción de este proyecto es la confidencialidad, ya que en él se maneja información muy importante de diferentes compañías, principalmente Unión Fenosa Distribución, cuya publicación podría repercutir considerablemente en su operación como empresa. Además, este TFM tiene la restricción de ceñirse exclusivamente a la parte social del problema en su conjunto como ya hemos explicado en apartados anteriores. A nivel técnico, también se han establecido restricciones como el uso de tecnologías abiertas que el resto de socios del proyecto SEPS conozcan, siempre con el fin de integrar con el menor trabajo posible todas las piezas que conforman el proyecto global. El ámbito de este proyecto se restringe a la región de Galicia, que es una de las zonas de donde Unión Fenosa Distribución dispone de más datos, pero se ha desarrollado de manera que pueda funcionar en cualquier región. Además, se trabajará a nivel de subestación, que es el más indicado para que la empresa pueda realizar los movimientos pertinentes para reducir el impacto de un incidente o incluso evitarlo.

Centrándonos en la parte social, este proyecto se ha desarrollado suponiendo que los expertos de Unión Fenosa estarán atentos y revisarán los datos que se obtengan del sistema final, pudiendo realizar ajustes en caso de que fuera necesario, no en cuanto a la lógica del sistema, sino en cuanto a la interpretación de los mismos. Además, se da por supuesto que la entidad que albergue este sistema tendrá recursos suficientes tanto a nivel de hardware, red, suministro eléctrico, etc., como para que éste funcione correctamente. El sistema ha sido diseñado para estar funcionando de forma continua y recuperarse ante fallos.

El proyecto global SEPS tiene diferentes dependencias, ya que está dividido en distintos paquetes de trabajo, cada uno centrado en una parte diferente. El sistema de previsión de riesgos sociales y análisis de opinión en Internet desarrollado en el marco de este TFM es una parte de un modelo de previsión de riesgos general que tiene en cuenta otras variables como meteorología, incidentes pasados, consumo, etc. Dado que la planificación de ejecución de SEPS va más allá de la de este TFM, la parte social ha sido desarrollada hasta el final de forma independiente (siempre teniendo en cuenta la facilidad de integración y las restricciones), adelantando trabajo a nivel de interfaz de usuario de cara al control del sistema y de la monitorización de sus resultados. Esta parte de interfaz puede que no sea la que se utilice finalmente, ya que esto se decidirá con el resto del consorcio del proyecto cuando proceda.

1.6 Trabajo realizado por el autor

Nos encontramos ante un proyecto en cuyo desarrollo han trabajado diversas personas y entidades. A pesar de ello, la labor realizada por el autor de este documento es totalmente nueva, utilizando las últimas técnicas y herramientas disponibles, y considerando en todo momento los acuerdos del consorcio. El trabajo del autor ha consistido en el desarrollo del trabajo asignado al BIFI dentro de este proyecto, que como ya hemos mencionado, se corresponde con la totalidad de la componente social del mismo.

Dicho trabajo se puede dividir en las siguientes tareas:

- Estudio de posibles fuentes en Internet de las que obtener información relevante para el sistema, tanto para extraer futuros eventos (deportivos, culturales, de ocio, etc.), como para obtener sentimientos y opiniones de la sociedad. Se analizaron varios aspectos como su estructuración, sus límites, sus restricciones legales, etc.
- Instalación y mantenimiento de una herramienta para la planificación y el seguimiento del proyecto
- Análisis, diseño, implementación y optimización de una arquitectura HW[3]+SW[4] que soportara toda la carga necesaria por el sistema
- Análisis de las diferentes técnicas de extracción de información de sitios web, siempre teniendo en cuenta sus límites y las restricciones del consorcio
- Diseño e implementación de un modelo de datos en el que se pueda almacenar toda la información que se recoja
- Diseño e implementación de los distintos scrapers necesarios para la extracción de datos de las fuentes elegidas
- Diseño de un sistema de control para la ejecución de los diferentes scrapers desarrollados en el punto anterior
- Definición de los valores por defecto, las consultas y filtros necesarios para la evaluación de los posibles riesgos, así como para su integración con el modelo general de previsión
- Estudio de las diferentes técnicas y herramientas existentes para el análisis de sentimiento o minado de opinión de textos
- Implementación de un analizador de sentimiento basado en el estudio anterior, y comparación de resultados con otras herramientas del mercado

[3] Hardware
[4] Software

- Análisis, diseño e implementación de una interfaz para controlar el sistema de recogida de datos y evaluación de riesgos subyacentes, así como para la monitorización de sus resultados
- Gestión de versiones y copias de seguridad de la aplicación, así como empaquetado para su fácil instalación
- Documentación de las reuniones realizadas durante el proyecto
- Planificación del proyecto

En el anexo C se puede ver el tiempo aproximado que se ha dedicado a estas tareas y el calendario inicial que se propuso.

1.6.1 Herramientas utilizadas

En esta sección resumimos algunas de las diferentes herramientas que se han utilizado para el desarrollo de este proyecto. En el anexo F podemos encontrar más detalles sobre ellas, así como la motivación que ha llevado a usarlas.

Entorno de desarrollo

- **Sistema operativo:** OS X Mavericks
- **Servidor web:** Django 1.6.5
- **IDE**[5]**:** Eclipse 4
- **Control de versiones:** Mercurial
- **Debuggers:** Chrome Development Tools

Entorno de producción

- **Sistema operativo:** Ubuntu 14.04 LTS[6]
- **Servidor web:** Django 1.6.5 a través de Apache 2.2

[5] Integrated Development Environment
[6] Long Term Support

Herramientas utilizadas

- **Lenguajes de programación, lado del servidor:** Python
- **Lenguajes de programación, lado del cliente:** HTML[7], CSS[8], Javascript, JSON[9]
- **Base de datos:** PostgreSQL 9.3
- **Tratamiento digital de imágenes:** Adobe Photoshop
- **Otras librerías:** JQuery, Google Maps, Google Charts, Bootstrap

Documentación

- **General:** LaTeXen TeXShop, Google Docs, Microsoft Word 2013, Microsoft Excel 2013
- **Gráficos, diagramas, y presentaciones:** Microsoft PowerPoint 2013

1.7 Contenido de la documentación

El presente documento se divide en dos partes:

- **Memoria**, página 1, que a su vez contiene los siguientes capítulos:
  - **Introducción**, el presente capítulo, en el que se resumen el contexto, la motivación y los objetivos del proyecto.
  - **Panorámica del proyecto**, página 15, que ofrece una visión global del problema planteado para su desarrollo y de la solución propuesta.
  - **Previsión de riesgos sociales**, página 19, en el que se explica el estudio realizado y la solución desarrollada para la parte de previsión de riesgos.
  - **Análisis de opinión social**, página 27, que recoge el estado del arte del análisis de sentimiento y la implementación de un sistema con tal fin.
  - **Conclusiones**, página 37, en el que se contemplan las conclusiones alcanzadas tras la realización del mismo.
- **Anexos**, página 41, que son los siguientes:
  - **Propuesta**, página 43, en el que se incluye parte del documento presentado como propuesta al Ministerio de Economía y Competitividad.

[7] HyperText Markup Language
[8] Cascading Style Sheets
[9] JavaScript Object Notation

- **Especificaciones Detalladas del Modelo de Severidad**, página 67, en el que se incluyen las especificaciones acordadas para el modelo de severidad dentro del paquete de trabajo cuatro por parte del consorcio.
- **Planificación y seguimiento**, página 77, en el que se detallan la planificación y la dedicación para el desarrollo del proyecto realizado.
- **Aplicación piloto**, página 81, en el que se incluyen capturas de la interfaz piloto creada para el uso del sistema.
- **Modelo de datos**, página 93, en el que se detalla el modelo de datos utilizado para el almacenamiento.
- **Marco tecnológico**, página 99, en el que se recogen las tecnologías utilizadas.
- **Palabras y usuarios clave**, página 111, en el que se explica qué palabras y usuarios fueron escogidos y con qué criterios para la extracción de información relevante para el proyecto.
- **Evaluación de otros sistemas de análisis de sentimiento**, página 117, que contiene una comparativa entre la herramienta desarrollada y algunas de las alternativas disponibles.
- **Acrónimos**, página 127, que recoge los acrónimos utilizados en este documento.
- **Bibliografía**, página 129, en el que se encuentran las referencias bibliográficas utilizadas en el desarrollo del proyecto.

Capítulo 2

Panorámica del proyecto

2.1 Introducción

En este capítulo se da al lector una visión global del desarrollo de este proyecto, describiendo el problema que ha motivado su realización, introduciendo las ideas más importantes que aparecen a lo largo de su evolución, y proporcionando una visión general de la solución adoptada.

2.2 Planteamiento del problema

Como hemos descrito previamente, el suministro eléctrico es un elemento importantísimo para la sociedad actual. Por ello, las compañías eléctricas deben de satisfacer una serie de requisitos en cuanto a continuidad y seguridad que han ido creciendo con los años, y se prevé que así continúe. Del suministro eléctrico no sólo dependen nuestras comodidades, sino también nuestras vidas.

2.2.1 Cuestiones generales

Para satisfacer estos requisitos, actualmente se invierte mucho dinero y esfuerzo en diseño, instalaciones, investigación, seguridad, etc., pero sigue existiendo un vacío en cuanto a previsión. En otros ámbitos como la meteorología, la predicción resulta ser un concepto más familiar, y todos entendemos su utilidad. Pero si analizamos cuidadosamente el problema, podemos ver que metodologías análogas se pueden aplicar a otras disciplinas. Basándonos en hechos históricos, y observando lo que está sucediendo en un momento dado, podemos tratar de prever qué puede pasar en distintos puntos clave de una red eléctrica según el clima, las características de las instalaciones, averías sucedidas en el pasado, la demanda, o los

posibles eventos que vayan a tener lugar cerca de distintos puntos esenciales para el suministro eléctrico. Si juntando todo esto fuéramos capaces de prever posibles incidentes eléctricos, las empresas suministradoras tendrían capacidad suficiente para anticiparse, movilizar escuadras de técnicos a zonas con más riesgo, sustituir equipos con la antelación necesaria, etc., lo que se traduciría en un suministro mucho más seguro, eficaz, y fiable. Esto puede evitar desde que un barrio entero se quede sin luz con los inconvenientes que ello conlleva, hasta que un hospital acabe sin suministro tras una catástrofe natural.

2.2.2 Cuestiones específicas

Este TFM se centra exclusivamente en la parte de previsión de riesgos y estimación de severidad sociales. Con ello se pretende anticipar eventos que puedan tener una mayor repercusión social (partido de fútbol de Celta de Vigo, concierto de famoso artista, etc.) ya sea por asistencia como por sensibilidad de su audiencia, para que las distribuidoras puedan centrar los esfuerzos en estos puntos clave, con el fin de satisfacer mejor a sus clientes. También se analiza el impacto de un incidente una vez que ha sucedido para ver si ha sido correctamente gestionado, cuánto se podría mejorar, y si su estimación a priori estaba correctamente establecida. Es una parte muy novedosa, de la que el consorcio no tiene constancia que se haya aplicado en ámbitos de suministro eléctrico.

2.3 Solución propuesta

La solución propuesta por el BIFI y la empresa, y desarrollada íntegramente por el autor de este TFM, consta de dos partes:

La primera de ellas trata la previsión de incidentes. Consiste en la identificación de una serie de fuentes de Internet que contengan información relevante para el suministro eléctrico sobre todo tipo de eventos (deportivos, culturales, de ocio, etc.) con el fin de extraerla y analizarla para identificar las zonas más vulnerables en un horizonte de 72 horas. Esta información es capturada mediante una serie de scrapers o robots (también llamados crawlers o arañas) que almacenan los datos de interés estructurados en una base de datos diseñada con este fin. Los operarios tienen acceso a un sistema web a través de cuya interfaz pueden:

- Controlar el sistema de captura de datos, las estimaciones de audiencia y su sensibilidad
- Monitorizar los puntos críticos en los que es susceptible la aparición de incidentes con mayor repercusión social, con el fin de poder movilizar los equipos de mantenimiento a zonas cercanas, establecer rutas alternativas de suministro, etc.

Por otro lado, encontramos la parte de análisis de sentimiento. En ella, constantemente se

capturan datos de diferentes fuentes de Internet que puedan estar relacionados con el suministro eléctrico e incidentes que hayan surgido, lo que permite analizar a posteriori cuál es el sentir de la sociedad ante dichos incidentes y hacia su posterior gestión. Esto permitirá a las empresas distribuidoras manejar mejor estos incidentes en el futuro, ajustar la estimación de severidad a priori en caso de que haya sido errónea, así como tener una idea de lo que piensa la gente respecto a su servicio.

Para que todo esto funcione, se ha diseñado e implementado una arquitectura HW+SW en un nodo principal en el que se realizan todas las operaciones de recogida y procesado de datos. En dicho nodo también se encuentra desplegado el servidor web desde el que se controla el sistema y se consultan los datos analizados. Además, se dispone de una serie de pequeñas máquinas virtuales de cloud computing[1] que serán utilizadas dinámicamente en caso de que se alcance alguno de los límites de las fuentes que se recogen. Podemos ver un esquema de la arquitectura de la solución en la figura 2.1

2.4 Ventajas de la solución

Los problemas que pueden ocasionar los incidentes en redes eléctricas en términos de pérdidas económicas y de reputación pueden ser muy altos, por lo que disponer de un sistema como el propuesto que, aunque no sea 100% fiable, pueda mitigar las consecuencias de los incidentes, e incluso reducir su número, es de gran ayuda.

Aunque la inversión necesaria para desarrollar un sistema de estas características es elevada, ésta puede ser amortizada con la previsión/anticipación de un número no muy alto de incidentes, ya que la repercusión de los mismos tanto en el aspecto económico como en el de imagen de marca suele ser muy importante, especialmente si éste afecta a determinados clientes, si es de larga duración, y si no se gestiona correctamente.

Este sistema permite además optimizar los recursos de la empresa en cuanto a equipos de mantenimiento, pudiendo ajustarlos según las previsiones, así como los desplazamientos, ya que se pueden realizar ciertos movimientos en base a las mismas para responder más rápido y evitar largos viajes innecesarios.

[1] Computación en la nube

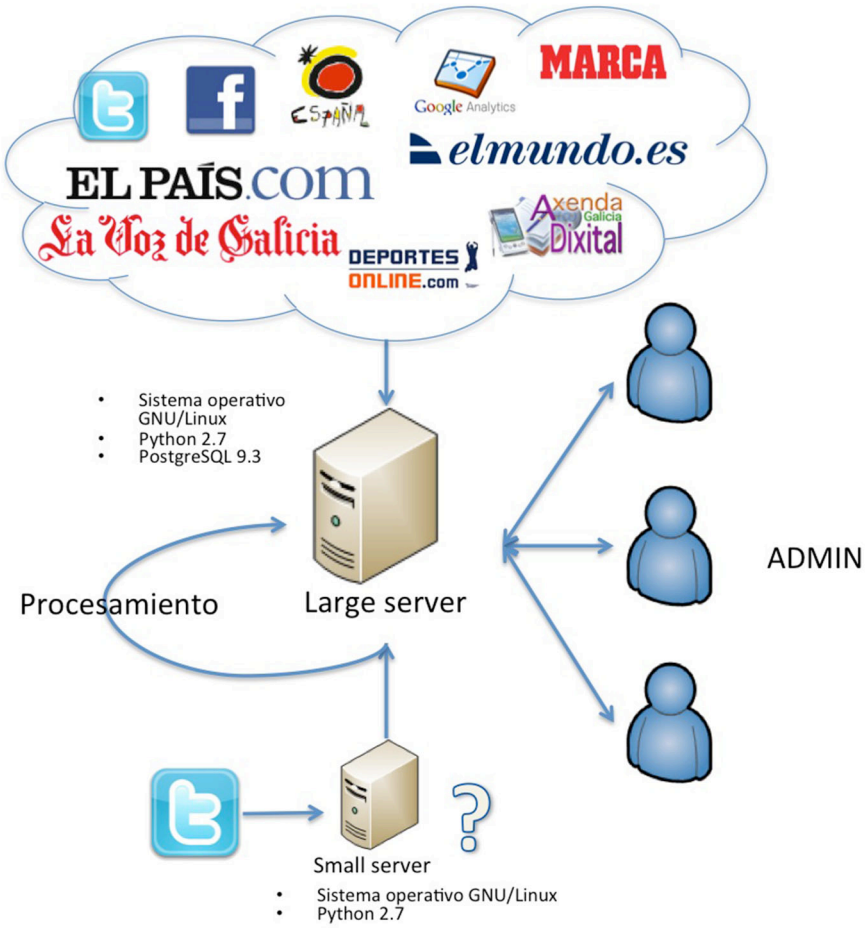

Figura 2.1: Arquitectura general de la solución

Capítulo 3

Previsión de riesgos sociales

3.1 Introducción

El sistema desarrollado en este TFM permite a las empresas de distribución eléctrica conocer potenciales riesgos antes de que estos sucedan, lo que les da un margen de maniobrabilidad del que no dispondrían de otro modo. Actualmente, Unión Fenosa Distribución se basa en la opinión de los expertos para tratar de prever posibles incidentes en base a meteorología, estructura de la red, características de los equipos, etc. Es un método poco contrastado que no les funciona del todo mal, pero no tiene en cuenta la componente social del problema, ni tampoco manejan los grandes volúmenes de información de los que disponen.

3.2 Estudio previo

Para abordar la previsión de riesgos sociales, el primer paso que se dio fue hablar con los expertos de Unión Fenosa. Con su experiencia, y la información que nos facilitaron, se trataron de buscar las fuentes de datos de Internet de las que se pudiera extraer la información más relevante para el proyecto. Se elaboró una lista inicial que comprendía `http://www.spain.info` para eventos de carácter general; `http://www.marca.es` para eventos deportivos; `http://www.acadomia.es` para disponer del calendario escolar; `http://www.guiadelocio.com` para conciertos; y Facebook eventos (`https://www.facebook.com/events`) para eventos creados por la propia sociedad.

Una vez fijadas una serie de fuentes de las que extraer información, se pasó a estudiar de qué forma se podría utilizar para estimar la gravedad de los riesgos que podrían conllevar los incidentes que sucedieran en los eventos extraídos. Para ello, se analizaron trabajos que tratan temas similares como se describe en la siguiente sección.

3.2.1 Trabajos relacionados

Uno de los problemas con los que hemos encontrado similitudes con la estimación de la audiencia, y que más ha sido trabajado, es el de la estimación de demanda de plazas de estacionamiento.

Algunos de los trabajos analizados se basan en la parte económica del problema, y otros, simplemente en tratar de encontrar el número de plazas que deberían de estar disponibles para satisfacer las necesidades de un área concreta. En ellos se utilizan distintos tipos de datos como históricos reales sobre población, número de plazas de parking disponibles, ocupación media, datos del censo, puestos de trabajo en la zona, etc., para la estimación de la demanda. Los modelos desarrollados para esta estimación están basados en regresión. En estos trabajos resulta más sencillo conseguir datos concretos, ya que se dispone de ellos al tratarse de información pública y en muchos casos se cuenta con históricos. En nuestro caso, los eventos suelen ser muy dispares y aleatorios, les afectan variables muy diferentes cada vez, y por ello no encontramos la forma de aplicar este tipo de técnicas a nuestro problema. A continuación se listan algunos de los trabajos que se analizaron:

- Madsen, Edith and Mulalic, Ismir and Pilegaard, Ninette: A model for estimation of the demand for on-street parking. Munich Personal RePEc Archive, 2013.
- GEOK K. KUAH. Estimating Parking Demand for Mixed-Use Developments Subject to TSM Ordinances. ITE Journal. Febrero de 1991.
- Choy Peng NG, Dadang Mohamad MA'SOEM. The Development Of Model Estimation To Determine Paring Needs At LRT Stations In Suburban Area. Proceedings of the Eastern Asia Society for Transportation Studies, Vol. 5, pp. 877 - 890, 2005.
- Nicholas J. Garber, Hua Wang. Demand for Commercial Heavy Truck Parking on Interstate Highways: A Case Study of I-81 in Virginia. TRB 2003 Annual Meeting.
- Christopher F. Dumas, John C. Whitehead, James H. Herstine, Robert B. Buerger, Jeffery M. Hill. Estimating Peak Demand for Beach Parking Spaces Under Capacity Constraints. 2005, working paper University of North Carolina-Wilmington.

3.2.2 Conclusiones

Tras el análisis de los trabajos anteriores, y junto con el consorcio del proyecto, se decidió integrar la parte social del problema como una serie de variables más en el modelo general de estimación de severidad en el que se incluyen otras variables relacionadas con la meteorología o la previsión de demanda en base a históricos. Para ello, se utilizan sólo dos variables cualitativas, una de ellas dividida en tres, que son la cantidad de audiencia de un evento por un lado, y la sensibilidad de la misma por otro. A continuación se describen los criterios utilizados para definirlas:

- **Audiencia**, la cantidad de gente que va a tener noción del evento y que es susceptible de cambiar su opinión si sucede un incidente eléctrico en él (muy poca, poca, normal, bastante, mucha). Se divide en:

  - **In-situ**: Aquélla que asista en persona al evento, pague entrada por él, se desplace, etc.

  - **Directa**: Aquélla que vea el evento ya sea a través de TV, radio, Internet... Es importante ya que hay eventos para los que hay que pagar aunque no se asista en persona, como los abonos de televisión para ver partidos de fútbol.

  - **Indirecta**: Aquélla que aunque no asista al evento en el momento en el que suceda, tenga noción del mismo, y a posteriori pueda estar afectada por lo que suceda en él.

- **Sensibilidad**: el grado de afección estimado si algo interrumpe el evento (muy poca, poca, normal, bastante, mucha). Esto dependerá del tipo de evento, del precio que hayan pagado los asistentes por presenciarlo, etc.

Para su estimación, ya que no se dispone de estos valores, además de la experiencia de los expertos, se han tratado de utilizar dos métodos. El primero de ellos consiste en usar el aforo de los lugares donde se celebra en los casos en los que este dato está disponible. Con esta información, y catalogando la importancia del visitante (por ejemplo, no será lo mismo que Real Madrid o FC Barcelona visiten Balaidos a que lo visite el colista), se estima la afluencia. A los eventos que se sabe de antemano que van a tener mucha repercusión se les asignan automáticamente los máximos valores posibles en todas las variables.

El otro método consiste en utilizar datos históricos disponibles de asistencia media en webs como `http://www.european-football-statistics.co.uk/` o el foro `http://foro.delcelta.com/` y tomar sus valores medios como el valor estimado. Estos datos sólo los hemos encontrado con facilidad para partidos de fútbol de ligas importantes.

Dado que es un tema muy complejo, finalmente lo que se decidió es categorizar los eventos, y asignar valores por defecto a cada categoría para cada valor, pudiéndose luego visualizar los próximos eventos en un panel de control desde el que se permite modificar manualmente su estimación por parte de los expertos. Algunas fuentes ya disponen de distintas categorías que son automáticamente capturadas por el sistema, y otras se definieron en el transcurso del proyecto. Hay un caso especial, que es Facebook eventos, en el que sí que se dispone de un dato numérico con los usuarios que han marcado que sí que van a asistir a cada evento, y se utiliza como referencia para la estimación de audiencia de los mismos.

3.3 Diseño y desarrollo

Una vez elaborada la lista inicial de fuentes de datos junto con los expertos de Unión Fenosa, se pasó a estudiar las opciones para extraer su información. Prácticamente al mismo tiempo, se comenzó a diseñar un modelo de datos (ver anexo E) para almacenar toda la información extraída de una manera uniforme y estructurada. Dado que las tecnologías ya estaban acordadas con el resto de integrantes del consorcio (Python y PostGreSQL[1]), nos centramos en la búsqueda tanto de librerías para el acceso a la base de datos desde Python, que fue Pyscopg2 por tratarse de la más extendida y mantenida; como de alternativas para la extracción de datos a partir de código fuente HTML. Para las fuentes que disponen de API[2] que ofrecen los datos en formato JSON, se utiliza la librería nativa de Python para el tratamiento de los mismos. Las peticiones HTTP[3] se realizan mediante la librería también nativa llamada urllib. Las alternativas disponibles para la extracción de datos fueron, entre otras:

1. Scrapy, framework de código abierto para realizar scraping
2. BeautifulSoup4, librería para realizar scraping en su versión 4
3. Consultas de XPATH[4] a través de la librería lxml

Finalmente la última fue la elegida, debido principalmente a la experiencia del autor, a la flexibilidad que ofrece, y al mejor rendimiento que se obtiene frente a las alternativas. Por contra, puede resultar un poco más compleja de utilizar.

Para cada una de las fuentes de las que se extraen datos desde el código HTML, se ha realizado un análisis exhaustivo de su navegación para poder construir automáticamente todas las URLs[5] necesarias para el volcado de los datos relevantes, siempre obviando elementos duplicados.

Analizando las fuentes, pronto se apreció que había algunas cuyos eventos no disponían de geoposición, para las que se decidió utilizar información adicional como dirección, ciudad, código postal, etc., para geocodificarla con la mayor precisión posible a través de la API que provee Google Maps. El error cometido en este paso no es demasiado importante, ya que una subestación (nivel al que se trabaja en este sistema), puede comprender una ciudad completa, por lo que no se buscó ninguna alternativa mejor, ni se realizaron comparaciones con otras APIs, al escaparse la complejidad de este problema del ámbito de este TFM.

Como hemos comentado en el capítulo anterior, se diseñó una arquitectura con un nodo principal en el que se recogen y procesan todos los datos, además de usarse también para al-

[1] Structured Query Language
[2] Application Programming Interface
[3] Hypertext Transfer Protocol
[4] XML Path Language
[5] Uniform Resource Locator

bergar el servidor web con la interfaz diseñada. Se decidió crear un scraper distinto para cada fuente, con partes comunes como el acceso a la base de datos o el uso de la API de Google Maps. Dependiendo de la fuente, se utilizaron distintas técnicas de extracción de datos como se detalla a continuación. En alguna de ellas hubo que realizar algo de ingeniería inversa ya que utilizan sesiones, identificadores, peticiones AJAX[6], etc., que no son tan fácilmente identificables y reproducibles. Notar que algunas de las fuentes seleccionadas inicialmente fueron sustituidas por otras, ya fuera por estar las últimas más completas, más actualizadas, o con la información mejor estructurada:

1. **Agenda digital de Galicia** (`http://agenda.galiciadigital.com`), es una web muy completa con eventos culturales de Galicia como conciertos, actuaciones, cursos, fiestas, exposiciones, etc. Dado que los datos se encuentran en formato HTML, se creó un scraper que utiliza lxml y consultas xpath para extraer la información deseada, y se ejecuta cada 24 horas ya que la frecuencia de actualización de la web es baja.

2. **Deportes Online** (`http://www.deportesonline.com`), se trata de una web en la que se encuentran la inmensa mayoría de eventos deportivos de muy distintas categorías (incluso ligas inferiores), muy bien estructurada y actualizada. El robot también utiliza lxml y xpath para extraer datos del HTML. Se lanza cada 24 horas ya que su frecuencia de actualización es diaria y los eventos son publicados con bastante antelación.

3. **Facebook eventos** (`https://www.facebook.com/events/`), sección de eventos de una de las redes sociales más grande del mundo. Los eventos están geoposicionados y tienen información sobre la audiencia de los mismos, lo que resulta muy útil para el proyecto. Por contra, no están categorizados. El scraper usa la API de JSON de Facebook y se lanza cada hora dado que los usuarios participan activamente en ella.

4. **Spain.info** (`http://www.spain.info/`), otra web bastante completa de eventos de todo tipo con mucha información, contenido geoposicionado en algunas ocasiones, múltiples categorías, y actualizada con un gran horizonte temporal. Dado que no dispone de API, se extraen los datos del HTML usando de nuevo consultas xpath. La frecuencia de lanzamiento de su robot es de 24 horas también.

Estas fuentes cubren prácticamente la totalidad de la información que se puede encontrar en Internet que resulta relevante para el proyecto en cuanto a eventos. Se podrían añadir más, pero nos hemos encontrado con mucha información redundante y ruido que distorsionan el propósito del proyecto.

El diseño modular del sistema permite controlar los scrapers de manera independiente, lanzarlos, detenerlos, modificarlos, e incluso añadir nuevas fuentes sin afectar al funcionamiento de las demás. Para la tarea propiamente de control, existe un proceso de tipo demonio encargado de ella en base a la información de cada robot que hay en la base de datos, cuyos valores

[6] Asynchronous JavaScript And XML

pueden ser modificados desde la interfaz. Mediante un sistema de monitorización, en base a los eventos recogidos y a su información, se pueden visualizar geoposicionadas las zonas más comprometidas en el horizonte temporal seleccionado. También se pueden ver estadísticas de los datos recogidos. En la figura 3.1 podemos ver un ejemplo y en el anexo D podremos encontrar más pantallazos de todo el sistema de monitorización. Los marcadores de colores son los eventos, los distintos colores indican la severidad estimada, y los recuadros azules son las subestaciones del sistema de suministro.

Todas las tecnologías mencionadas están recogidas en detalle en el anexo F.

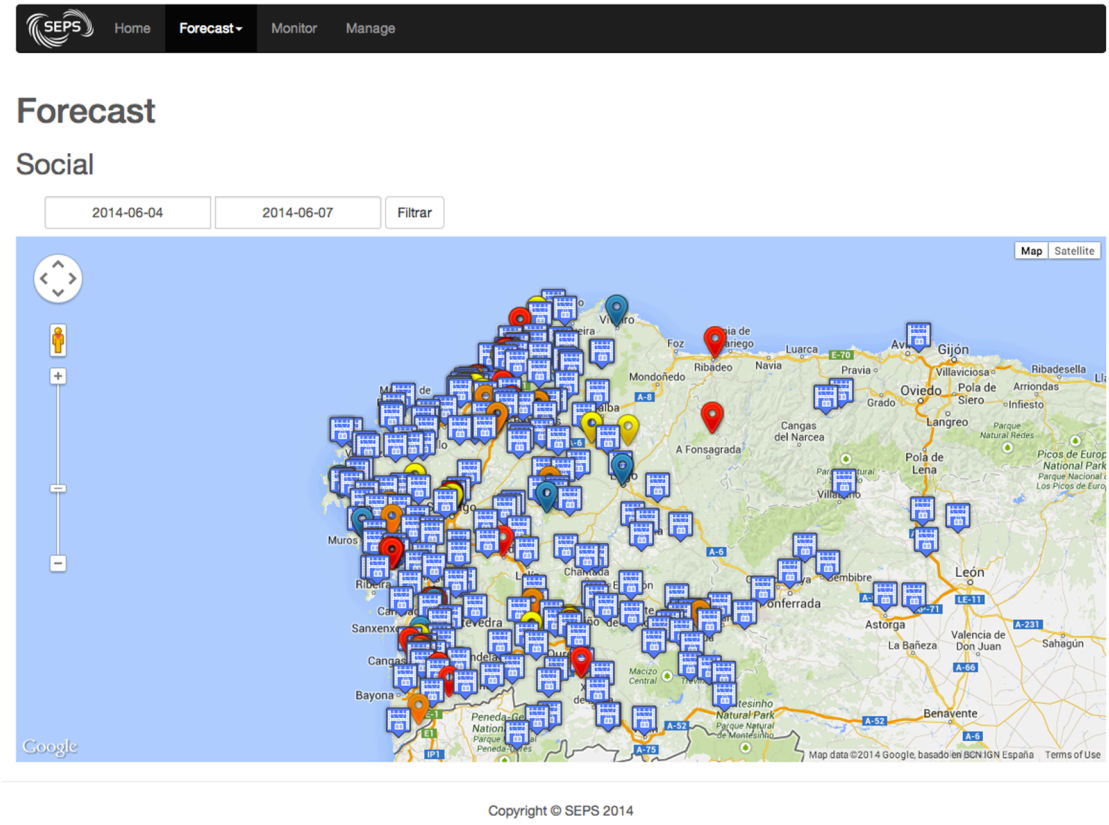

Figura 3.1: Pantalla para la visualización de eventos geoposicionados

3.4 Pruebas

Las pruebas de la parte de recogida de datos se realizaron para asegurar la robustez de los scrapers. Para ello, fueron diseñados con una serie de módulos comunes (uso de la base de

datos, registro de sucesos, acceso a otros servicios como Google Maps...) dejando sólo la parte diferencial en cada uno de ellos. De esta forma, se evita replicar código, reduciendo su mantenimiento, y se facilita encontrar posibles errores.

Todos los robots tienen un modo verboso cuya función es guardar en un fichero de registro todo lo que están capturando sin almacenar nada en la base de datos para evitar errores e inconsistencias. Una vez que un robot era desarrollado, se dejaba en funcionamiento en este modo verboso durante las horas que fuera necesario, y se analizaba su salida hasta asegurarse de que éste funcionara correctamente y capturara los datos deseados. Esto se contrastaba en la medida de lo posible realizando las búsquedas de forma manual, y comparando los resultados encontrados con los capturados por el robot. Cuando se daba por bueno, el scraper se incorporaba a la base de datos central con el fin de que el demonio diseñado para encargarse de la gestión de la ejecución de todos los robots lo utilizara en producción.

Una vez en producción, los scrapers, además de guardar los datos en la base de datos, disponen de un fichero de registro en el que van indicando los errores que han ido surgiendo durante su ejecución. Para ello, todos están preparados para recuperarse de posibles fallos capturando las excepciones generadas con los detalles de lo que las ha provocado.

Capítulo 4

Análisis de opinión social

4.1 Introducción

En este apartado se describe el trabajo realizado para la fase del análisis de la repercusión de un incidente eléctrico una vez que éste ya ha sucedido en base a la información recopilada en redes sociales, prensa, y otras fuentes de Internet.

Dicha fase está dividida a su vez en dos partes. Por un lado, la recogida de información, que se ha realizado de forma muy similar a la anteriormente descrita para la previsión de riesgos, y por otro, el propio análisis de la opinión social.

4.2 Estudio previo

Como hemos comentado, esta fase consta de una parte de recogida de datos muy similar a la descrita en el capítulo anterior, y dado que se han utilizado las mismas técnicas y tecnologías para su desarrollo, sólo entraremos en detalle para comentar las fuentes seleccionadas.

Para analizar todos estos datos y extraer algún tipo de conclusión útil para el proyecto, se plantearon diferentes posibilidades.

En un extremo, teníamos la opción de tratar de aplicar algún tipo de análisis semántico a los contenidos, pero dada la complejidad del tema, la menor fiabilidad de sus técnicas por la ambigüedad del lenguaje, y la limitación de recursos del proyecto, decidimos que se escapaba del alcance del mismo.

En el otro extremo, existía la posibilidad de realizar simplemente estadísticas de aparición de ciertas palabras, correlación de las mismas, acceso a las webs y perfiles de redes sociales de la empresa, pero consideramos que esto se quedaba demasiado corto para un proyecto de este calibre. Por ello se decidió ir a un punto intermedio, y además de realizar estadísticas, se

propuso aplicar análisis de sentimiento sobre los contenidos extraídos. Con esta información se podría tener una noción bastante fiable de lo que la sociedad siente hacia la empresa, especialmente hacia la gestión de incidentes en sus instalaciones una vez que estos hubieran sucedido.

Notar que los contenidos de los que se pretendía extraer conclusiones consistían en mensajes cortos que la sociedad publica en Internet (tweets[1] y pequeños comentarios de noticias), así como noticias publicadas en diarios online que en su título contuvieran ciertas palabras clave.

4.2.1 Trabajos relacionados

Antes de comenzar a probar técnicas, se realizó un estudio del estado del arte, analizando trabajos de grupos de investigación dedicados al tema. Ésta es la parte que con diferencia ha tenido más carga "investigadora" de todo el proyecto. A continuación se introducen los distintos trabajos divididos en tres grupos según la temática y el papel que han jugado para el desarrollo de este TFM.

Trabajos generales

Dado que el autor de este trabajo no tenía mucho conocimiento sobre el tema, comenzó a buscar trabajos de carácter más general en los que se tratara de analizar el comportamiento de la humanidad y sus comunicaciones en Internet. En ellos se habla de los términos más utilizados dentro del mundo del estudio del comportamiento humano, de organización en la red, incluso de tratar de dar otro enfoque a la inteligencia artificial.

- Saif M. Mohammady and Tony (Wenda) Yang, Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. arXiv:1309.6347v1 [cs.CL] 24 Sep 2013. Trabajo en el que se estudian los emails intercambiados entre géneros. Se analizan distintos tipos de emails, y realizando estadísticas de palabras utilizadas, se pueden extraer conclusiones de cuáles son las más utilizadas entre hombres y mujeres, cuando se dirigen al mismo o al sexo opuesto, y en distintas situaciones (odio, amor, cartas de suicidio...).
- Javier Toret, @Datanalysis15m, Antonio Calleja, Óscar Marín Miró, Pablo Aragón, Miguel Aguilera. Tecnopolítica: la potencia de las multitudes conectadas. El sistema red 15M[2], un nuevo paradigma de la política distribuida. IN3 Working Paper Series, Internet Interdisciplinary Institute. 2013. Trabajo un poco más general de cómo extraer conclusiones de redes sociales como Twitter.
- Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz, Francisco Sanz,

[1] Mensaje de la red social Twitter`http://www.twitter.com`

[2] Denominación que recibió el movimiento social de protesta del 15 de mayo de 2011

Fermín Serrano, Cristina Viñas, Yamir Moreno. Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study. PLoS ONE 6(8): e23883, 2011. Similar al trabajo anterior, se utilizó para coger ideas de cómo buscar mejor la información a extraer dentro de la voluminosa cantidad de datos que albergan estas redes sociales.

- González Bedia, M., y García Carrasco, J. (2006). Arquitecturas emocionales en Inteligencia Artificial : una propuesta unificadora. [Versión electrónica]. "Teoría de la Educación : educación y cultura en la sociedad de la información", 7 (2), 156-168. Pequeño trabajo de un profesor del I3A que ha aportado algunas de sus ideas a este TFM que ayuda a ver desde otro punto de vista el comportamiento del ser humano.

Trabajos sobre análisis de sentimiento

En la siguiente fase, se buscaron trabajos un poco más específicos, en los que directamente se tratara el tema del análisis de sentimiento. Hablan de los conceptos necesarios para comprenderlo, de qué factores influyen, de las limitaciones del problema, y de distintas técnicas y recursos existentes.

- Yi-jie Tang, Hsin-Hsi Chen, Mining Sentiment Words from Microblogs for Predicting Writer-Reader Emotion Transition. LREC 2012: 1226-1229. Trabajo que analiza mensajes en el sitio de microblogging llamado Plurk con el fin de intentar prever cómo cambia el sentimiento de los intervinientes en una conversación. Este estudio se basa en los mensajes que disponen de emoticonos. Aparecen conceptos nuevos para el autor de este TFM que luego serán utilizados, como etiquetado POS[3].

- Rui Fan, Jichang Zhao, Yan Chen and Ke Xu, Anger is More Influential Than Joy: Sentiment Correlation in Weibo. arXiv:1309.2402. Septiembre 2013. En este trabajo se estudia cómo se suceden las emociones, y se llega a la conclusión de que el enfado de un usuario provoca más enfado en el resto de individuos que se comunican con él. El estudio se realiza en la red social Weibo, y se utilizan estadísticas con mensajes que contienen emoticonos.

- Proceedings of the Workshop on Semantic Analysis in Social Media, 13th Conference of the European Chapter of the Association for Computational Linguistics. Se trata de otro estado del arte en el que se habla de todos los conceptos que entran en juego en el análisis de sentimiento en redes sociales.

- Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and

[3] Part Of Speech

Technology. Vol 61 Issue 12. Trabajo sobre un algoritmo que además de detectar la polaridad de un texto, intenta identificar la intensidad de su sentimiento. Está basado en machine learning.

- Bo Pang and Lillian Lee, Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. Vol. 2, No 1-2 (2008) 1–135. Estado del arte en el que por primera vez el autor de este TFM lee sobre SentiWordNet, que es un recurso léxico que se utiliza en este proyecto.

Trabajos sobre SentiWordNet

Una vez contextualizado el problema, y tras asimilar la evolución que habían tenido las diferentes soluciones planteadas a lo largo de los últimos años, se pudo observar que la tendencia actual era el uso del recurso léxico llamado SentiWordNet, recomendado por el ponente de este TFM. Anteriormente habían tenido más peso otros métodos como la utilización de técnicas heurísticas o sistemas de aprendizaje. También se estudia en alguno de ellos como la combinación de estas técnicas puede conseguir mejores resultados que usarlas individualmente.

- Andea Esuli and Fabrizio Sebastiani, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06). Trabajo sobre la creación del recurso léxico utilizado para este trabajo. Consiste en un diccionario de palabras en inglés con su polaridad positiva y negativa en distintos contextos o conjuntos de sinónimos (synsets).

- Marco Guerini, Lorenzo Gatti, Marco Turchi. Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. arXiv:1309.5843v1 [cs.CL] 23 Sep 2013. Un resumen muy bueno de por qué usar SentiWordNet y las distintas formas de hacerlo, evaluándolas, lo que determinó la forma en que se usó el recurso en este proyecto.

- Monalisa Ghosh, Animesh Kar. Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using Sentiwordnet 3.0. International Journal of Engineering Research & Technology. Vol.2 - Issue 9 (September - 2013). Es un trabajo muy reciente, clave para el desarrollo de este TFM, en el que se describe la forma en la que nosotros utilizamos SentiWordNet para evaluar la polaridad de los distintos tweets o comentarios recogidos de las diversas fuentes, y cómo su versión 3.0 es mejor que las anteriores.

- Aurangzeb khan, Baharum Baharudin, Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs. International Journal of Computer Science & Emerging Technologies. Vol 2, No 4 (2011). Trabajo sobre cómo utilizar dicho recurso léxico tratando de utilizar información contextual para enlazar frases.

- Kerstin Denecke, Using SentiWordNet for Multilingual Sentiment Analysis. Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference. Se trata de un trabajo sobre cómo utilizar SentiWordNet, que está en inglés, para extraer la opinión en textos en otros idiomas. La propuesta, que obtiene buenos resultados, consiste en pasar primero el texto por un traductor al inglés, técnica que utilizamos en este TFM.

- Alaa Hamouda, Mohamed Rohaim. Reviews Classification Using SentiWordNet Lexicon. The Online Journal on Computer Science and Information Technology (OJCSIT) Vol. (2) - No. (1). Otro trabajo más de cómo utilizar SentiWordNet para evaluar automáticamente revisiones de películas.

4.2.2 Conclusiones

Tras analizar todos estos textos, algunos otros, y diversas webs relacionadas con el tema, vimos que teníamos dos alternativas. La primera era realizar un analizador de sentimiento que se basara en el entrenamiento, por lo cual habría que proporcionar un corpus inicial en castellano, evaluarlo, etc. La segunda era usar un recurso léxico que fuera completo, contrastado y actualizado, con el que se pudiera extraer la polaridad de cada palabra. Finalmente se decidió implementar un analizador semántico utilizando el recurso léxico SentiWordNet, ya que según lo que se puede extraer de los trabajos científicos anteriormente mencionados, es el más completo, con el que mejores resultados se obtiene, y el mejor para el rendimiento. Además, es el más adecuado para el ámbito de este proyecto, ya que se van a analizar muchos mensajes cortos sin relación entre ellos.

Como SentiWordNet está hecho en inglés, se buscaron alternativas en castellano, pero sus corpus no eran tan completos ni tan precisos. Por ello, se encontró la alternativa de traducir las frases al inglés mediante herramientas estándar del mercado, para posteriormente analizarlas, y por las conclusiones del trabajo [19], los resultados obtenidos son muy buenos.

4.3 Diseño y desarrollo

Como hemos comentado anteriormente, previo al análisis de sentimiento, se han de obtener los contenidos relevantes para el proyecto. Analizando el volumen de tráfico de las distintas fuentes posibles, pronto vimos que a nivel nacional la referencia en cuanto a contenido es Twitter. Además, proporciona una API en streaming que permite capturar contenido en tiempo real, y filtrarlo geográficamente, por palabras clave, o por usuario, justamente lo que necesitábamos. También decidimos añadir algunas fuentes de noticias de prensa, buscando ciertas palabras clave en los titulares o en el propio contenido.

Las técnicas utilizadas para la extracción de datos son las mismas que para el apartado anterior, con alguna ligera diferencia, pero dado que ya están descritas, no vamos a profundizar

de nuevo en ellas.

1. **Twitter** (`http://www.twitter.com`), la red social más importante y con mayor tráfico generado a nivel nacional. Se capturan en tiempo real tweets geoposicionados en Galicia y tweets por palabras clave susceptibles de estar relacionados con temas relevantes para el proyecto. También se capturan tweets de las cuentas más importantes de Unión Fenosa. Se utiliza la API de streaming que devuelve los datos en formato JSON.

2. **La Voz de Galicia** (`http://www.lavozdegalicia.es`), uno de los diarios más importantes a nivel local. Se tratan de obtener noticias con respecto a la empresa y su gestión de incidentes eléctricos. Se extraen los datos directamente del HTML y su scraper se ejecuta periódicamente cada hora.

3. **El País en Galicia** (`http://ccaa.elpais.com/ccaa/galicia.html`), es un diario nacional similar al anterior, pero que dispone de sección para Galicia, y que además admite comentarios en las noticias, por lo que se puede extraer una idea más concreta sobre cuánto se ha hablado de una noticia, y si se ha hecho positiva o negativamente.

4. **El Mundo en Galicia** (`http://www.elmundo.es/galicia.html`). Similar al anterior.

5. **Marca** (`http://www.marca.com`), web de noticias deportivas a nivel nacional en la que han aparecido algunos incidentes eléctricos que han afectado a eventos deportivos. Dispone de comentarios. Como en las anteriores, se extraen los datos directamente del HTML, y el robot se ejecuta periódicamente cada hora.

Además de estas fuentes, disponemos de acceso a la cuenta de Google Analytics de Unión Fenosa, de la que obtenemos el volumen de visitas que tiene su web. Con esta información podemos detectar picos de acceso a la misma tras un incidente que sabemos que se ha dado, lo que será, según los expertos, un indicador claramente negativo de cara a la empresa.

Para la parte de análisis de sentimiento y opinión hemos definido un proceso en cadena que dispone de los elementos que se pueden ver en la figura 4.1.

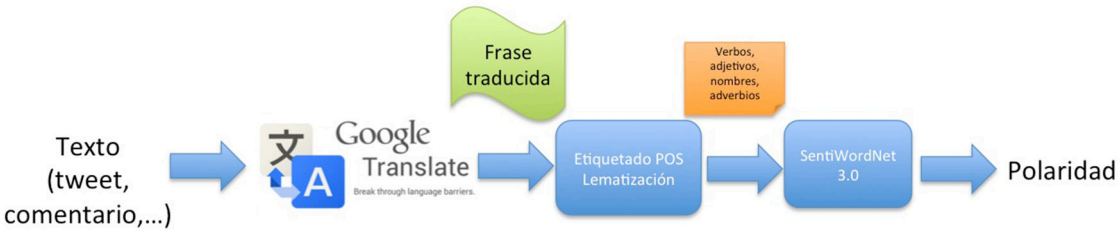

Figura 4.1: Proceso de evaluación de la polaridad de un texto en nuestro sistema

Para implementarlo, se ha desarrollado un paquete en Python dividido en diferentes módulos, cada uno de ellos ocupándose de una de las funciones.

Por un lado, se utilizan los servicios de Google Translate para traducir la frase al inglés. Se eligió este servicio ya que tras probar otros similares (Yahoo, Babel, ...), fue el que consideramos que realizaba mejor las funciones e introducía menos ruido. Para ello, se utiliza una API en JSON a través de peticiones en las que se le envía la frase original, el idioma original, el idioma destino, y devuelve la traducción codificada en UTF8[4].

A continuación fue necesario integrar un clasificador de la parte de la expresión (POS tagger), ya que el recurso léxico SentiWordNet necesita la información de si una palabra es un adjetivo, un nombre, etc. Dado que la tecnología que se iba a utilizar en el proyecto era Python, se buscó cuál era el más indicado tanto en términos de precisión como de rendimiento, y tras visitar diferentes artículos y webs, el elegido fue [27]. Forma parte de un paquete de Python y se llama textblob-aptagger. Para utilizarlo simplemente se crea un instancia del objeto para etiquetar, y se llama a un método con la frase original el inglés. El resultado es una lista de tuplas, una por cada palabra de la frase, con la función que ocupa cada una en el texto. A continuación, mediante la librería textblob, se lematiza cada uno de los términos, ya que el recurso léxico los tiene registrados de esta forma.

Una vez tenemos esta información, entra en juego SentiWordNet. Este recurso contiene la polaridad de muchas palabras en distintos contextos. El módulo que se desarrolló carga este recurso desde un fichero en memoria en una estructura de tipo diccionario para un fácil y rápido acceso. A este módulo se le pasa una palabra con la función que ocupa en dicha frase, y el módulo calcula la polaridad de dicha palabra con esa función. En caso de que la palabra realice una función irrelevante o no esté en el diccionario, devuelve un 0. Para calcular la polaridad de una palabra, se utiliza el promedio de la polaridad de todos los contextos en los que aparece pesada por la frecuencia de aparición de cada significado para dicha palabra. Según las mediciones del trabajo [16], es un método que introduce poco error, muy rápido, y sencillo. Hay otras alternativas como coger la polaridad del significado más frecuente, la correspondiente con la mediana, etc.

Una vez disponemos de la polaridad de todas las palabras, se realiza un promedio de la suma de los resultados obtenidos dividido por el número de términos encontrados en el diccionario, y ésta será la polaridad final del texto.

Este proceso es ejecutado para los tweets, comentarios y noticias recogidos las últimas horas, y contrastando el número de positivos con el de negativos, se obtiene una visión general del sentimiento global de la sociedad respecto a lo que concierne a este proyecto.

Dado que la información recogida en este apartado está filtrada por geolocalización y por aparición de palabras, éstas fueron elegidas cuidadosamente junto con los expertos de Unión Fenosa. Además, se realizaron varias iteraciones para identificar los términos que más ruido

[4] 8-bit Unicode Transformation Format

capturaban con el fin de establecer los parámetros más adecuados, como introducir coocurrencias de palabras en lugar de términos sueltos (por ejemplo no capturar todos los mensajes y noticias que contengan la palabra luz, sino la secuencia "corte de luz"). Los criterios con los que se definieron y seleccionaron las palabras y los usuarios clave para la captura de tweets, comentarios, noticias, etc., se pueden encontrar detallados en el anexo G.

El método implementado es lo suficientemente rápido como para evaluar mensajes en poco tiempo, siendo la carga de SentiWordNet en memoria el proceso más costoso de la ejecución del módulo. En el anexo H podemos ver una comparación de los resultados obtenidos por el analizador implementado en este TFM frente a otras alternativas que podemos encontrar.

En la herramienta piloto, los robots están constantemente capturando y analizando datos para ofrecer una serie de estadísticas con el fin de ayudar al operario a comprender mejor la opinión de la sociedad. La figura 4.2 muestra una captura de la pantalla desde la que los operarios pueden ver información de los tweets, noticias, y comentarios capturados y analizados tanto positiva como negativamente, así como observar las estadísticas de acceso a la web de Unión Fenosa.

4.4 Pruebas

La recogida de datos fue probada y validada de forma similar a como se hace en el capítulo anterior, por lo que no vamos a entrar de nuevo en ella.

El proceso de validación del análisis de sentimiento fue un poco más laborioso y a medida. Por un lado, se identificaron otras herramientas disponibles para realizar pruebas como `http://www.mrtuit.com`. Por otro, se probaron otros recursos léxicos similares a SentiWordNet en castellano como SentiWordNet-BC. La opción de crear un recurso propio fue totalmente descartada, ya que es con diferencia la parte más complicada y costosa de este tipo de sistemas.

Para ambos escenarios, se seleccionaron una serie de frases, y fueron probadas en varios sistemas para analizar tanto el resultado obtenido, como el tiempo que llevaba la obtención del mismo. Este estudio se encuentra detallado en el anexo H, y como podemos observar, los resultados obtenidos por el sistema implementado son bastante buenos.

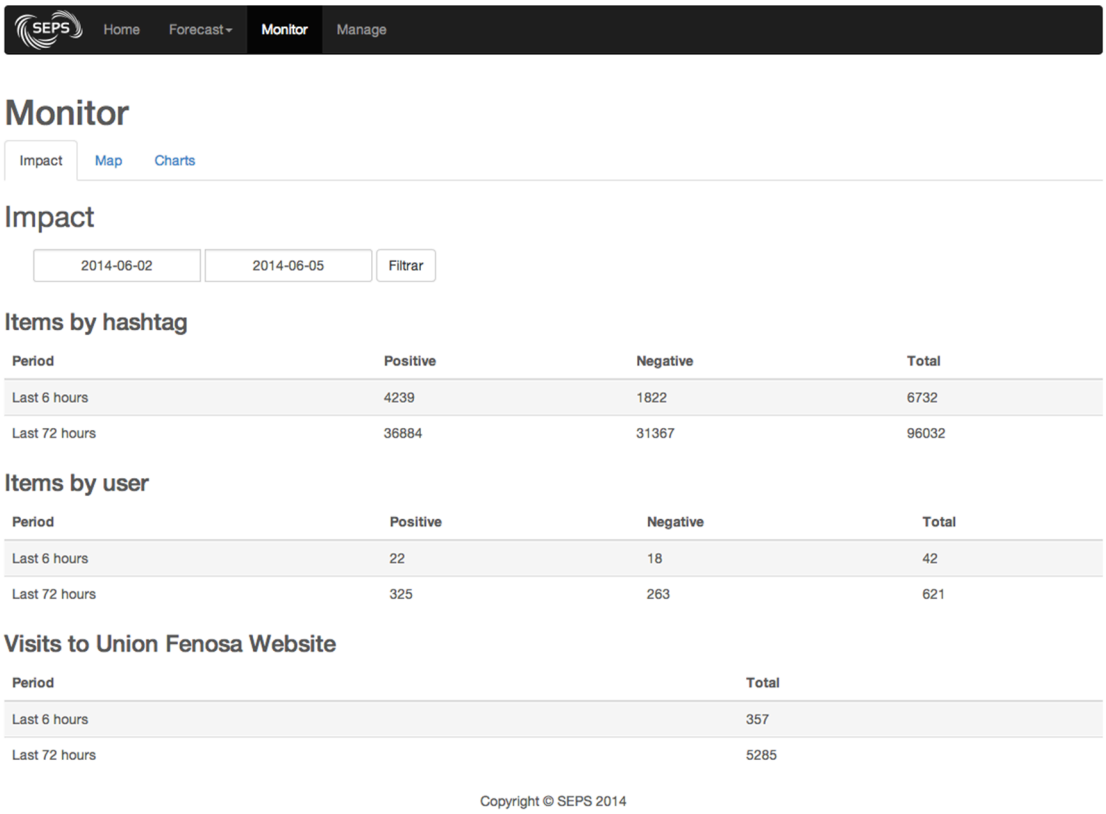

Figura 4.2: Pantalla de análisis del impacto de los elementos capturados recientemente

Capítulo 5

Conclusiones

5.1 Cumplimiento de objetivos

El principal objetivo de este TFM consistía en dotar al proyecto SEPS de una componente social basándose en información de Internet. Para ello, se han desarrollado dos vertientes. Por un lado, la previsión posibles incidentes en situaciones de mayor importancia social, y por otro, la evaluación del sentimiento de la sociedad hacia la empresa y su gestión de dichos incidentes. Para ello, era necesario el estudio del estado del arte de sistemas y técnicas similares, con el fin de que lo que se desarrollara fuera útil, actual, y estuviera lo suficientemente contrastado para poder utilizar los resultados obtenidos.

Dada la complejidad del tema, la propia empresa Unión Fenosa Distribución no tenía puestas en esta parte demasiadas expectativas, pero tras los resultados obtenidos, se ha mostrado más que satisfecha con el trabajo realizado, e incluso sorprendida de lo que se puede llegar a extraer de Internet en relación a temas que a priori pueden parecer muy específicos.

Además, los resultados de las pruebas realizadas sobre el sistema ratifican el correcto funcionamiento del mismo y por tanto el de los diferentes subsistemas que lo componen. Se ha incidido en la parte de monitorización del sistema, con el fin de facilitar a los usuarios tanto la comprensión de su funcionamiento como la búsqueda de la información deseada.

5.1.1 Objetivos desestimados

El único objetivo desestimado en el marco de este TFM, ha sido la realimentación del sistema de previsión de riesgos con los resultados del sistema de análisis de opinión social. No se ha desechado por completo, pero por el momento se ha dejado a un lado. Esto es debido a que es un tema muy complejo, y automatizar este proceso puede llevar a una espiral de resultados incorrectos al basarse en valoraciones sociales que no son del todo fiables para riesgos futuros, situación que no interesa en absoluto a la empresa.

5.2 Ampliaciones y mejoras para el futuro

Como acabamos de comentar, la principal ampliación que se puede realizar en este sistema, y que además se está estudiando actualmente, es la realimentación del sistema de previsión de riesgos con los resultados del análisis de opinión. Esta funcionalidad convertirá al sistema en más autónomo todavía, reajustando las valoraciones que se hagan a priori sobre posibles incidentes en eventos con lo que ha sucedido en el pasado.

También se podrían implementar otras mejoras como:

- **Fuentes:** aunque actualmente se recogen datos de las fuentes más representativas para el proyecto, siempre se pueden añadir más. No obstante, tampoco es conveniente incrementar demasiado el volumen de información, ya que si se añade demasiado ruido puede ser contraproducente.
- **Alertas:** ahora mismo el piloto está diseñado para estar observando constantemente y que el operario lo consulte cuando desee, pero se podría desarrollar un sistema de alertas de manera que avisara a los responsables de las escuadras de técnicos para movilizarlos a las zonas más susceptibles de sufrir incidentes, con el fin de reducir su impacto, subsanándolos en el menor tiempo posible.
- **Sistema de aprendizaje:** en algunos de los trabajos estudiados como en [16], se habla de que con sistemas de aprendizaje se puede mejorar algo la precisión del análisis de sentimiento de una frase, y es una mejora que se podría valorar para su incorporación.

5.3 Incidencias

La principal incidencia que se ha encontrado en el desarrollo de este proyecto ha sido la dependencia de otros socios del consorcio. Aunque las funcionalidades desarrolladas en el contexto de este TFM no necesitaban de gran interacción con el resto de paquetes de trabajo, todas las que sí que lo hacían han sufrido algún pequeño retraso, aunque en ningún caso se ha tratado de nada importante ni ha afectado por el momento seriamente a la planificación inicial.

El aspecto que sí que ha repercutido algo en el desarrollo de este TFM ha sido la integración. Dicha tarea tendría que haber comenzado a principios del año 2014 con el fin de ocupar toda la anualidad para disponer de tiempo para terminarla y optimizarla lo máximo posible. Esta tarea lleva un retraso de unos 3 meses, que tampoco va a afectar a la entrega del proyecto, pero ha obligado al autor de este TFM a desarrollar por su cuenta la interfaz del sistema referente a la parte social de SEPS, en lugar de desarrollar los módulos correspondientes dentro de la aplicación piloto general. Por otro lado, esto es una ventaja de cara al consorcio, ya que se ha adelantado trabajo que podrá ser reaprovechado en el resto de paquetes.

5.4 Valoraciones

El autor de este TFM concluye que el desarrollo de la parte social del proyecto “SEPS” ha sido satisfactorio.

El prototipo desarrollado cumple las expectativas para las que se fue diseñado, y las herramientas conseguidas utilizan técnicas y tecnologías que se encuentran en el estado del arte.

La parte de previsión de riesgos es la más acotada, los datos de los que se dispone provienen de fuentes bien conocidas que son las que más información relevante contienen para el proyecto, y las tecnologías utilizadas son estables y llevan siendo desarrolladas durante años. La forma en la que estos datos son recogidos y analizados ha sido definida conjuntamente con los expertos de Unión Fenosa Distribución para que los resultados se ajusten lo máximo posible a sus conocimientos. Con este trabajo el autor ha aprendido bastante sobre la extracción de datos a partir tanto de código HTML como a través de APIs, manejando llamadas complejas, simulando el comportamiento de un usuario a través de un navegador, utilizando cookies y sesiones, etc.

La parte de análisis de sentimiento es la más novedosa y experimental. Por ello, durante el transcurso de este TFM, se realizó un estudio del estado del arte, analizando trabajos recientes sobre el tema, probando y adaptando diferentes herramientas para evaluar sus resultados, y eligiendo finalmente la que mejor se adaptó a las necesidades del proyecto en cuanto a fiabilidad y rendimiento. Esta última parte ha contenido la vertiente más investigadora del proyecto, y es con la que el autor del TFM considera que más ha aprendido, abordando un problema totalmente nuevo para él partiendo de otros trabajos, artículos, conferencias, etc., publicados en medios importantes del ámbito investigador que ni siquiera conocía al comenzar esta titulación.

Parte II

Anexos

Anexo A

Propuesta del proyecto

En este apartado se incluyen dos fragmentos de la memoria técnica presentada al Ministerio de Economía y Competitividad para el desarrollo del proyecto.

El primero de ellos consiste en la introducción del mismo, que puede ayudar a la contextualización de este TFM, y comienza a partir de la página siguiente.

INTRODUCCIÓN

El suministro eléctrico es un elemento esencial en la vertebración y el crecimiento socio-económico de la sociedad actual. Las exigencias que deben satisfacer las redes eléctricas en cuanto a la continuidad y seguridad de suministro son crecientes, en un entorno en el que la necesidad de una mayor integración de las energías renovables en la red forma parte de la necesaria sostenibilidad medioambiental del sector.

La gestión que han de realizar las empresas distribuidoras debe estar alineada con estos requerimientos de mejora continua en la continuidad y seguridad de suministro e integración eficiente de fuentes de energías renovables y distribuidas en la red.

En la consecución de este crucial objetivo existen importantes retos técnicos que los gestores de las redes de distribución deben afrontar para conseguir que la importante evolución tecnológica hacia redes eléctricas inteligentes que están experimentando las redes actuales vaya acompañada de crecientes estándares de calidad.

Dentro de dichos retos, sin duda, de los más complicados de gestionar están algunos factores extrínsecos que por su naturaleza no es posible corregir como es el caso de la variabilidad con la que la meteorología impacta sobre las redes eléctricas.
En este sentido, se puede indicar que la meteorología afecta a la gestión de las redes eléctricas fundamentalmente en dos ámbitos:

- El número de incidencias que se producen en las redes, lo que tiene una correlación directa sobre la continuidad de suministro (ver Fig. 1).

- Los flujos de demanda y generación conectada a la red, que repercuten de forma sustancial en los márgenes de seguridad de suministro dado que determinan posibles restricciones en la explotación.

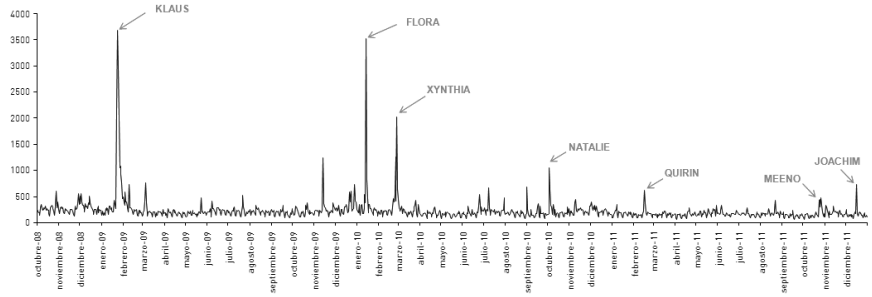

ILUSTRACIÓN 1. NÚMERO DE INCIDENCIAS DIARIAS EN LA RED DE UNIÓN FENOSA DISTRIBUCIÓN EN LOS AÑOS 2009-2011 INDICANDO, EN ALGUNOS CASOS, LOS NOMBRES DE CICLO GÉNESIS EXPLOSIVAS ACAECIDOS.

El objetivo fundamental del proyecto es desarrollar un sistema experto que a partir de la información de previsión meteorológica y previsión de demanda y de generación, genere unos índices de probabilidad de ocurrencia de incidencias y restricciones en la red de distribución y a partir de información sobre el mercado afectado y eventos relevantes, el impacto social de dichas incidencias.

Para ello se elaborarán algoritmos de previsión de incidencias basados en la información histórica y las previsiones meteorológicas, junto con otros para identificar la severidad y la repercusión social de los incidentes sobre la red a partir de aspectos intrínsecos (potencia y clientes afectados, tiempos de reposición, etc.) y extrínsecos (sensibilidad social, eventos relevantes, etc.).

De esta forma se podrían establecer estrategias preventivas de explotación de las redes que permitan gestionar de forma anticipada condiciones meteorológicas extremas o excepcionales y minimizar su impacto sobre la continuidad y seguridad de suministro.

Los beneficios fundamentales que proporcionará su utilización pueden resumirse, por tanto, como:

- Mejora de la calidad de suministro
- Reducción del impacto social / mediático de los incidentes y restricciones de red
- Maximización integración de renovable en la red

El objetivo final del proyecto es el desarrollo una herramienta informática que implemente este sistema experto de previsión de incidentes y la severidad de los mismos. Se espera con esta herramienta dar un salto tecnológico importante sobre las soluciones comerciales ya existentes. Las principales novedades o mejoras que proporcionará se detallan a continuación:

- Número y precisión de fuentes de información que van utilizar
- Detalle y precisión en las previsiones tanto en las probabilidades de ocurrencia como en la severidad
- Estimación precisa de la potencia afectada por los incidentes basado en una previsión de la demanda y generación distribuida.
- Ayuda en la integración de energías renovables
- Uso de información desestructurada, en parte procedente de Internet para la recopilación de eventos de relevancia social
- Utilización de información de redes sociales, como Twitter o Facebook, para la medida de la relevancia social de los eventos.

La presencia en este consorcio de una de las principales empresas distribuidoras de electricidad del país permitirá que el sistema desarrollado pueda ser ya probado por expertos en la materia y en condiciones realistas de operación.

Posteriormente, la herramienta informática que implemente el sistema desarrollado estará preparada para su comercialización a cualquier empresa de distribución eléctrica con redes tanto en el territorio nacional como el extranjero. Es importante destacar que el coste de adaptación al mercado internacional es pequeño debido a la universalidad de las tecnologías utilizadas en las redes de distribución eléctricas.

Consorcio

Para el desarrollo de este ambicioso proyecto, se requiere contar con un importante equipo multidisciplinar que aúne experiencia técnica y profesional muy específica con capacidad de I+D en campos muy diversos como la previsión meteorológica, modelos de redes eléctricas de potencia, modelos predictivos complejos, sistemas de inteligencia artificial y gestión del conocimiento o modelos de análisis espacial.

Con este fin, se crea un consorcio formado por empresas y grupos universitarios de referencia en cada uno de los campos a cubrir.

En primer lugar, el proyecto cuenta con la participación de Unión Fenosa Distribución (Grupo Gas Natural Fenosa - GNF), una de las principales distribuidoras eléctricas del país y con importante presencia internacional del grupo, aportando toda su experiencia y conocimiento de la actividad, fundamental para guiar las necesidades y objetivos durante el desarrollo de todo el proyecto.

Se cuenta también con otra empresa líder en su sector y con gran proyección internacional, como Telvent, experta en sistemas de información para sector energético, transportes o meteorología y medio ambiente.

Por otra parte, otras dos empresas expertas en análisis científico de datos y desarrollos informáticos participan en el consorcio. Por un lado, AIA (Aplicaciones en Informática Avanzada), compañía con una importante experiencia tanto en modelado de la red eléctrica, como en previsión de demanda y generación. Por otro lado, SCIEN Analytics, aportando su conocimiento en modelos de previsión, sistemas de información geográfica y análisis espacial, y desarrollo e integración de sistemas de información.

El consorcio se completa con grupos de investigación de tres centros universitarios. En primer lugar, el equipo de la Universidad de Girona colaborará con su conocimiento en gestión de información, sistemas expertos y descubrimiento de patrones. La Universidad Carlos III contribuirá con su experiencia en investigación aplicada a diversos modelos para el sistema eléctrico y en análisis de incidentes. Por último, la Universidad de Zaragoza aportará su conocimiento investigador en análisis de redes complejas y extracción de conocimiento a partir de fuentes desestructuradas procedentes de Internet.

1. ESTADO DEL ARTE

Los problemas de toma de decisiones consisten en la selección en la una acción *apropiada* de entre un conjunto de posibles alternativas de acuerdo a la *información disponible*. El objetivo de los sistemas informáticos de ayuda a la toma de decisiones (DSS, *decision support systems*) es proporcionar al responsable de la decisión todo el soporte necesario para facilitar la selección de la mejor opción. En consecuencia, los DSS abarcan numerosas tareas: (i) adquisición e integración de datos/mediciones/indicadores; (ii) selección y resumen de información; (iii) evaluación automática de alternativas de acuerdo a criterios de optimalidad; (iv) presentación de resultados; (v) almacenamiento de histórico; etc.

Las técnicas de Inteligencia Artificial tienen aplicación principalmente en las tareas de selección y resumen de información y de evaluación de alternativas en los DSS. Así, los DSS inteligentes determinan automáticamente qué datos de entrada proporcionan información relevante para la decisión y estiman, en función de esta información, la plausibilidad de diferentes situaciones, el impacto de las mismas y/o el coste que repercutiría en la organización responsable.

1.1 METEOROLOGÍA

La predicción meteorológica es uno de los campos de la ciencia en los que más se está avanzando en estos últimos años. Prestigiosas instituciones norteamericanas y europeas están consiguiendo unos resultados con un nivel de detalle impensable hace décadas.

Los métodos de predicción meteorológica se pueden dividir en métodos estadísticos y en modelos numéricos de predicción (MNP). Los primeros se basan en establecer correlaciones entre series temporales de variables meteorológicas que permitan extrapolar valores más allá del rango de las medidas utilizadas para determinar un conjunto de coeficientes.

Los MNP utilizan las ecuaciones fundamentales de la física atmosférica para predecir el valor de las variables en el futuro, partiendo para ello de un buen conocimiento del estado previo de la atmósfera, las denominadas condiciones iniciales, y de las condiciones laterales, de frontera y de contorno, que son aquellas que se establecen en el área que está siendo modelada. Cuentan con el

inconveniente de que los cálculos a realizar requieren una potencia de cálculo importante, pero son la única manera de basar los pronósticos en el conocimiento físico de los procesos en operación

La predicción meteorológica en España está basada en modelos que corren sobre superordenadores instalados en las oficinas de la Agencia Española de Meteorología y en instituciones similares en toda Europa como el Centro Europeo de Predicción a Medio Plazo (ECMWF). Basados en los millones de datos (temperatura, humedad, presión, dirección y velocidad de viento, etc) provenientes de diferentes localizaciones se generan las predicciones a nivel global, nacional y regional.

El tiempo sin embargo, puede variar mucho sobre distancias de pocos kilómetros. Los Servicios Meteorológicos no se encargan de generar predicciones adaptadas a la escala temporal y espacial requerida por diferentes usuarios a nivel masivo, ya que ello implicaría unas capacidades de cómputo y sobre todo un trabajo de adaptación a las necesidades de miles de usuarios que no está incluida entre sus prioridades y objetivos.

No existe un único modelo numérico cuya fiabilidad y exactitud resulte la más apropiada en todas las situaciones que podemos encontrar. Existen aproximaciones en las se realiza la incorporación en el sistema de diferentes modelos numéricos para con ello lograr un mejor pronóstico inicial.

Empleando la técnica de predicción por “ensembles” se consigue modelar un análisis de la sensibilidad a las condiciones iniciales que presentan los modelos atmosféricos. Siendo la idea general el proporcionar un conjunto de diferentes condiciones iniciales (perturbaciones) al modelo, para cada una de las cuales el modelo genera un pronóstico diferente.

Esta predicción por ensembles supone una estrategia abordada en los últimos años por el ECMWF (European Centre for Medium Range Weather Forecasting/Centro Europeo de Pronóstico a Medio Plazo), en la que se realizan un número de predicciones simultáneas (50) de la evolución atmosférica posible, variando en cada una de ellas ligeramente las condiciones iniciales de las que parte la predicción. De esta forma, se obtiene no sólo pronóstico a través de valores concretos, sino que se llega a conocer la oscilación que puede presentar esa variable en cada momento Con ello se pueden considerar las probabilidades de ocurrencia de las distintas evoluciones previstas, algo especialmente necesario cuando se predice a más de 72 horas.

Otro aspecto importante es la definición de las diferentes estrategias para los diferentes intervalos de predicción, que se mueve entre las 0 horas hasta las siguientes 72 horas generalmente (pudiendo ser hasta 15 días si se requieren predicciones para hacer estimaciones a largo plazo). Diferentes estrategias de predicción requieren diferentes fuentes de información, como muestra la siguiente ilustración, a medida que la predicción se aleja del presente la observación va perdiendo peso mientras que los modelos de predicción meteorológica van ganando peso.

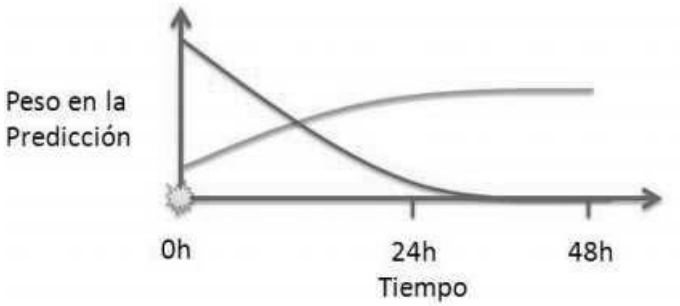

ILUSTRACIÓN 2. LA LÍNEA ROJA MARCA LA IMPORTANCIA O PESO DE LA OBSERVACIÓN EN LA PREDICCIÓN Y LA VERDE LA DE LOS MODELOS NUMERICOS DE PREDICCIÓN.

La predicción meteorológica es la fuente de información más importante que usarán los modelos de previsión de demanda y generación y de previsión de averías.

Por ejemplo, la predicción de la demanda de energía se ve afectada por diversos aspectos estacionales como la hora del día, día de la semana, o la laboralidad pero con diferencia el factor que más afecta a su precisión es el pronóstico meteorológico. Valores típicos de esta dependencia nos indican que hasta un 90% del error en la predicción de la demanda es generado directamente por errores en la predicción meteorológica.

Con respecto a las energías renovables, se están aplicando diferentes tecnologías de predicción. Hay numerosos estudios comparativos para el sector eólico donde se están combinando modelos físicos, estadísticos e híbridos. En el sector de la energía solar se está modelizando la irradiación solar directa con resultados bastante limitados para días cubiertos o poco estables.

En el apartado de las averías el gran reto es integrar todas las fuentes de información meteorológica disponible, observación y predicción, de forma que puedan ayudar a una toma de decisiones ágil y efectiva.

1.2 PREVISIÓN DE DEMANDA Y GENERACIÓN

Previsión de generación

- **Eólica**: Se utilizan varios métodos para la predicción a corto plazo de la generación eólica. Los más simples se basan en la climatología o en los promedios de los valores históricos de producción. Ellos pueden ser considerados como los métodos de pronóstico de referencia, ya que son fáciles de implementar, así como punto de referencia en la evaluación de los enfoques más avanzados. El más popular de estos métodos de referencia es, sin duda el de *persistencia*. Esta predicción simplificada - comúnmente conocida como 'lo que ves es lo que obtienes "- afirma que la generación de energía eólica futuro será el mismo que el último valor medido. A pesar de su aparente sencillez, este método podría ser difícil de superar para las horas que se avecinan en un rango de 4-6 horas de antelación. Los enfoques más avanzados para predicciones a corto plazo de energía eólica requieren como entrada pronóstico de variables meteorológicas. Entre ellos se diferencian en la forma en las predicciones de las variables meteorológicas se convierten en las predicciones de producción de energía eólica, a través de la llamada *curva de potencia*. Tales métodos avanzados se dividen tradicionalmente en dos grupos. El primer grupo, conocido como el planteamiento **físico**, se centra en la descripción del flujo del viento alrededor y dentro del parque eólico, y utilizar la curva de potencia del fabricante, para proponer una estimación de la producción de energía eólica. Paralelamente, el segundo grupo, conocido como método estadístico, se centra en la captura de la relación entre las predicciones meteorológicas y las medidas (posiblemente históricos) y la potencia a través de modelos estadísticos cuyos parámetros deben estimarse a partir de los datos, sin hacer ninguna hipótesis sobre los fenómenos físicos.
- **Solar.** Algunas técnicas para predecir la energía solar se centran en los modelos numéricos de predicción meteorológica que requieren una potencia de cálculo considerable. A pesar de esto, estos modelos son actualmente incapaces de hacer pronósticos precisos de densidad

de las nubes, su formación y su movimiento. Otros en cambio se han centrado en la predicción de la radiación solar a partir de imágenes satelitales de movimientos de las nubes, pero estos modelos no pueden prever la disipación o formación de nubes, ni la opacidad de la nube a la radiación solar. Los modelos más antiguos, que datan de 30 años atrás, emplean modelos estadísticos de series temporales con observaciones meteorológicas como entrada. Estos estudios han utilizado técnicas clásicas, así como nuevas tecnologías tales como las redes neuronales y los modelos híbridos. Ver por ejemplo Reikard, quien llevó a cabo una comparación de varios modelos de series temporales. Casi la totalidad de esta investigación fue reportado en la etapa de desarrollo inicial.

Previsión de demanda

Las metodologías para la previsión de demanda a corto plazo se dividen en tres grandes grupos (Al-Hamadi, 2004):

- los modelos que son independientes de los cambios de clima (modelos no-meteorológicos sensibles),
- los modelos en función de los cambios de clima (modelos sensibles al cambio del tiempo), y
- los modelos híbridos.

Las metodologías basadas en Redes Neuronales Artificiales han sido ampliamente utilizados con, en cierta medida, resultados satisfactorios. Sin embargo, las opciones de diseño no siempre se justifica plenamente y con frecuencia los modelos tienen un nivel de alta complejidad (Hippert, 2001).

El tipo más importante de variable incluida en el vector de entrada (IV) es el de la serie temporal de los últimos tiempos de serie de la previsión variable (Hippert de 2001, Senjyu de 2002, Papalexopoulos, 1994 y Khotanzad, 1994). Otras variables, de carácter auxiliar que se utilizan y, al no estar directamente relacionados con el consumo de electricidad, que suelen estar representados por las funciones del tipo sinusoidal o binaria, con el objetivo de ayudar a la red neuronal artificial para detectar las características periódicas del comportamiento de la carga (Drezga, 1998 y Fidalgo, 1999). La mayoría de los métodos de predicción hacen uso de técnicas estadísticas o algoritmos de inteligencia artificial tales como la regresión, redes neuronales, lógica difusa, y sistemas expertos. Dos de los métodos, los llamados de uso final y de enfoque econométricos son ampliamente utilizados para la previsión a medio y largo plazo. Una variedad de métodos, que incluyen el llamado de día similar, varios modelos de regresión, series temporales, redes neuronales, algoritmos de

aprendizaje estadístico, lógica difusa y sistemas expertos, se han desarrollado para la predicción a corto plazo.

1.3 PREVISIÓN DE INCIDENTES

Los análisis de fiabilidad y estimación de fallos son una parte crítica en la gestión de grandes infraestructuras distribuidas como las redes eléctricas. Disponer de las herramientas adecuadas para realizar estos análisis de la forma más precisa posible permite anticiparse a posibles contingencias y planificar en distintos horizontes las operaciones de mantenimiento necesarias, optimizando costes y recursos al tiempo que se minimizan las afecciones para los clientes y su repercusión.

Métodos teóricos aplicables

Por la importancia de estas tareas, existe una importante literatura con diferentes enfoques para abordar dicha problemática.

Las aproximaciones más clásicas parten de los trabajos sobre análisis de supervivencia y tasas de fallo en equipos. Estos se basan en el ajuste individual de modelos paramétricos para estimar el conteo o tasas de fallo de distintos componentes del sistema (p.ej. mediante modelos Weibull, Poisson, etc.), basándose en datos históricos. Variantes más refinadas incluyen asunciones adicionales sobre la dependencia de las variables explicativas y la tasa de fallo o función de riesgo, p.ej. los modelos de riesgos proporcionales o de Cox.

Estos métodos, en su aplicación habitual, presentan algunos inconvenientes, como la necesidad de realizar asunciones sobre la distribución que siguen los tiempos entre fallos o de reparaciones. Tampoco recogen satisfactoriamente las posibles dependencias entre determinados tipos de fallos y la evolución temporal del sistema. Por último, los índices que proporcionan son estimadores puntuales (p.ej. de la tasa o del tiempo promedio hasta el siguiente fallo), sin información acerca de la variabilidad o incertidumbre que permita evaluar los riesgos de manera más adecuada.

Para solventar la carencia de información sobre la variabilidad, se suelen aplicar técnicas basadas en simulación estocástica (p.ej. métodos Monte-Carlo) para obtener estimaciones de las distribuciones de los índices de fallo.

Por otra parte, para la captura de la evolución temporal del sistema y dependencias entre fallos de sus componentes se han propuesto, entre otras técnicas, métodos basados en cadenas de Markov con distintas variantes, así como en sistemas de autómatas celulares.

También cabe destacar el uso de métodos bayesianos para estimar las probabilidades de fallo en contextos en los que no hay información suficiente a priori de todas las variables implicadas. Estos métodos permiten reajustar el modelo conforme se producen nuevos eventos u observaciones, incorporando de una forma bien definida el nuevo conocimiento al modelo.

Por último, hay que apuntar que estas técnicas son en general aplicables cuando se trabaja con variables continuas y asumiendo dependencias lineales. Para casos en los que aparecen variables categóricas o dependencias no lineales, se han propuesto también diferentes alternativas, que pasan por el uso de redes neuronales, modelos de regresión logística o técnicas difusas.

Un paso más allá en la construcción de modelos predictivos consistiría en analizar las dependencias temporales entre eventos, considerando los eventos no de forma independiente sino en forma de secuencias de eventos. Este enfoque busca descubrir la existencia de patrones que permitan pronosticar la aparición de futuras incidencias. La búsqueda de secuencias de patrones es un problema de minería de datos con amplias aplicaciones. En el ámbito de la calidad de potencia eléctrica, el enfoque de la minería de secuencias de datos se ha introducido con diferentes objetivos. Por ejemplo, (CJ Kim, 2004) presenta una herramienta basada en análisis de tendencias en eventos para aislar parámetros que pudieran predecir y anticipar faltas en sistemas de distribución eléctrica. (Benner, 2004) recoge numerosos ejemplos de faltas incipientes que predecían el fallo de un elemento a partir de largas campañas de monitorización. (Yasar, 2008) exploró la viabilidad de utilizar técnicas como las transformadas wavelet y Hilbert-Huang para implementar minería de datos en la monitorización de faltas incipientes. El grupo de investigación eXIT de la Universidad de Girona tiene experiencia en el descubrimiento de patrones en secuencias de eventos (principalmente huecos de tensión) registrado en alimentadores de distribución y en el estudio de registros de incidencias de red (Quiroga, 2012).

Recientemente, se han desarrollado varios proyectos de investigación, con participación de socios del consorcio, con el objetivo de automatizar el procedimiento de evaluación de riesgos en las redes

de distribución eléctricas, que consiste en determinar qué componentes son más susceptibles de fallo, con vistas a la optimización de las tareas de reparación y mejora (CENIT ENERGOS 2009/12). Se combinaban técnicas de decisión analítico-jerárquicas y lógica borrosa para determinar cuál es el riesgo físico de red asociado a un equipo a partir de mediciones físicas y estimaciones cualitativas sobre su estado actual y las repercusiones de un posible fallo. Una vez calculado el riesgo de red para cada uno de los componentes de la red, puede establecerse una lista ordenada de equipos para priorizar las actuaciones de mantenimiento. Los resultados obtenidos a nivel de investigación son prometedores y son uno de los motivos de esta propuesta.

Sistemas comerciales

Existen en el mercado sistemas informáticos para el análisis, modelado y estimación de fallos y duración operativa de equipos e infraestructuras genéricas.

Dentro de estos sistemas generales se puede citar la *suite* de herramientas de ReliaSoft (www.reliasoft.com), entre las que se incluyen aplicaciones para modelar el tiempo de vida de distintos componentes de un sistema, usando diferentes técnicas según su naturaleza y la disponibilidad de datos históricos, así como aplicaciones para la simulación y análisis de ciclos de vida, fallos y consecuencias a nivel de daños en los equipos, estado operativo del sistema y duración de las afecciones hasta su restablecimiento. En base a estos análisis permite planificar programas de mantenimiento.

Otro conjunto de aplicaciones similares es el proporcionado por Isograph (www.isograph-software.com). Esta empresa ofrece distintas herramientas para análisis de riesgos de incidentes en un sistema o infraestructura, análisis y seguimiento de causas, o simulación de escenarios y evaluación de operatividad del sistema.

Respecto a herramientas específicas para redes eléctricas, la empresa ETAP (www.etap.com) dispone entre sus soluciones para gestión integral de redes de un módulo para análisis de fiabilidad del sistema. A partir de modelos físicos de la red y de sus distintos elementos, y del estado operativo de cada componente, calcula índices de fiabilidad para cada uno de ellos, así como medidas del riesgo en términos de energía, duración y clientes afectados, a partir de distintos escenarios.

La empresa SKM (www.skm.com) también ofrece un programa para evaluación del riesgo de incidentes en sistemas de potencia, identificando componentes críticos y estimando el coste de la posible falla en pérdida de energía, reparación y tiempo hasta la reposición.

En otro ámbito relacionado, también se pueden citar las especificaciones e implementaciones de procedimientos de análisis de fiabilidad y riesgo de equipos de Telcordia (http://telecom-info.telcordia.com). Esta compañía ha desarrollado una serie de procedimientos y técnicas para auditoría y previsión de riesgos en equipos electrónicos y redes, en las que se modelan y caracterizan físicamente distintos elementos, permitiendo controlar la fiabilidad de un sistema y elaborar planes de actuación.

Las herramientas enumeradas basan su funcionamiento en uno de dos enfoques. O bien el modelado se realiza a partir de datos históricos, o bien se basan en modelos de estado físico de cada elemento y de la red en su conjunto, pero no integran de manera natural y simple ambos tipos de información.

Por otro lado, la introducción de variables de contexto (p.ej. datos meteorológicos) para explicar y prever la posible ocurrencia de fallos no está generalizada en los sistemas específicos para el sector eléctrico.

Finalmente, el impacto de los posibles fallos suele estimarse con indicadores de energía dejada de suministrar y duración del incidente hasta la recuperación del sistema En algún caso se ofrece también una valoración de posibles usuarios afectados. Sin embargo, en ninguno de ellos se proporciona una visión global que combine toda esta información para priorizar actuaciones según múltiples criterios. Por supuesto, tampoco son capaces de inferir cuál puede ser la repercusión por derivadas como el impacto en imagen y reputación social de la empresa.

La estimación de la severidad de los incidentes en la red eléctrica tiene una doble vertiente. Por un lado se tiene el impacto directo que provoca en la propia red de la compañía distribuidora, ya sea en términos de coste por los elementos que han podido verse dañados o en la potencia que se ha dejado de suministrar. Por otro lado se tiene el impacto social que la falta de electricidad durante un tiempo determinado provoca, tanto desde el punto de vista de los problemas que la falta de un bien tan importante crea en los usuarios domésticos o industriales como desde el punto de vista del daño que causa a la imagen de la compañía distribuidora.

1.4 ANÁLISIS DE IMPACTO SOCIAL A TRAVÉS DE DATOS DE INTERNET

Para analizar el impacto social, contamos con una potente herramienta y fuente de datos: Internet. Hoy en día, el conocimiento de lo que ocurre en la Red es imprescindible para una imagen completa de la realidad en sus diferentes facetas: social, económica, política, científica, académica, etc.

El desarrollo de Internet ha permitido, por primera vez en la historia, disponer de información sobre las opiniones de millones de usuarios, e incluso extraer conocimiento sobre lo que están haciendo prácticamente en tiempo real. Para ello se emplean técnicas encuadradas en lo que se define como Sentiment Analysis u Opinion Mining.

Estas técnicas se pueden combinar con análisis de redes sociales para estudiar cómo se organizan estructuras y polos de opinión, cómo se difunden y se solapan en la misma red y con otras redes/medios (Tanev, 2011).

Otro enfoque, más orientado a la previsión, utiliza la información generada en la red por los usuarios para tratar de explicar la evolución de ciertas variables, construyendo modelos que permitan analizar y hacer previsiones de tendencias en indicadores de negocio (ventas, demanda,…). Aunque no hay un término global extendido, podríamos denominarlo “Social Media Based Forecasting”. Se trata de aprovechar el conocimiento agregado de la masa de usuarios de los Social Media (Harvey, 2009).

Podemos dividir las tecnologías necesarias para este tipo de estudios en cuatro categorías: primero aquellas, de software y estandarización, que permiten la recogida de datos de la red; en segundo lugar el software en el que se apoya el análisis de datos, en tercer lugar las tecnologías matemáticas y lógicas que se usan para el análisis de datos, y finalmente en un apartado propio los recursos propiamente computacionales, de acceso al hardware y organización de los batches de trabajo, que emplean las anteriores.

En la cuestión de recogida de datos el estado del arte presenta ciertas dificultades: son escasas las organizaciones que ofrecen sus datos en un formato directo para su uso por el ordenador, y cuando lo hacen no suele ocurrir que sigan estándares de construcción de ontologías (RDFs, etc), por lo que es necesario programar desde cero el acceso a la mayoría de las fuentes de datos. En muchos casos

hay un rudimentario espacio semántico definido por palabras clave proporcionadas por los redactores del documento, o incluso subrayadas al vuelo en el propio texto (hashtags).

No obstante, el número de fuentes accesibles sigue creciendo, y hay algunos sistemas, como la dbpedia -que tiene un fuerte grupo en español, véase es.dbpedia.org-, dedicados en exclusiva a hacer accesible programáticamente la información que se encuentra liberada en la red.

También en este lado positivo, es interesante resaltar que se ha superado el bache que constituían las páginas web basadas en javascript con consultas asíncronas (AJAX), en las que los datos a consultar no aparecen directamente en el documento HTML que se descarga en navegador. La solución ha consistido en herramientas de scraping que son capaces de actuar como navegadores y ejecutar el código javascript correspondiente.

Dentro del software, pero ya en la cuestión de herramientas de análisis, hay que mencionar la aparición de librerías con propósitos específicos, sea para tratamiento de RDFs o de entradas JSON, (el parsing de JSON se entiende ya a nivel de servidor también), o para cuestiones más matemáticas como el análisis de redes (ej networkx o igraph). Y el auge de herramientas generales como R y librerías Python como Numpy/Scipy.

Con estas herramientas, y como se ha dicho, con desarrollos propios de los investigadores, se aplican algunas de las técnicas más potentes de data mining y aprendizaje automático. Así tenemos el uso de técnicas exploratorias: Análisis multivariable, Técnicas de Reducción Dimensional: Análisis de componentes principales, análisis factorial, correspondencias múltiples). Y técnicas para modelado y predicción: Árboles de decisión, Redes Neuronales, Support Vector Machines, Random Forests, Redes Bayesianas, Regresión lineal y no lineal, Regresión probabilística, Regresión no paramétrica. Modelización de Series Temporales: ARIMA, SARIMA, GARCH, Holt-Winters).

Por último, en el nivel de plataforma física, hay que comentar el uso ya estándar de organización de tareas con Hadoop y las ideas de paralelización basadas en MapReduce, y a nivel de administración la aparición de sistemas de nube abiertos (cloudstack) que permiten organizar los clientes necesarios bien para las tareas de búsqueda y scraping, bien para el análisis. La alternativa a la paralelización basada en nube es el uso de clusters locales y sistemas de memoria compartida; es posible un cluster de proceso que albergue hoy en día varios teraflops de datos en RAM.

1.5 BIBLIOGRAFÍA

1. Application of Improved BP Network in Failure Forecasting. Zhi Gang Li, Bo Wei Shi. 2012, Advanced Materials Research, Vols. 490 - 495, págs. 373-377.

2. Fuzzy reliability estimation using Bayesian approach. Wu, Hsien-Chung. 3, s.l. : Elsevier, 2004, Journal of Computers and Industrial Engineering, Vol. 46, págs. 467 - 493.

3. On Distribution Asset Management: Development of Replacement Strategies. Miroslav Begovic, Joshua Perkel, Nigel Hampton, Rick Hartlein. Johannesburg : s.n., 2007. IEEE PES PowerAfrica.

4. Evaluación de riesgos en redes eléctricas: una aproximación difusa. Juan Gómez-Romero, Antonio Berlanga, José M. Molina, Ángel Ramos Gómez. Valencia : s.n., 2010. Actas del III Simposio sobre Lógica Fuzzy y Soft Computing. págs. 123-130.

5. A Study of Log-Logistic Model in Survival Analysis. Gupta, Ramesh C., Akman, Olcay y Lvin, Sergey. 4, s.l. : Wiley, 1999, Biometrical Journal, Vol. 41, págs. 431–443.

6. Regression Models and Life-Tables. Cox, David R. 2, 1972, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 34, págs. 187–220.

7. Casteren, Jasper van. Power System Reliability Assessment using the Weibull-Markov Model. Göteborg : Chalmers University of Technology, 2001.

8. Brown, Richard. E. Electric Power Distribution Reliability. 2nd Edition. s.l. : CRC Press, 2008.

9. Analysis of Survival Data under the Proportional Hazards Model. Breslow, Norman E. 1, 1975, International Statistical Review, Vol. 43, págs. 45-47.

10. Failure rate prediction with artificial neural networks. Bevilacqua, Maurizio, y otros. 3, s.l. : Emerald Publishing, 2005, Journal of Quality in Maintenance Engineering , Vol. 11, págs. 279-294.

11. Ayman Z. Faza, Sahra Sedigh, Bruce M. McMillin. Reliability Modeling for the Advanced Electric Power Grid. Computer Safety, Reliability, and Security. s.l. : Springer Berlin / Heidelberg, 2007, Vol. 4680, págs. 370-383.

12. Evaluation of feeder monitoring parameters for incipient fault detection using Laplace trend statistic. C. J. Kim, S.J. Lee, S.H. Kang. IEEE Trans. Industry Applications, Vol. 40, No. 6, Nov-Dec 2004, pp. 1718 – 1724.

13. Distribution incipient faults and abnormal events: case studies from recorded field data. C.L. Benner, B.D Russell. 2004 57th Annual Conference for Protective Relay Engineers, 30 Mar-1 Apr 2004, pp. 86 – 90.

14. Trend detection and data mining via wavelet and Hilbert-Huang transforms. M. Yasar, A. Ray. American Control Conference 2008, 11-13 June 2008, pp. 4292 – 4297.

15. Analysis of sequence of events for the characterization of faults in power systems. O. Quiroga, J. Meléndez, S. Herraiz. Electric Power System Research, Vol. 87, 2012, pp. 22 – 30.

16. Analysis of frequent episodes in sequences of incidences in power distribution systems. O. Quiroga, J. Meléndez, S. Herraiz, A. Ferreira, A. Muñoz. 2011 2nd IEEE PES Int. Conf. Innovative Smart Grid Technologies, 5-7 Dec 2011.

17. Hristo Tanev, Bruno Pouliquen, Vanni Zavarella and Ralf Steinberger "Automatic Expansion of a Social Network Using Sentiment Analysis", bookchapter in Annals of Information Systems, Special Issue on Data Mining for Social Network Data, Springer Verlag, 2010.

18. Daniel Kristopher Harvey, "Forecasting the belief of the population: Prediction Markets, Social Media & Swine Flu", MSc Thesis, U.of Edinburgh, 2009

El segundo describe el paquete de trabajo en el que se ha enmarcado el desarrollo de este TFM, y comienza a partir de la página siguiente.

Esta validación se hará a partir de datos históricos y de casos de prueba representativos, aplicando métodos estadísticos y de simulación realista de los procesos de previsión, obteniendo medidas de la precisión y error de los modelos.

El análisis de estos resultados servirá a su vez para evaluar distintas variantes, permitiendo refinar la definición de los modelos, y seleccionar la más adecuada según los criterios de calidad, robustez y operatividad que se establezcan.

4.1.4 PT4 - DATA MINING Y ESTIMACIÓN DE SEVERIDADES

La estimación de la severidad de los incidentes en la red eléctrica tiene una doble vertiente. Por un lado se tiene el impacto directo que provoca en la propia red de la compañía distribuidora, ya sea en términos de coste por los elementos que han podido verse dañados o en la potencia que se ha dejado de suministrar. Por otro lado se tiene el impacto social que la falta de electricidad durante un tiempo determinado provoca, tanto desde el punto de vista de los problemas que la falta de un bien tan importante crea en los usuarios domésticos o industriales como desde el punto de vista del daño que causa a la imagen de la compañía distribuidora.

Para realizar esa estimación, dividiremos por tanto el problema en dos modelos, el primero de severidad física (gravedad del incidente desde el punto de vista de la red eléctrica) y el segundo de impacto social (repercusión en la sociedad y en la imagen de la compañía). Estos dos modelos serán después integrados en uno solo para ofrecer una medida conjunta de la severidad.

Se llevarán a cabo las tareas relacionadas con la minería de datos procedentes de Web para determinar la repercusión social de los posibles incidentes y la incorporación de esta componente en el cálculo del riesgo para dar lugar a un modelo más amplio de evaluación de la severidad de un fallo. Como resultados de este PT tendremos: (a) un conjunto de técnicas, algoritmos y programas para la minería de eventos en Web; (b) un modelo de severidad basado en AHP y reglas difusas, junto con un conjunto de herramientas para la creación y actualización de ese modelo y algoritmos para la realización de la inferencia de salidas a partir de las entradas.

4.1.4.1 ESPECIFICACIONES DETALLADAS DEL MÓDULO.

Especificar el modelo de severidad general que incorpora las severidades calculadas de los elementos de red afectados y la reputación. Se establecerá el conjunto de variables, su dominio y las estrategias o conocimientos para calcularlo.

El modelo de severidad permitirá obtener reglas que conformarán el sistema experto. A este fin, se propone evaluar dos paradigmas distintos que permiten inferir un valor para la severidad. Por un lado, la obtención de un modelo de razonamiento de lógica borrosa que permite establecer relaciones no lineales entre la variables de entrada para realizar inferencias sobre los valores de las variables de salida y por otro, la aplicación del proceso analítico jerárquico. Si bien, la relación final entre las variables del modelo analítico jerárquico será lineal, tiene la ventaja de facilitar la incorporación de ponderaciones cualitativas entre las variables del modelo.

4.1.4.2 IDENTIFICACIÓN DE FUENTES Y NIVELES DE AGREGACIÓN.

Determinar los sistemas y repositorios de información que servirán como datos de entrada para el cálculo de los modelos de severidad física y de reputación. Se hará un estudio y selección de las fuentes de Internet (buscadores, bases de datos generales o sectoriales, redes sociales, etc.) sobre las que se va a extraer la información. Así mismo habrá que definir con qué nivel de agregación es necesario obtener y analizar los datos y presentar los resultados.

4.1.4.3 MODELO DE SEVERIDAD FÍSICA.

Determinación del modelo que permite el cálculo de la severidad física de un elemento de la red. El objetivo de esta tarea es diseñar y desarrollar un modelo que permita calcular el riesgo de los diferentes elementos de la red a partir de los parámetros físicos y de funcionamiento de los equipos. Una vez modelada la relación entre las variables que determinan el riesgo de un elemento y la posibilidad de fallo del equipo, se propondrán diversas alternativas/técnicas para trasladar dicho modelo conceptual a un diseño formal que permita su implementación. El resultado final será un conjunto de modelos evaluados que permita probar la evaluación del riesgo de los equipos en diversos escenarios reales a lo largo del tiempo.

Para ello, es necesario definir:

- Variables: Definidas por un rango de valores, etiquetas lingüísticas asociados a los valores que están dentro del rango y los conjuntos de valores asociados a las etiquetas
- Relación entre las variables: Definidas con en una estructura jerárquica, con reglas que incorporan el conocimiento de los expertos que permiten calcular los valores de salida dados unos valores de entrada

Como se ha mencionado en apartados anteriores, se elaborará un modelo para inferencia borrosa y otro según el proceso analítico jerárquico.

4.1.4.4 PLATAFORMA DE DESCARGA Y ANÁLISIS DE DATOS DE INTERNET.

Se hará una evaluación de las necesidades de cálculo y almacenamiento de información necesaria para poder asignar los recursos de computación. Esta plataforma hará uso de los recursos computacionales del BIFI y estará basada en tecnologías de computación cloud de manera que permita la asignación dinámica de recursos en función de las necesidades de análisis y almacenamiento de información. En cuanto al análisis de la información, se usarán tecnologías de crawling, indexación, marcado e identificación de temas, técnicas de minería de datos y correlación de datos de internet, análisis de grafos y de redes sociales, etc. para el enriquecimiento y alineamiento con datos estructurados y el diseño de modelos de detección de conceptos y precisión.

4.1.4.5 MODELO DE PÉRDIDA DE IMAGEN REPUTACIONAL.

Este modelo se ocupará de obtener una valoración cuantitativa de la posible pérdida de imagen reputacional debida a una incidencia o fallo en un equipo de la red. A partir de los eventos extraídos en la tarea anterior, representados como un conjunto de variables con sus correspondientes valores instanciados según los eventos sociales esperados, este modelo aplicará un proceso de inferencia para obtener un valor de impacto en la reputación de la compañía de un posible fallo. Este modelo se integrará con el modelo de cálculo del riesgo de mercado clásico, que incluye parámetros como el tamaño de la población afectada.

También en este caso, como se menciona en los modelos de severidad física, es necesario definir estas variables de entrada y salida del modelo, así como caracterizar el proceso de inferencia según las recomendaciones de los expertos.

4.1.4.6 GEORREFERENCIACIÓN DE ÁREAS AFECTADAS

De manera adicional a las valoraciones de severidad obtenidas de los modelos físicos y reputacionales, se calculará una estimación del área geográfica potencialmente afectada por las consecuencias de un determinado incidente. Esta información permitirá a posteriori visualizar y entender mejor hasta dónde se extendería físicamente su impacto, incluso delimitando distintas zonas en función del grado de severidad, duración prevista del incidente, etc.

4.1.4.7 MODELO INTEGRADO DE SEVERIDAD.

El modelo integrado de severidad se ocupará de obtener una única valoración de riesgo de un fallo a partir de los sub-modelos de incidencia de carácter físico y de impacto en el mercado.

4.1.4.8 VALIDACIÓN Y REFINAMIENTO.

Una vez modelada la relación entre conceptos para representar el riesgo de un elemento y la relación entre elementos para modelar la severidad del elemento de la red y de la severidad en la reputación, se propondrán diversas alternativas/técnicas para trasladar dicho modelo conceptual a un diseño formal que permita su implementación. El resultado final será un conjunto de modelos evaluados que permita probar la evaluación del riesgo en diversos escenarios reales a lo largo del tiempo.

4.1.5 PT5 - MODELO INTEGRADO Y DEMOSTRACIÓN

El objetivo de este paquete de trabajo consistirá en integrar los distintos módulos desarrollados en los bloques anteriores, para obtener un modelo que permita estimar de forma conjunta el riesgo de posibles incidentes en el sistema y el impacto asociado, tanto en términos de negocio medibles directamente (p.ej. clientes afectados, potencia, duración), como de forma indirecta por su repercusión en la imagen y reputación de la empresa.

Anexo B

Especificaciones Detalladas del Modelo de Severidad

A continuación se adjunta el documento creado por todo el consorcio mediante Google Docs que contiene las especificaciones detalladas del modelo de severidad global.

Especificaciones Detalladas del Modelo de Severidad

PT4. DATA MINING Y ESTIMACIÓN DE SEVERIDADES

1. Introducción

El fin de este documento es especificar las características del modelo de severidad general que incorporará tanto las severidades físicas estimadas sobre los elementos de red afectados como el impacto social y reputacional de los incidentes. Se establecerá el conjunto de variables, su dominio y las estrategias o conocimientos para calcularlo.

El modelo de severidad permitirá obtener reglas que conformarán el sistema experto. A este fin, se propone evaluar dos paradigmas distintos que permiten inferir un valor para la severidad. Por un lado, la obtención de un modelo de razonamiento de lógica borrosa que permite establecer relaciones no lineales entre la variables de entrada para realizar inferencias sobre los valores de las variables de salida y por otro, la aplicación del proceso analítico jerárquico. Si bien, la relación final entre las variables del modelo analítico jerárquico será lineal, tiene la ventaja de facilitar la incorporación de ponderaciones cualitativas entre las variables del modelo.

El trabajo desarrollado en este paquete estará liderado por el BIFI, aunque en él participarán todos los socios. Se prevé una interacción más activa con Gas Natural Fenosa, Scien, y especialmente la Universidad Carlos III.

Este documento se basa en las especificaciones detalladas en el documento entregable titulado "Especificaciones funcionales de alto nivel", profundizando en ellas, de manera que sirva como punto de partida de trabajo del PT4.

2. Especificaciones funcionales PT4

Recordamos aquí las especificaciones funcionales finales del PT4 del entregable correspondiente:

CÓDIGO	ESPECIFICACIÓN FUNCIONAL	COD EF GLOBAL	GNF	AIA	SCIEN	TELVENT	UC3M	UdG	BIFI
EF.PT4.01	El modelo de severidad integrará valores de severidad física (PxT) con severidad social, reputacional y la opinión de los expertos	EF.G.01	X	X	X		X	X	X

EF.PT4.02	El modelo de severidad asignará un valor de severidad a un incidente de un tiempo determinado y bajo unas circunstancias dadas	EF.G.03	X		X		X	X	X
EF.PT4.03	El modelo de severidad será capaz de ofrecer resultados a distintos niveles de agregación, desde base hasta zonas geográficas determinadas	EF.G.03	X		X		X	X	X
EF.PT4.04	El sistema recogerá información periódicamente de distintas fuentes (detalladas en el documento pertinente) sobre eventos de gran repercusión social	EF.G.01	X				X		X
EF.PT4.05	Los posibles eventos se caracterizarán por título, estampilla de tiempo, categoría, aforo aproximado y geolocalización	EF.G.01	X				X		X
EF.PT4.06	Utilizando la localización de dichos eventos, se mostrarán en un mapa los lugares en los que una incidencia puede tener mayor gravedad social	EF.G.02	X		X		X		X
EF.PT4.07	La plataforma recogerá datos relacionados con la red eléctrica en medios en los que los usuarios transmitan su opinión (definidos en el documento correspondiente)	EF.G.05	X						X
EF.PT4.08	Estos datos vendrán caracterizados por la fecha, la fuente, y el contenido en sí	EF.G.05	X						X
EF.PT4.09	A partir de estos datos el sistema dará una visión de la repercusión social que tienen los incidentes eléctricos y su dependencia con diferentes factores (geográficos, magnitud, eventos o localizaciones que se han visto afectados, etc)	EF.G.05	X		X		X		X
EF.PT4.10	El usuario podrá consultar estos datos siempre que lo desee tanto en el mapa como en formato lista filtrados por distintos criterios de búsqueda	EF.G.06							X
EF.PT4.11	El usuario podrá configurar las fuentes de las que se tomarán los datos, así como el nivel de periodicidad	EF.G.06	X		X				X

EF.PT4.12	El usuario final podrá exportar los datos a formatos explotables desde otros programas como XLS, CSV, etc.	EF.G.03	X		X				X

Descripción de especificaciones funcionales del PT4

EF.PT4.01: El modelo de severidad integrará valores de severidad física (PxT) con severidad social, reputacional y la opinión de los expertos

Para estimar la severidad de un incidente se tendrá en cuenta tanto la potencia no suministrada como la importancia que la compañía le otorgue al mismo por su posible impacto social o reputacional. Para estos últimos conceptos se tendrán en cuenta tanto los datos capturados de Internet para conocer la importancia que los clientes otorgan al incidente como la opinión de los expertos de la compañía distribuidora.

EF.PT4.02: El modelo de severidad asignará un valor de severidad a un incidente de un tiempo determinado y bajo unas circunstancias dadas

El modelo de severidad no será un modelo de previsión sino que, dado un incidente de una cierta duración en un sitio determinado y bajo unas circunstancias dadas (por ejemplo, hay o no partido de fútbol), se le otorgará una determinada severidad.
Por tanto, la previsión final de riesgo vendrá dada por la previsión de probabilidad de incidente y su duración combinada con la severidad asociada.

EF.PT4.03: El modelo de severidad será capaz de ofrecer resultados a distintos niveles de agregación, desde base hasta zonas geográficas determinadas

De forma similar al modelo de previsión de incidentes, y en combinación con él, la herramienta podrá ofrecer valores de severidad asociados tanto a bases como a otras zonas geográficas determinadas.

EF.PT4.04: El sistema recogerá información periódicamente de distintas fuentes (detalladas en el documento pertinente) sobre eventos de gran repercusión social

La base del sistema de previsión y prevención de riesgos será el análisis de los datos obtenidos de las fuentes definidas previamente. Para ello, realizará ciertas tareas periódicamente que se ocuparán de de recogerlos y almacenarlos de forma que queden disponibles y en un formato óptimo para su tratamiento y consulta.

EF.PT4.05: Los posibles eventos se caracterizarán por título, estampilla de tiempo, categoría, aforo aproximado y geolocalización

Los campos necesarios para el correcto tratamiento y la consulta de los datos analizados para el modelo deberán de disponer de los campos título, en el que se resumirá de qué va el evento; estampilla de tiempo, que indicará cuándo se producirá; categoría, que podrá ser de deportes, cultural, etc.; el aforo aproximado, que ayudará a dar una idea cuantitativa del impacto que puede tener el evento; y la geolocalización, para su posterior muestra en un mapa. Todos estos atributos serán extraídos de las fuentes de datos seleccionadas.

EF.PT4.06: Utilizando la localización de dichos eventos, se mostrarán en el mapa los lugares en los que una incidencia puede tener mayor gravedad social

Mostrar los eventos en un mapa es una de las herramientas más importantes de las que dispondrá el sistema, ya que permite ver de un sólo vistazo cuáles pueden ser las zonas de la red más afectadas y concentrar los esfuerzos. El sistema facilitará la identificación de dichas zonas jerárquicamente según el grado de severidad, ya sea a través de un código de colores o de otras técnicas similares.

EF.PT4.07: La plataforma recogerá datos relacionados con la red eléctrica en medios en los que los usuarios transmitan su opinión (definidos en el documento correspondiente)

Se recogerán datos de diversas fuentes ya definidas en las que los usuarios puedan dar su opinión, y que puedan estar relacionadas con los incidentes eléctricos. Se realizará un análisis que permita obtener una valoración general dentro de unos límites previamente especificados.

EF.PT4.08: Estos datos vendrán caracterizados por la fecha, la fuente, y el contenido.

Los datos necesarios para realizar dicho análisis serán la fecha de emisión de la opinión; el origen de la misma, ya sea una noticia, un foro, etc.; y el contenido. Todo esto se guardará para, además de la valoración, poder consultar los datos de forma manual cuando se desee.

EF.PT4.09: A partir de estos datos el sistema dará una visión de la repercusión social que tienen los incidentes eléctricos y su dependencia con diferentes factores (geográficos, magnitud, eventos o localizaciones que se han visto afectados, etc

La valoración de la repercusión social de los incidentes se basará principalmente en la búsqueda de palabras clave y el conteo de posibles coocurrencias de palabras en los elementos recogidos. Normalmente la aparición de una empresa energética en la red cuando ocurre un incidente suele tener connotaciones negativas, ya que la gente no arranca a opinar positivamente sobre éstas. En cambio, al contrario sí que suele darse esa situación, y en cuanto aparece un problema se suceden las opiniones negativas.

EF.PT4.10: El usuario podrá consultar estos datos siempre que lo desee tanto en el mapa como en formato lista filtrados por distintos criterios de búsqueda

El usuario podrá consultar todos los datos tanto en formato mapa, ya que es la forma más visual y directa de hacerlo, como en formato listado. Con este último facilitaremos la búsqueda de datos concretos a través interfaz y de una serie de filtros por los atributos que se dispongan de los mismos.

EF.PT4.11: El usuario podrá configurar las fuentes de las que se tomarán los datos, así como el nivel de periodicidad

El sistema ha de ser configurable para que, de todas las fuentes de datos disponibles, el usuario pueda escoger cuáles le interesan más. También se podrá indicar al sistema la periodicidad con la que recogerá los datos de dichas fuentes.

EF.PT4.12: El usuario final podrá exportar los datos a formatos explotables desde otros programas como XLS, CSV, etc.

El usuario podrá obtener dichos datos del sistema en diferentes formatos con el fin de que sean más manejables desde otras herramientas con el fin de ser tratados de distinta manera. Esto permitirá elaborar gráficas, informes, etc., así como exportarlos a otros sistemas.

3. Especificaciones del modelo

En este apartado vamos a especificar todas las características necesarias para modelar la estimación de severidades, que se divide en el modelo de severidad física, el modelo de mercado, así como el de pérdida de reputación y el impacto social producidos por la aparición de determinados incidentes en la red.

3.1 Modelo de severidad física y de mercado

Este modelo definirá la gravedad que tendrían incidentes de una duración determinada acaecidos en una cierta instalación y bajo unas condiciones concretas (de carga, etc) desde un punto de vista físico (pérdida de potencia suministrada..) y de mercado (quizás no tiene la misma gravedad que se quede sin luz un determinado tipo de clientes que otro, etc). No se trata por tanto de un cálculo de probabilidad de fallo sino, dado un fallo determinado, qué severidad le asignamos.
Se trabajará a partir de los modelos jerárquicos de UC3M y permitirá la incorporación también en la jerarquía de la parte de pérdida reputacional que describimos en el apartado siguiente.

3.2 Modelo de pérdida reputacional

El modelo de pérdida reputacional pretende valorar el perjuicio que para la imagen de GNF provoca la ocurrencia de determinados incidentes eléctricos. Para ello, se recogerá información de distintas fuentes en las que la sociedad refleja su opinión, y en las que se pueden encontrar referencias tanto a la empresa que se está estudiando como a eventos y sucesos directamente relacionados con ella.

Este modelo se puede dividir en dos partes bien diferenciadas. Por un lado, se realizará un análisis a priori, en el que se tratará de valorar el impacto social de un posible incidente en base a la importancia del evento en el que se pueda dar, y por otro, el análisis a posteriori, que observando el comportamiento de los usuarios nos permitirá valorar la repercusión de un incidente una vez que haya sucedido. Esta segunda parte servirá para retroalimentar y validar la primera.

Análisis a priori

El principal objetivo de este análisis es identificar aquellos lugares o eventos susceptibles de sufrir un incidente que pueden tener más repercusión social. Por ejemplo, no es lo mismo dejar sin suministro eléctrico de 100 MW a 5.000 clientes de un barrio que a un solo cliente, pero es muy probable que si ese único cliente es un hospital, la repercusión y severidad de ese incidente sea mucho mayor que la del primero.

Para ello, y basándonos en la experiencia de los expertos de GNF, se configurarán una serie de categorías en las que se enmarcarán todos los eventos que se identifiquen. Dichas categorías tendrán un valor de severidad predefinido. Cada evento tendrá también un valor de severidad definido, y en caso de que éste no pueda ser estimado por no disponer de los datos suficientes, se utilizará el de la categoría a la que pertenezca. La severidad de un incidente que pueda suceder durante un evento vendrá definido inicialmente por la afluencia al mismo, aunque la audiencia no tiene porqué ser únicamente la que asista presencialmente, sino que también puede haber audiencia a través de TV, internet, radio, etc. Una primera clasificación podría ser:

- Grandes eventos deportivos: partidos principalmente de fútbol, y partidos de gran relevancia de otros deportes como tenis, baloncesto, grandes premios, etc.
- Macroconciertos o fiestas de gran aforo
- Fiestas o celebraciones de alto standing: como ópera, funciones de teatro muy conocidas, conciertos en lugares más limitados, etc.
- Eventos de gran calado social: visita de famosos, manifestaciones, elecciones, presentaciones
- Celebración de fiestas locales

Con el fin de hacer un modelo más fiable, se permitirá añadir de forma manual eventos que no se hayan encontrado en las fuentes definidas y que se tenga certeza de que son relevantes.

Análisis a posteriori

Por otro lado, se estudiarán y analizarán datos recogidos de distintas fuentes como noticias, foros, redes sociales, para dar una valoración de la pérdida de imagen de la empresa debido a un incidente que haya sucedido que podría categorizarse en distintos niveles como por ejemplo: “Muy grave”, “Grave”, “Media”, “Leve”, y “Muy leve”. Esta valoración se basará principalmente en la búsqueda de palabras clave y el conteo de posibles co-ocurrencias de las mismas en los elementos recogidos. También se utilizarán otras herramientas como Google Trends y Google

Analytics para estudiar los patrones de acceso a la web de GNF así como de las búsquedas relacionadas con la empresa en base a estos incidentes.

Datos

Las fuentes de datos que se utilizarán para evaluar la reputación de GNF en la red son:

- Fuentes para eventos: este caso es bastante específico, y ya se han localizado diferentes fuentes que contemplan eventos de las categorías descritas previamente. Son principalmente webs de internet que disponen de calendarios de eventos con geolocalización y otros datos interesantes sobre los mismos. Estas fuentes habrá que redefinirlas con la opinión de los expertos, siempre teniendo en cuenta que la zona geográfica de estudio será Galicia.
- Facebook places, Foursquare, Tuenti Sitios y otras: en estas aplicaciones online se etiquetan gran parte de eventos que tienen cierta repercusión social y que es muy probable que no aparezcan en otras fuentes de datos más estructuradas como las del punto anterior. También se puede conseguir información en ellas de la asistencia esperada a los mismos, incluso tener feedback a posteriori del transcurso del mismo, información que se puede utilizar para estudiar posibles incidentes.
- Twitter: en el BIFI se recogen constantemente datos de esta red social relacionados con las palabras clave que se pretende estudiar. Identificando nuevas palabras clave y hashtags que pueden ser interesantes para nuestro modelo, se modificará el web crawler que ya está en funcionamiento para recoger estos datos, almacenarlos en una base de datos, para posteriormente analizarlos estudiando tiempos de aparición, número de veces que se ha dado cada término, enlaces publicados relacionados, etc.
- Elmundo.es, Elpaís.es, Marca.com: en esta fuente aparecen noticias a nivel nacional que pueden estar relacionadas con GNF y concretamente con posibles incidentes eléctricos. Analizando las noticias que se publican constantemente, buscando determinadas palabras clave en ellas, y analizando los comentarios de sus usuarios, nos puede servir para mejorar la estimación de la reputación de la empresa. Esto se realizará gracias a la creación de un nuevo web crawler específico para estas webs de noticias. Además, estos sitios incorporan ciertas estadísticas publicadas dentro de cada noticia que pueden ayudar a esta tarea.
- Se analizarán también fuentes de datos más locales como pueden ser los periódicos y otros medios de prensa nacionales (o sólo de Galicia)
- Se realizará una tarea similar con los foros de opinión más importantes en los que pueden aparecer también referencias a GNF y a sucesos que pueden afectar a su reputación. Algunos de los más grandes de habla hispana y en los que se tratan todo tipo de temas con bastante repercusión social son Forococoches y Mediavida, que podrían ser analizados con técnicas similares a las anteriores.
- Se realizarán estudios a través de distintas herramientas de Google. Una de ellas es Google Trends, que permite analizar el volumen de búsquedas que ha tenido un término históricamente, así como compararlo con otros términos diferentes. También se

estudiará a través de Google Analytics el patrón de accesos a la web de GNF, lo que nos permitirá ver si los usuarios la han visitado desde Google, escribiendo su dirección directamente, o través de un enlace de una noticia, una red social, etc.

3.3 Integración de los modelos

La fase de integración de los modelos se va a llevar a cabo desde el inicio con el fin de facilitar la tarea. La opción de crear dos modelos diferentes para posteriormente integrarlos pierde sentido, sobre todo en términos de fiabilidad de los mismos. Por ello se va a diseñar un modelo de severidad de mercado que incorpore los componentes social y reputacional.

Anexo C

Planificación y seguimiento

La planificación del proyecto fue planteada por el consorcio con mucho margen para evitar posibles imprevistos. Al tratarse de un proyecto de gran envergadura, con bastantes socios, y con distintos orígenes, la probabilidad de que hubiera algún desajuste o retraso por mala comunicación, porque una tarea se retrasara, etc., aumentaba considerablemente, sobre todo si tenemos en cuenta que se trataba de un proyecto cuya base es la investigación y la implantación de técnicas bastante nuevas y experimentales. En la figura C.1 encontramos la primera versión del calendario propuesto, que como hemos comentado, se realizó con mucha holgura en previsión de las circunstancias que se pudieran dar, y que se han dado (abandono de ciertos trabajadores de las empresas/instituciones del consorcio, dificultades técnicas, etc.).

Para organizar el trabajo y llevar el seguimiento del proyecto, el BIFI se ofreció a instalar, configurar y mantener una herramienta de gestión de proyectos. Dado que en el proyecto SEPS no se contemplaban fondos para una herramienta de este tipo, se optó por buscar alternativas gratuitas o libres. Después de barajar distintas opciones, se decidió utilizar Teambox en su versión 3, ya que es la última de código abierto que se publicó, otros socios estaban familiarizados con ella, y contenía todo lo necesario que en un principio se necesitaba para llevar a cabo el proyecto. Se creó un proyecto dentro de Teambox específico para SEPS, se invitó a los usuarios que iban a participar en él, se configuraron las alertas para que llegaran los cambios que sucedieran por correo electrónico, y se habilitó la creación de carpetas y la subida de archivos para mantener en esta herramienta los documentos oficiales y definitivos. En la figura C.2 podemos ver una captura reciente de la herramienta.

Como casi todas las herramientas, ésta tenía ciertas carencias, o simplemente existían otras alternativas más adecuadas para realizar ciertas tareas, como es la creación de documentos colaborativa. Para ello se decidió utilizar Google Docs, creando una carpeta compartida para el proyecto, así como una subcarpeta para cada paquete de trabajo, permitiendo ver los cambios en vivo, y evitando tener que enviar las distintas versiones de los documentos por correo electrónico o subirlas a Teambox.

En cuanto a la dedicación de las tareas listadas en el capítulo 1, en la tabla C.1 podemos ver

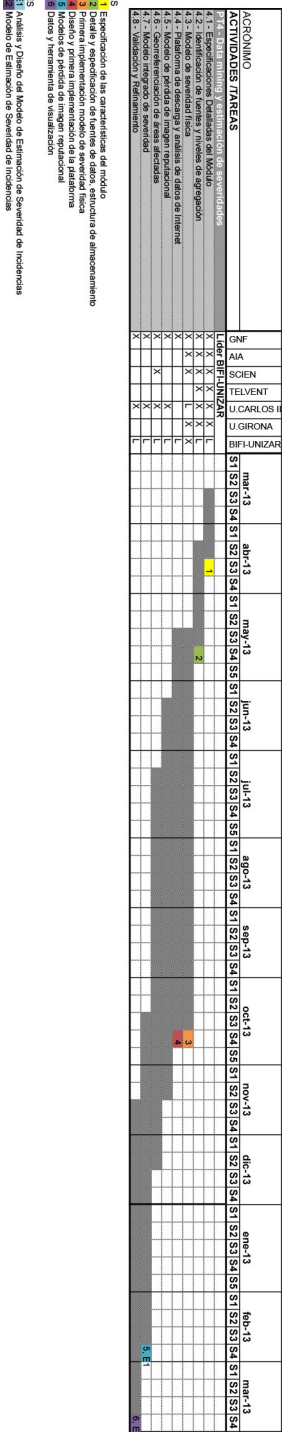

Figura C.1: Planificación inicial

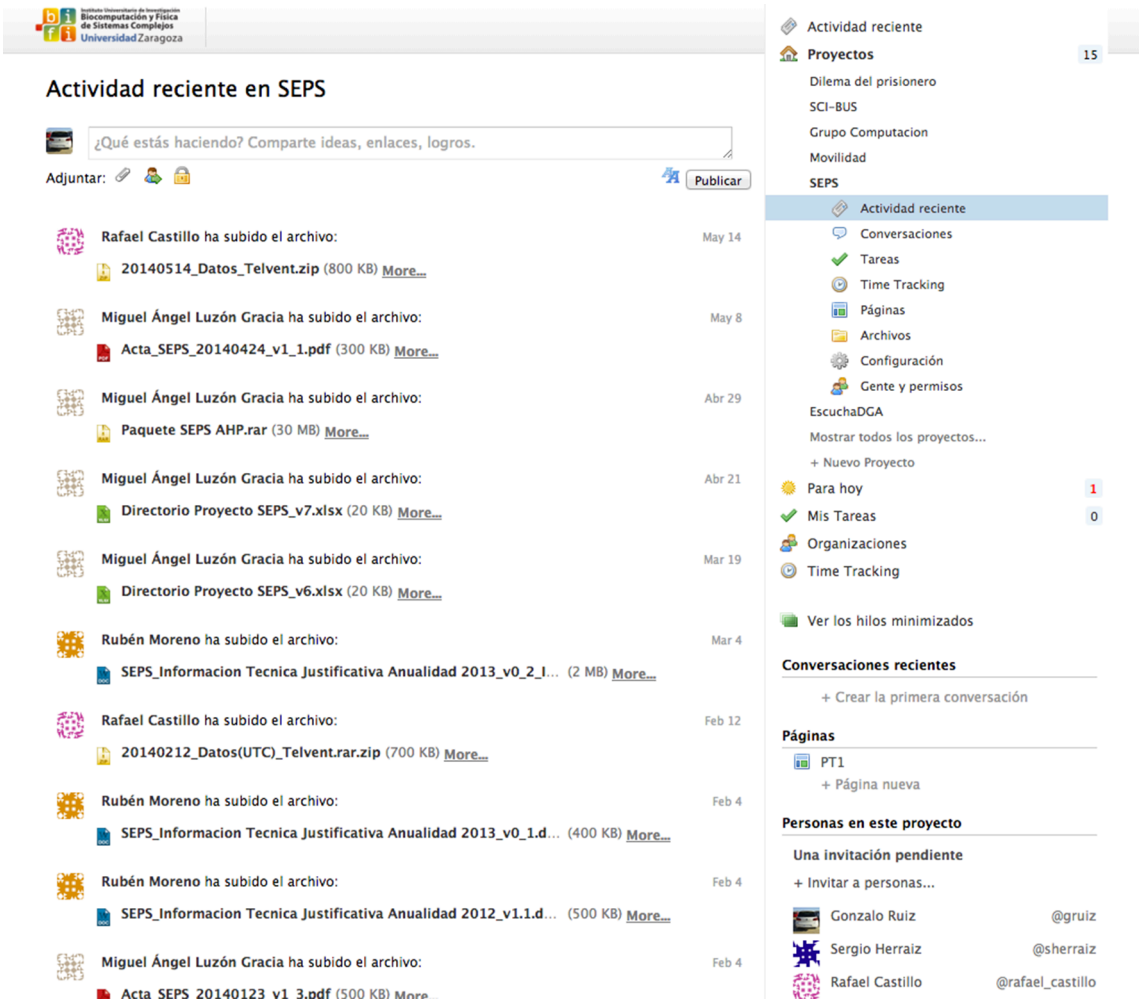

Figura C.2: Gestión del proyecto en Teambox

cuánto tiempo se ha invertido en cada una, y en qué período ha sido ejecutado debido a que el autor no se ha dedicado a tiempo completo a este TFM durante su desarrollo.

Tarea	Período	Dedicación
Estudio de posibles fuentes en Internet de las que obtener información relevante para el sistema	Jun '13 - Jul '13	32 horas
Instalación y mantenimiento de una herramienta para la planificación y el seguimiento del proyecto	Jun '13 - Jul '13	24 horas
Análisis, diseño, implementación y optimización de una arquitectura HW+SW que soportara toda la carga necesaria por el sistema	Sept '13 - Oct '13	48 horas
Análisis de las diferentes técnicas de extracción de información de sitios web	Oct '13 - Nov '13	64 horas
Diseño e implementación de un modelo de datos en el que se pueda almacenar toda la información	Nov '13 - Ene '14	40 horas
Diseño, implementación y pruebas de los distintos scrapers necesarios para la extracción de datos	Nov '13 - Feb '14	160 horas
Diseño de un sistema de control para la ejecución de los diferentes scrapers desarrollados	Feb '14 - Mar '14	56 horas
Definición de los valores por defecto, las consultas y filtros necesarios para la evaluación de los posibles riesgos	Abr '14	24 horas
Estudio de las diferentes técnicas y herramientas existentes para el análisis de sentimiento	Feb '14 - Abr '14	120 horas
Implementación de un analizador de sentimiento	Abr '14 - May '14	88 horas
Análisis, diseño e implementación de una interfaz para controlar el sistema de recogida de datos, evaluación de riesgos subyacentes y monitorización	Mar '14 - May '14	120 horas
Gestión de versiones y copias de seguridad de la aplicación, empaquetado para su fácil instalación	Ene '14 -	40 horas
Documentación de las reuniones realizadas durante el proyecto	Jun '13 -	8 horas
Planificación del proyecto	Jun '13 -	16 horas
Total		**840 horas**

Cuadro C.1: Dedicación de tiempo al desarrollo del TFM

Anexo D

Aplicación piloto

En este capítulo se incluyen capturas de pantalla de las diferentes secciones de las que consta la interfaz desarrollada para la gestión del sistema y que no han sido incluidas en la parte principal de la memoria.

En la figura D.1 podemos ver la página principal de la aplicación piloto, desde la que podemos acceder fácilmente al resto de secciones y funcionalidades disponibles.

La figura D.2 muestra una instantánea de la gestión de fuentes que permite la interfaz, facilitando detener una fuente, eliminarla, o añadir una nueva.

En la figura D.3 tenemos una captura de la gestión de categorías de eventos desde la que se pueden afinar los valores para cada una de ellas.

La figura D.4 es una instantánea de la pantalla de eventos capturados desde la que se permite afinar los valores de cada uno de ellos.

En la figura D.5 podemos ver la pantalla que posibilita ajustar finamente los valores de cada tweet, noticia o comentario capturado, de forma similar a la anterior con los eventos.

La figura D.6 es una captura de la pantalla de gestión desde la que podemos añadir o eliminar las palabras clave que deseemos para la captura de tweets, noticias, etc.

En la figura D.7 podemos ver una pantalla muy similar a la anterior desde la que se permite gestionar los usuarios de Twitter de los que se capturan datos.

La figura D.8 muestra la pantalla desde la que se visualizan los eventos que van a suceder próximamente y la cercanía de los mismos a las subestaciones de Unión Fenosa.

En la figura D.9 podemos ver la misma pantalla de visualización de eventos, en la que interactuando con el mapa se obtiene información adicional tanto de ellos como de las subestaciones.

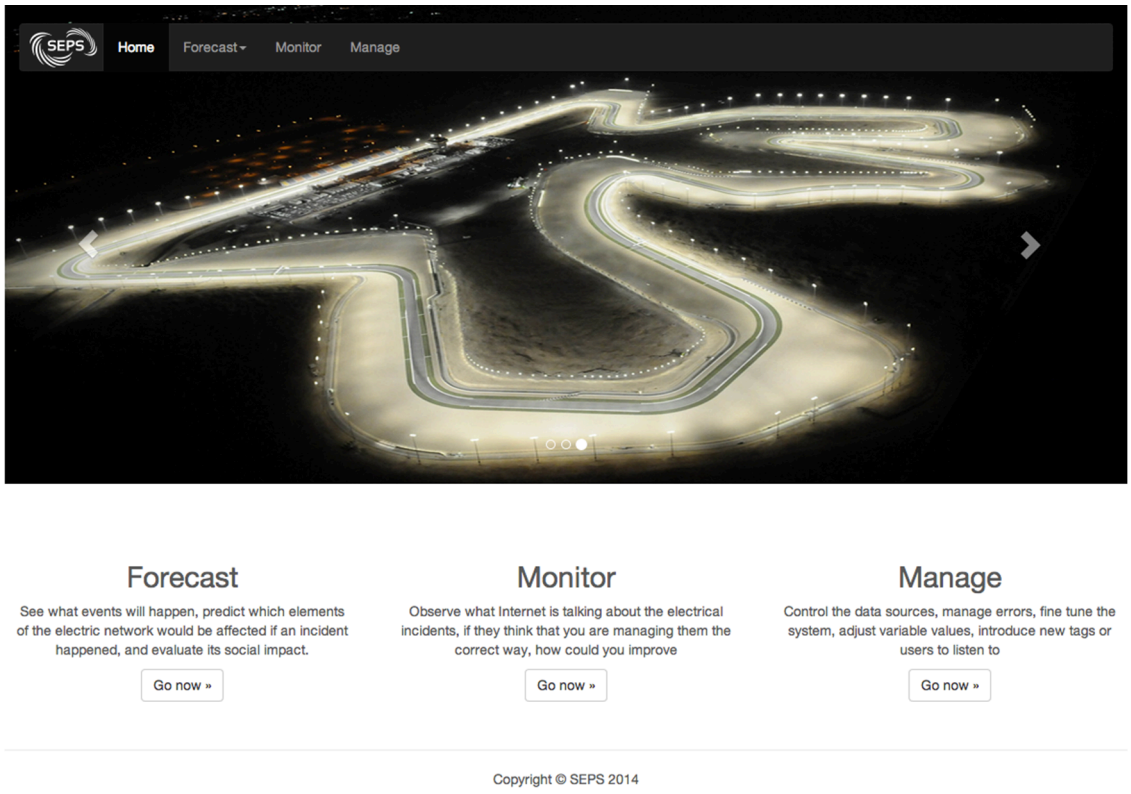

Figura D.1: Página principal de la aplicación piloto

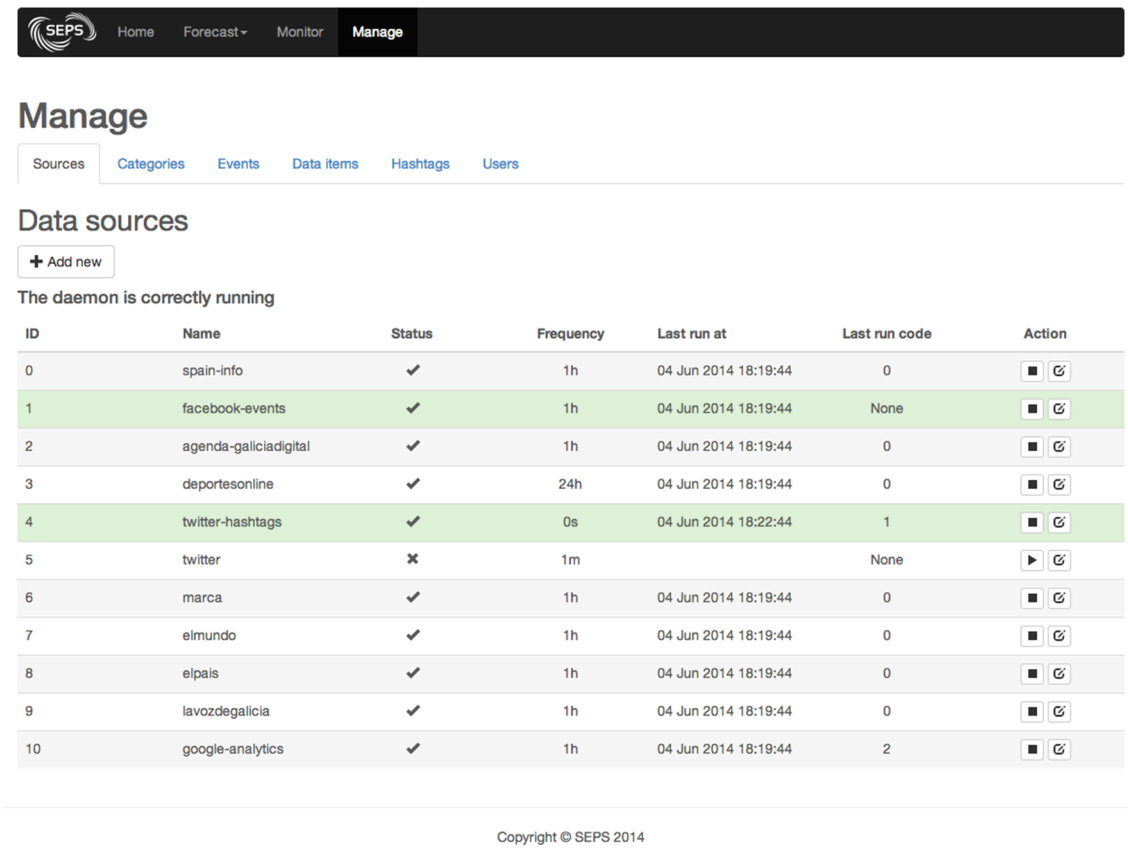

Figura D.2: Gestión de fuentes de datos a través de la interfaz

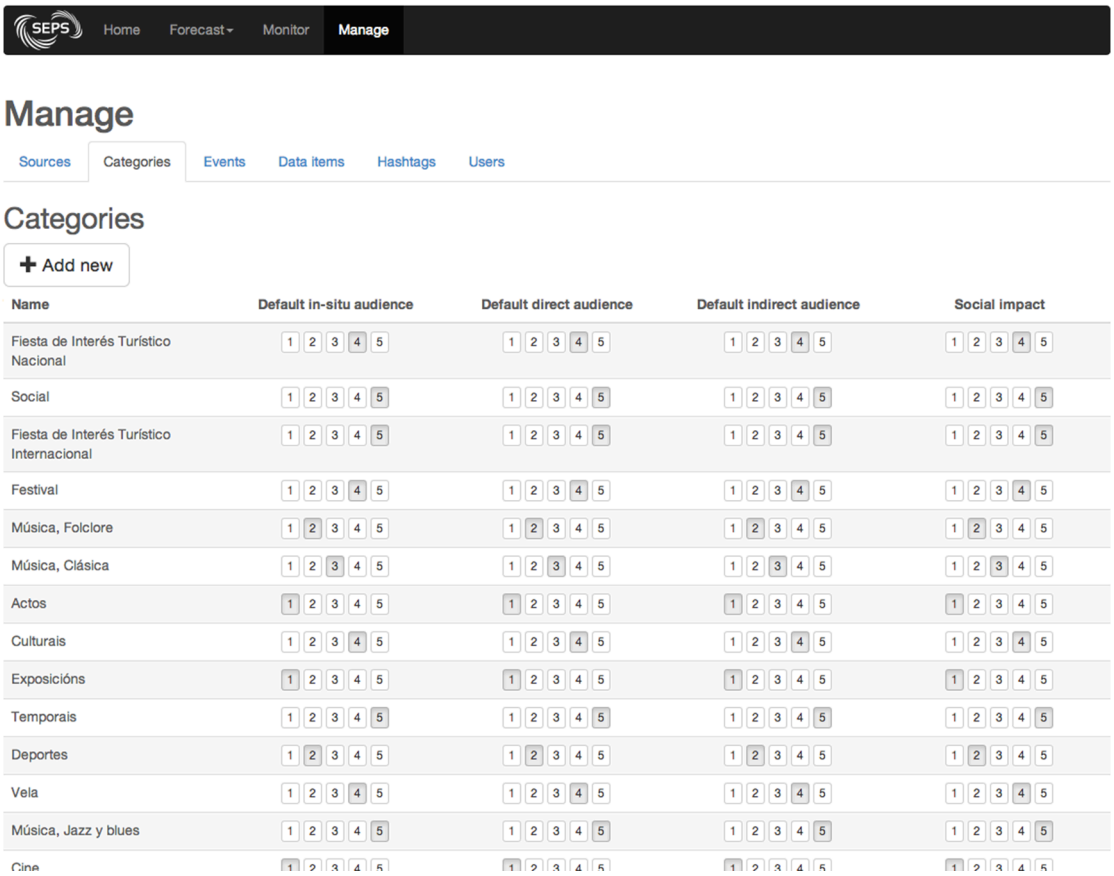

Figura D.3: Gestión de categorías de eventos a través de la interfaz

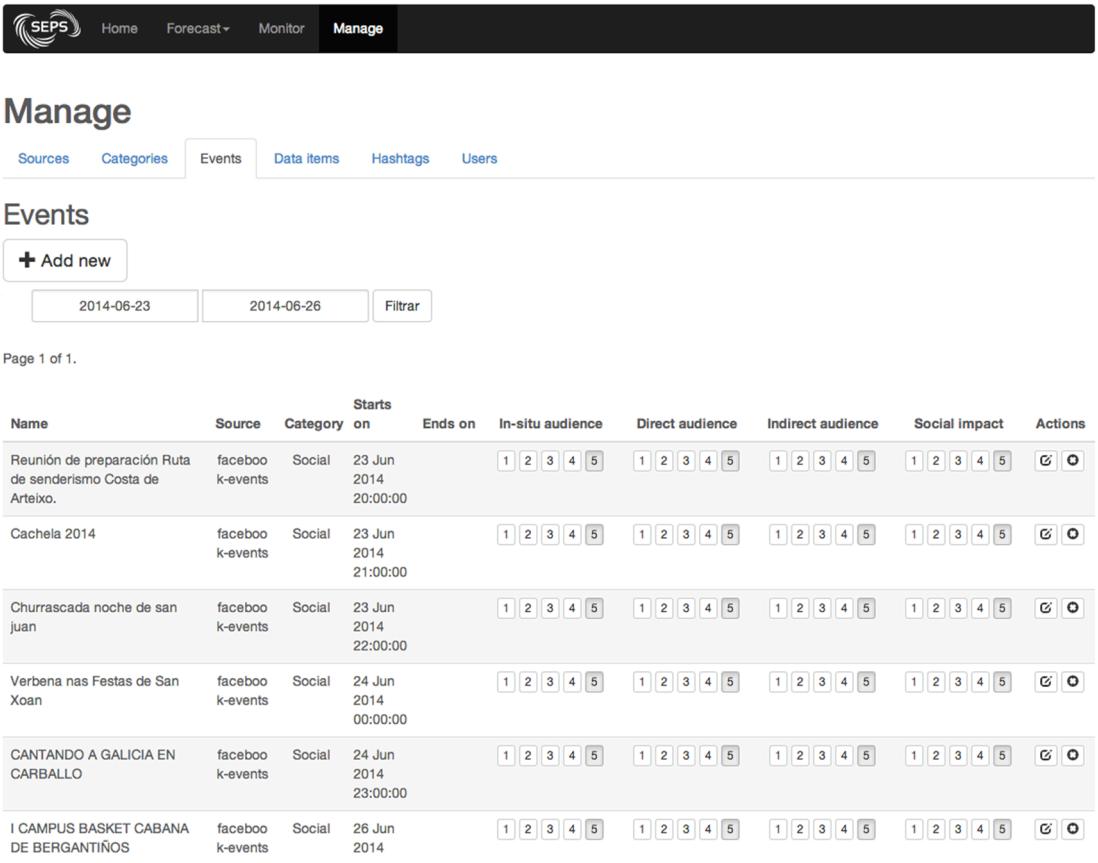

Figura D.4: Gestión de eventos a través de la interfaz

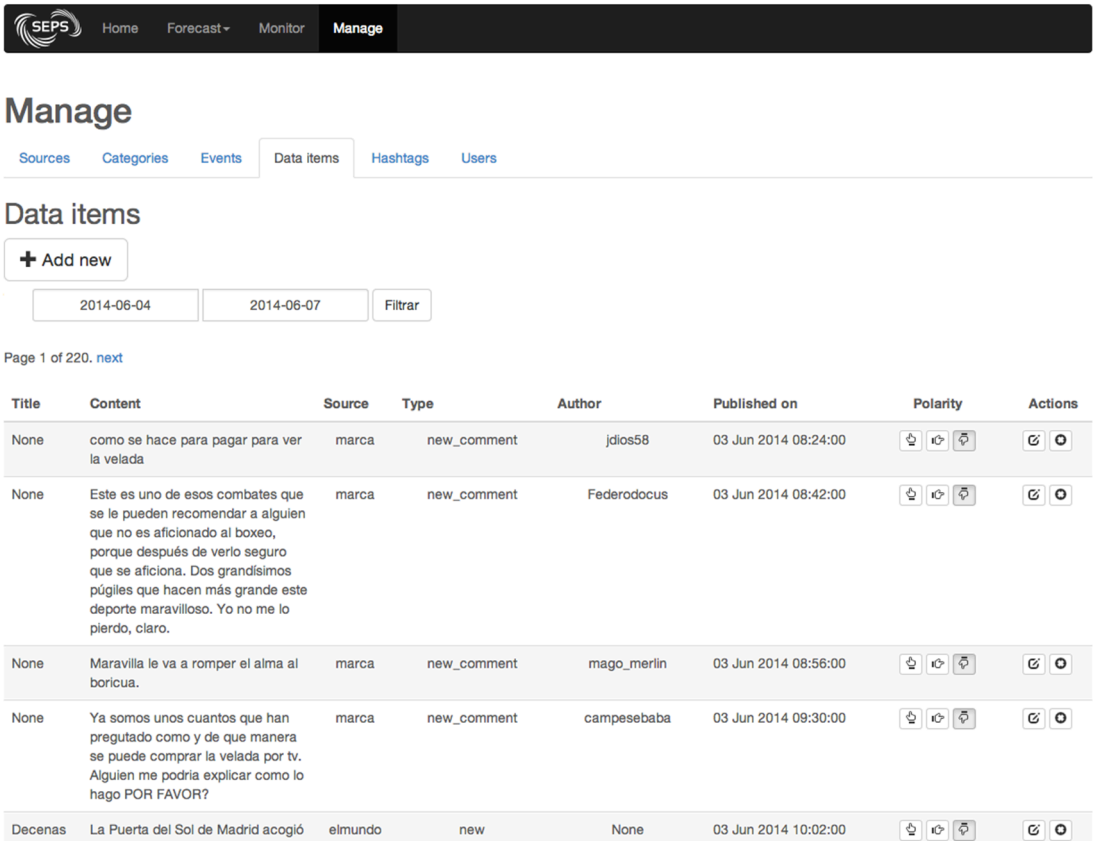

Figura D.5: Gestión de contenidos capturados a través de la interfaz

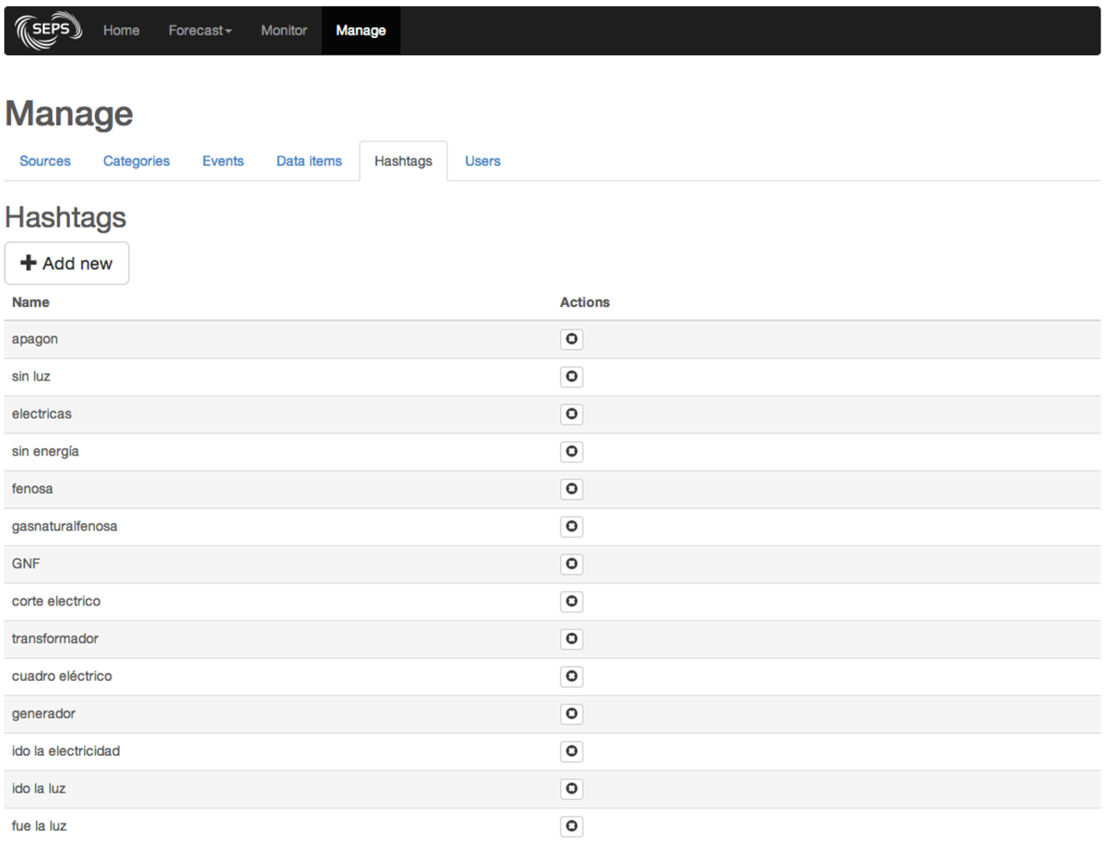

Figura D.6: Gestión de palabras clave utilizadas para la captura a través de la interfaz

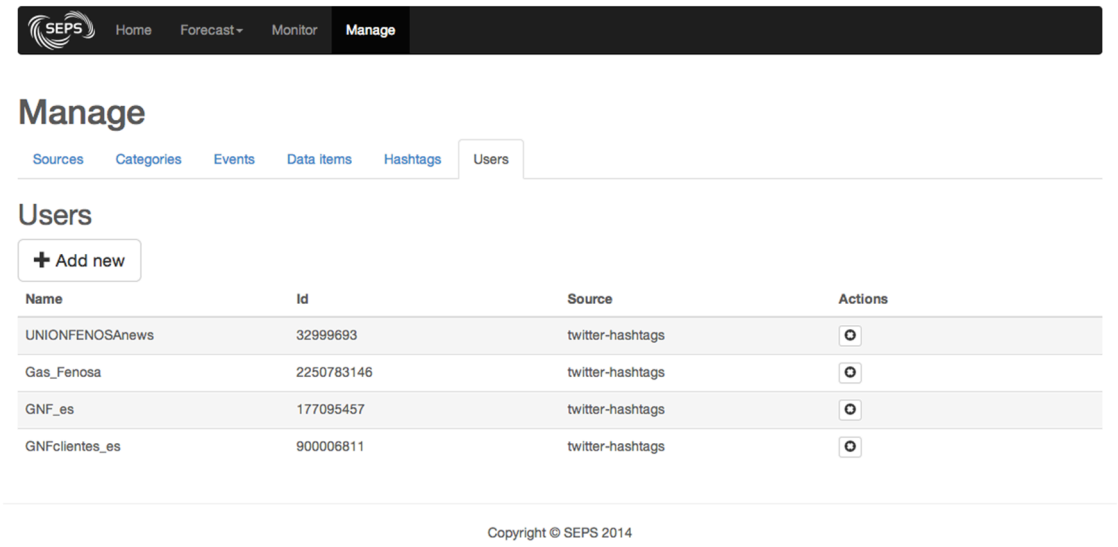

Figura D.7: Gestión de usuarios utilizados para la captura a través de la interfaz

En la figura D.10 podemos ver una captura de la pantalla en la que se muestra un mapa de calor con los datos geoposicionados recientemente. Esta captura ayuda a apreciar cuáles son las zonas de mayor tránsito y en las que un evento puede tener más repercusión.

En la figura D.11 podemos ver una captura de la pantalla desde la que los operarios pueden ver diferentes estadísticas de los datos capturados.

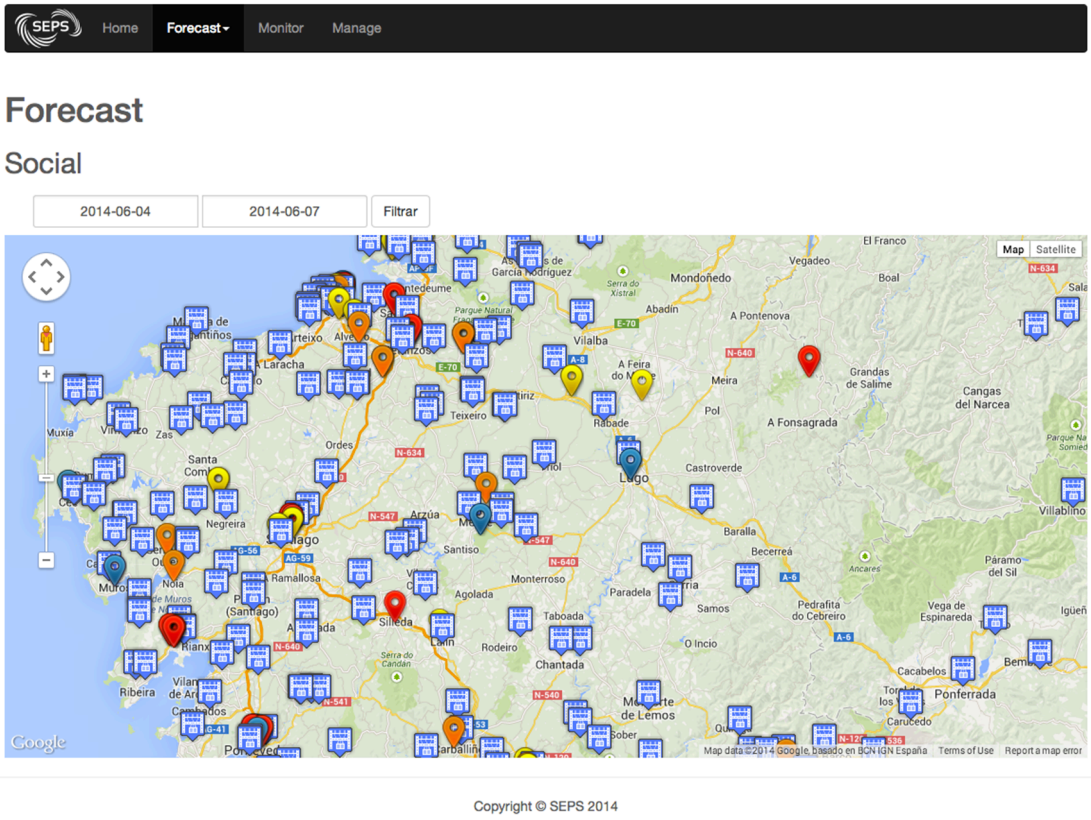

Figura D.8: Pantalla para la visualización de eventos geoposicionados

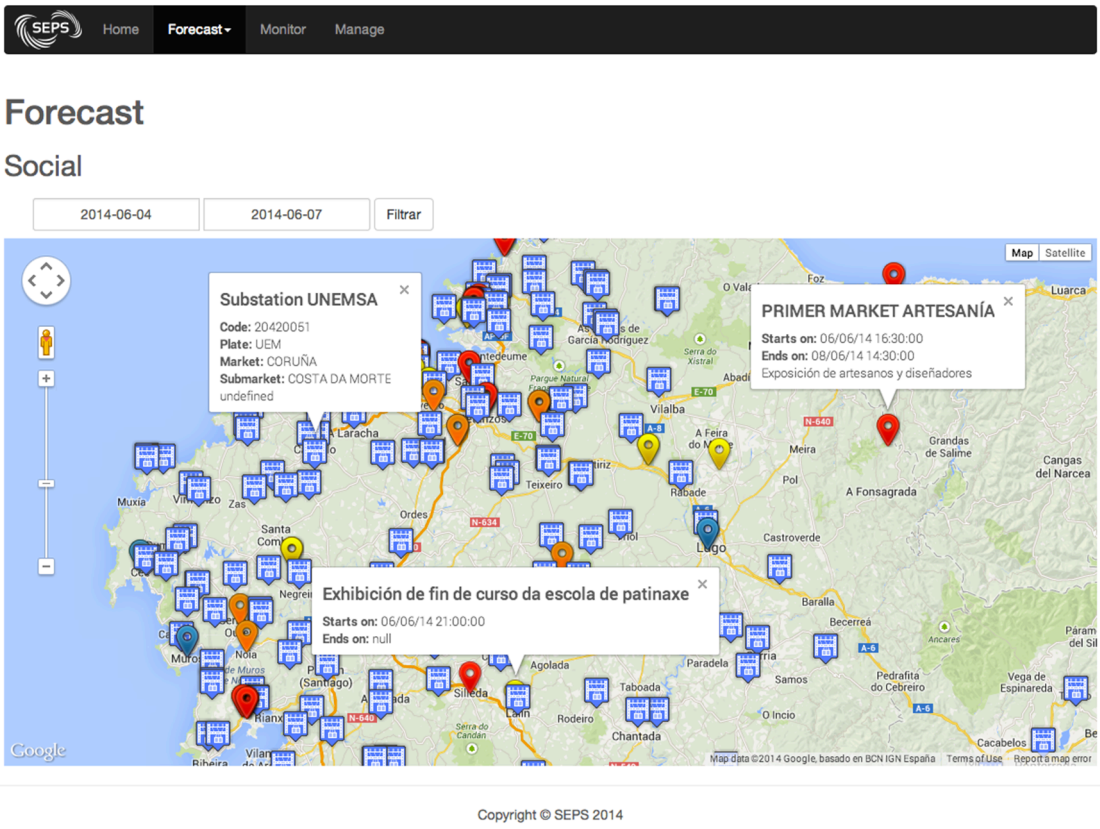

Figura D.9: Pantalla para la visualización de eventos geoposicionados con información adicional

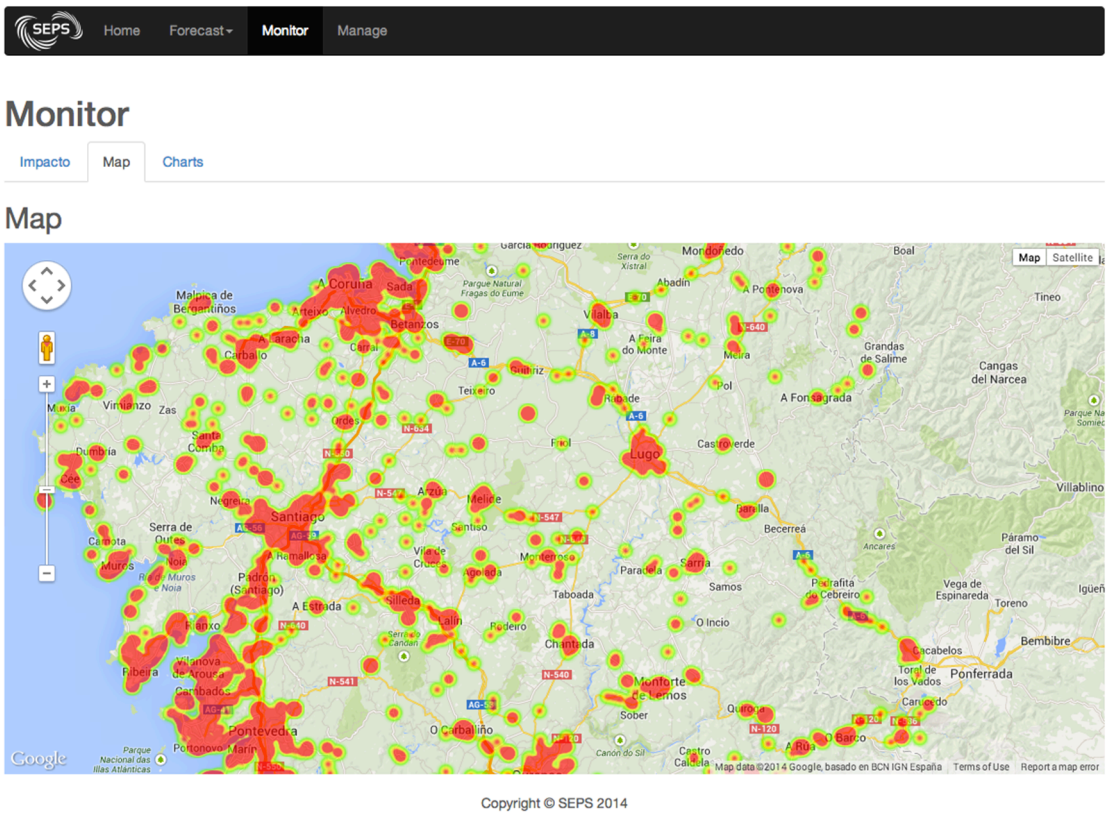

Figura D.10: Visualización en forma de mapa de calor de los elementos geoposicionados capturados

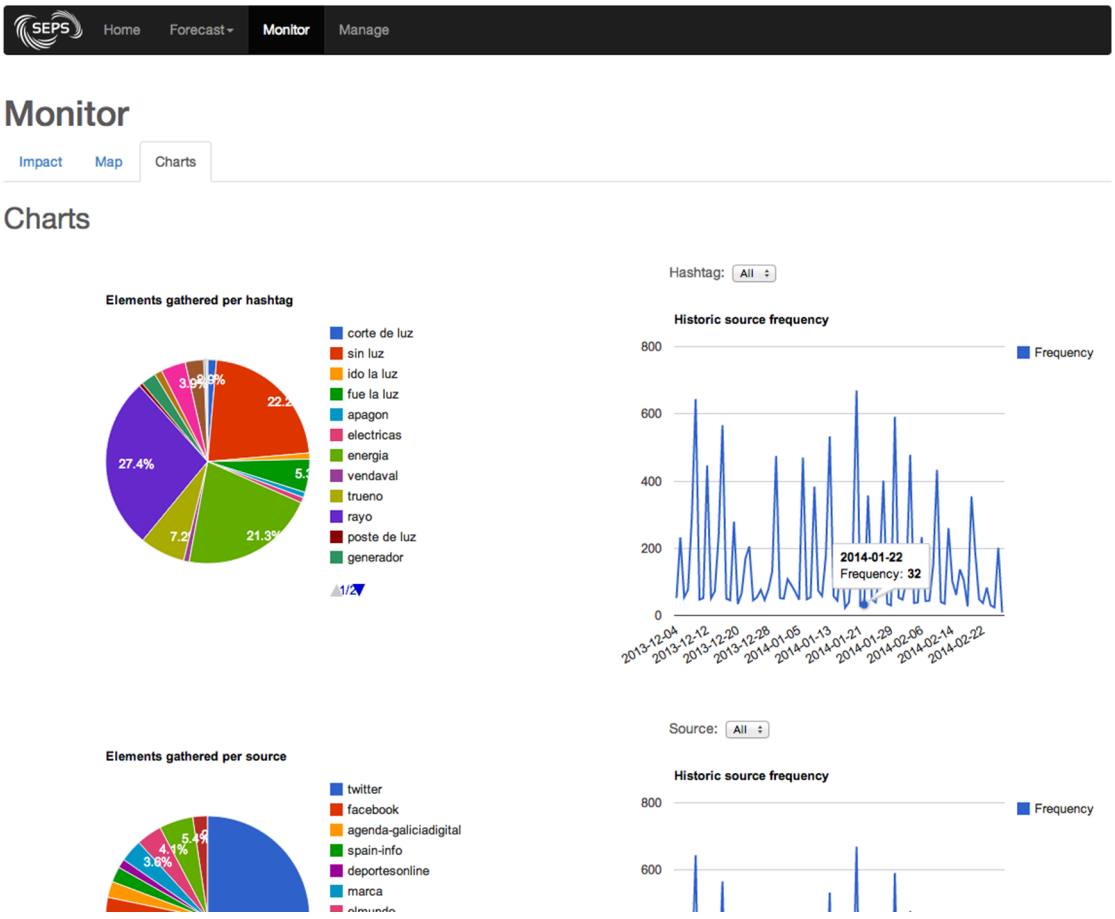

Figura D.11: Estadísticas de los datos capturados por fuente y por palabra clave

Anexo E

Modelo de datos

En este anexo se detalla el modelo de datos utilizado a lo largo del proyecto.

En la figura E.1 podemos ver un diagrama entidad relación del modelo de datos implementado.

También se adjunta un documento interno que describe qué representa tanto cada una de las entidades incluidas, como los campos que contienen, con el fin de entender qué información se almacena y el porqué.

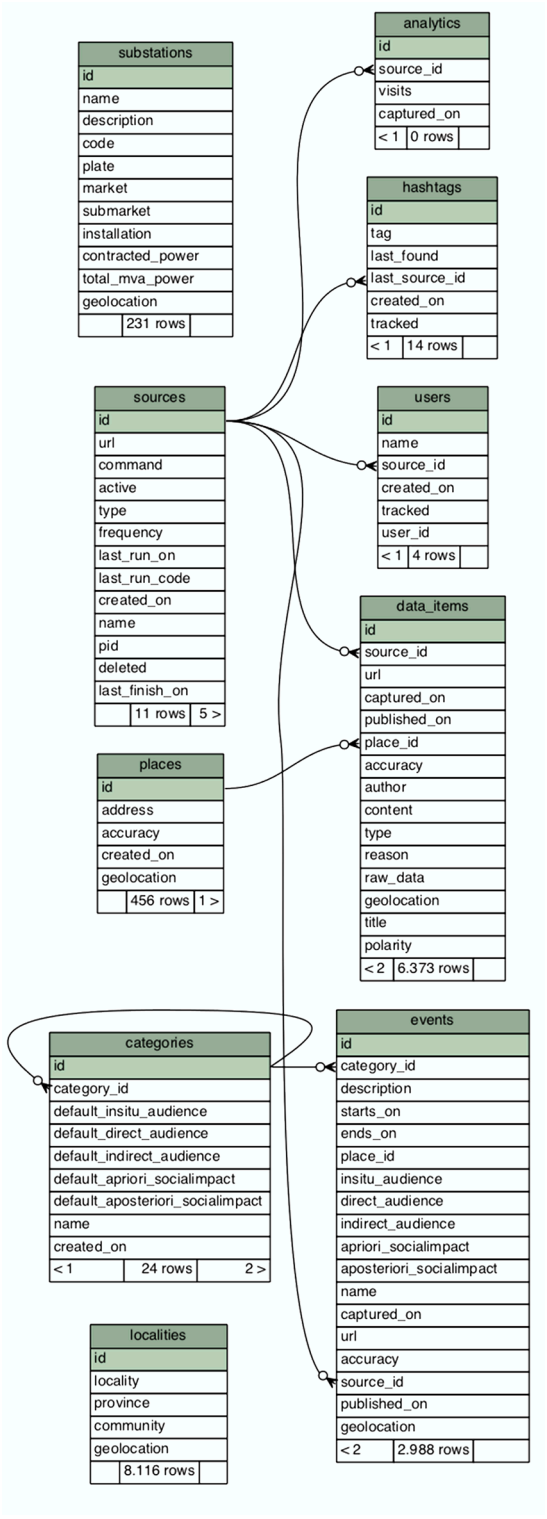

Figura E.1: Diagrama Entidad Relación de la base de datos

Modelo de datos

Introducción

Este documento pretende recoger la estructura que se utilizará en las bases de datos para almacenar toda la información relativa a eventos recogidos de diversas fuentes heterogéneas. En algunos campos, los campos tendrán definidos valores por defecto al no encontrarse la información necesaria disponible en la propia fuente.

Modelo

Tablas que se utilizarán para almacenar los datos:

Analytics (analíticas). Son los datos recogidos de fuentes como Google Analytics que se utilizan para medir el impacto de un evento una vez ha sucedido.

- id: identificador numérico único
- source_id: fuente de la que se han obtenido las analíticas
- captured_on: fecha de capturación de dicho dato
- visits: número de visitas obtenidas

Categories (tipos y subtipos de eventos). Los eventos capturados pertenecerán a distintas categorías dependiendo de la fuente de la que se hayan capturado, y éstas tendrán unos valores por defecto para las distintas variables utilizadas en el modelo de severidad. Éstas podrán ser configuradas desde el interfaz web, y si no se le asigna un valor específico al evento para cada variable, éste será el utilizado. Con esto se pretende facilitar la valoración de los eventos una vez capturados debido a su volumen, permitiendo recalificar todos los eventos de una categoría sin tener que ir uno a uno.

- id: identificador numérico único
- name: nombre de la categoría
- category_id: identificador de la categoría a la que pertenece, si fuera una subcategoría
- default_insitu_audience: número de personas estimado de asistentes físicamente al evento
- default_direct_audience: número de personas estimado de asistentes en directo al evento
- default_indirect_audience: número de personas estimado que hablaran del evento aunque no asistan
- default_apriori_socialimpact: número que representa una estimación del impacto que tiene el evento antes de que suceda
- default_aposteriori_socialimpact: número que representa una valoración del impacto que tiene el evento una vez sucedido
- created_on: fecha en la que se creó el dato

DataItem (elementos obtenidos de las fuentes seleccionadas). Son las noticias, tuits y comentarios que se recogen de las diferentes fuentes y son utilizados para la evaluación del sentimiento de la sociedad una vez un incidente ha sucedido.

- id: identificador numérico único
- source_id: fuente de la que proviene dicho elemento
- author: identificador del autor que ha producido el contenido
- url: url completa del origen de la información
- title: título del contenido, si es que se dispone de él
- content: contenido del elemento limpio
- raw_data: datos en crudo recogidos (texto del tweet, mensaje de Facebook, etc.)
- type: tipo de contenido (texto, noticia, comentario de noticia, etc.)
- published_on: fecha en la que el usuario creó el dato
- captured_on: fecha de capturación de dicho dato
- place_id: identificador que indica dónde se celebra
- accuracy: precisión de la geoposición del lugar en el que se ha publicado el elemento
- links: otros enlaces que puedan contener los datos, se realizará a través de una tabla intermedia
- reason: carácter que indica por qué motivo ha sido capturado el elemento
- geolocation: geoposición del elemento de tipo Geometry de PostGIS (SELECT AddGeometryColumn ('escucha','data_items','geom',4258,'POINT',2);)
- polarity: real que indica la polaridad calculada con el analizador de sentimiento para este elemento

Events (eventos). Son los eventos capturados de las distintas fuentes que se utilizan para prever posibles riesgos en base a su importancia y a sus audiencias.

- id: identificador numérico único
- source_id: fuente de la que proviene dicho elemento
- category_id: identificador que indica la categoría del evento
- name: nombre del evento
- description: descripción del evento para tener más información del mismo
- starts_on: estampilla de tiempo con la fecha/fora de inicio
- ends_on: estampilla de tiempo con la fecha/hora de fin
- captured_on: fecha de capturación de dicho dato
- published_on: fecha en la que el usuario creó el dato
- place_id: identificador que indica dónde se celebra
- url: url completa del origen de la información
- insitu_audience: número de personas estimado de asistentes físicamente al evento
- direct_audience: número de personas estimado de asistentes en directo al evento
- indirect_audience: número de personas estimado que hablaran del evento aunque no asistan
- apriori_socialimpact: número que representa una estimación del impacto que tiene el

evento
- aposteriori_socialimpact: número que representa la importancia final del evento para el sistema
- accuracy: precisión de la geoposición del lugar en el que se ha publicado el elemento
- geolocation: geoposición del elemento de tipo Geometry de PostGIS

Hashtags (palabras clave a buscar). Son las palabras clave que se utilizan para trackear contenidos en las distintas fuentes, y que estos estén relacionados con los fines del proyecto, ya que la cantidad de información accesible es demasiado grande y es necesario filtrar.

- id: identificador numérico único
- tag: cadena con la palabra clave
- last_found: fecha de la última aparición
- last_source: fuente en la que apareció por última vez
- tracked: booleano que indica si ess hashtag está siendo capturada o no
- created_on: fecha en la que el usuario creó el dato

Locality (localidades). Disponemos de información de todas las localidades de España, ya que son utilizadas en algunos de los scrapers como el de Facebook Eventos, que no permite filtrar por coordenadas si no se le ha especificado un campo de texto previamente.

- id: identificador numérico único
- name: nombre de la localidad
- province: nombre de la provincia a la que pertenece
- community: nombre de la comunidad autónoma a la que pertenece
- geolocation: geoposición del elemento de tipo Geometry de PostGIS

Places (lugares de interés). Son los lugares en los que se producen los distintos eventos capturados en las fuentes. Guardar esta información facilita ubicar los eventos si se producen varios en el mismo sitio, y reajustar la relevancia de los mismos si fuera necesario.

- id: identificador numérico único
- geolocation: geoposición del elemento de tipo Geometry de Postgresql (SELECT AddGeometryColumn ('escucha','data_items','geom',4258,'POINT',2);)
- address: texto extraído de la fuente que indica el lugar
- accuracy: precisión de la geolocalización (se utilizará al usar reverse geocoding)
- created_on: fecha en la que el usuario creó el dato

Sources (fuentes de datos). Son las fuentes de datos con toda la información necesaria para su control por parte del sistema, así como de su monitorización a través de la interfaz del mismo.

- id: identificador numérico único
- url: dirección de la web
- command: comando para lanzar el script que recoge datos de la fuente
- active: si el scrapper está recogiendo datos en este momento

- frequency: frecuencia de consulta la fuente de información (0 para fuentes de streaming)
- type: tipo de fuente (streaming, periódica…)
- last_run_on: fecha de última ejecución
- last_run_status: estado de la última ejecución
- created_on: fecha en la que el usuario creó el dato
- name: nombre de la fuente
- pid: identificador del proceso de la fuente
- deleted: si la fuente ha sido borrada, para no dejar huérfanos sus contenidos

Substations (subestaciones). Información facilitada por parte de Unión Fenosa sobre todas las subestaciones de distribución del territorio de Galicia.

- id: identificador numérico único
- name: nombre de la subestación
- description: breve descripción de la subestación
- code: código de identificación interno de la subestación
- plate: matrícula que utiliza Unión Fenosa Distribución para identificar las subestaciones
- market: cadena de texto que indica el tipo de mercado de la subestación
- submarket: cadena de texto que indica el submercado de la subestación
- installation: cadena de texto que indica la región o lugar donde está instalada la subestación
- contracted_power: potencia contratada de la subestación
- total_mva_power: potencia total que soporta la subestación
- geolocation: geoposición del elemento de tipo Geometry de PostGIS

Users (usuarios). Usuarios trackeados para capturar contenidos que son relevantes para el sistema en el apartado de análisis de sentimiento. Entre ellos encontraremos todos los usuarios de redes sociales de Gas Natural Fenosa.

- id: identificador numérico único
- name: nombre que utiliza el autor en las diferentes fuentes
- source_id: fuente a la que pertenece dicho autor (sería interesante distinguir un mismo usuario en distintas fuentes)
- tracked: booleano que indica si ess hashtag está siendo capturada o no
- created_on: fecha en la que el usuario creó el dato

Anexo F

Marco tecnológico

En este capítulo se resumen las diferentes tecnologías y herramientas que se han utilizado en este proyecto, así como los motivos de su elección. Recordamos que, por la filosofía del proyecto, se buscaron siempre alternativas libres, luego la labor de investigación tuvo un gran peso en el desarrollo del mismo.

F.1 Python

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible.

Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, usa tipado dinámico y es multiplataforma.

Es administrado por la Python Software Foundation. Posee una licencia de código abierto, denominada Python Software Foundation License,1 que es compatible con la Licencia pública general de GNU[1] a partir de la versión 2.1.1, e incompatible en ciertas versiones anteriores.

Las razones de su elección son las siguientes:

- Por acuerdo con el consorcio
- Por ser totalmente abierto
- Por ser muy utilizado y extendido
- Por disponer de una comunidad que lo mantiene

[1] GNU's Not Unix

- Por disponer de muchas librerías auxiliares que facilitan mucho el trabajo (acceso a base de datos, desarrollo web, parseo, etc.)
- Por ser muy sencillo su aprendizaje
- Por ser minimalista y rápido a pesar de ser interpretado
- Por ser multiplataforma
- Por la experiencia del autor en el lenguaje

A continuación se detallan algunos de los módulos o frameworks más importantes que se han utilizado en el desarrollo de este TFM.

F.1.1 Django

Django es un framework de desarrollo web de código abierto, escrito en Python, que respeta el paradigma conocido como Model Template View. 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[2] en julio de 2005; el framework fue nombrado en alusión al guitarrista de jazz gitano Django Reinhardt.

En junio del 2008 fue anunciado que la recién formada Django Software Foundation se haría cargo de Django en el futuro.

La meta fundamental de Django es facilitar la creación de sitios web complejos. Django pone énfasis en el re-uso, la conectividad y extensibilidad de componentes, el desarrollo rápido y el principio No te repitas (DRY, del inglés Don't Repeat Yourself). Python es usado en todas las partes del framework, incluso en configuraciones, archivos, y en los modelos de datos.

En este TFM se ha utilizado por su facilidad de uso, por todas las herramientas que ofrece (creación de plantillas, gestión de base de datos, autenticación, etc.), y porque al autor le resultaba interesante aprender un nuevo framework de desarrollo que estuviera en auge.

F.1.2 urllib

Es un módulo que provee una interfaz simple para acceder a recursos en red. En este proyecto nos permite realizar todas las peticiones HTTP que se requieren en los scrapers, enviando parámetros codificados, utilizando sesiones y cookies, modificando las cabeceras necesarias para simular el comportamiento de un usuario, etc.

[2] Berkeley Software Distribution

Además, permite un control de errores en las peticiones para gestionar aquéllas que no van a tener respuesta, las que han dado algún error, o para las que no se tiene permiso de acceso.

Una vez realizadas las peticiones, permite extraer los datos resultantes en forma de cadena, formato que puede ser utilizado para importarlos en otras librerías para realizar búsquedas más finas y precisas.

F.1.3 lxml

Son un conjunto de herramientas que enlazan con las librerías libxml2 y libxslt. Son librerías para el parseo de documentos XML[3] (HTML es un subtipo de XML) y también permiten la creación de documentos y la extracción de datos de los mismos.

En este proyecto se utiliza en casi todos los scrapers para extraer los datos de las fuentes seleccionadas que no proveen una API en JSON o en otro formato más fácilmente legible. Mediante consultas de XPATH son consultados los elementos de los documentos que tienen la información relevante para el proyecto.

Xpath

XPath es un lenguaje que permite construir expresiones que recorren y procesan un documento XML. La idea es parecida a las expresiones regulares para seleccionar partes de un texto sin atributos (plain text). XPath permite buscar y seleccionar teniendo en cuenta la estructura jerárquica del XML. XPath fue creado para su uso en el estándar XSLT[4], en el que se usa para seleccionar y examinar la estructura del documento de entrada de la transformación.

F.1.4 datetime

Es un módulo que permite el manejo de fechas y tiempos de manera muy sencilla intuitiva. En este proyecto se ha utilizado esencialmente para conversiones de cadenas con distintos formatos a un objeto de tipo tiempo, para sumar y restar cantidades de tiempo a una fecha determinada, para realizar consultas a la base de datos, para extraer información de una fecha dada, etc.

[3] eXtensible Markup Language
[4] Extensible Stylesheet Language Transformations

F.1.5 json

Es un módulo para el tratamiento de datos en formato JSON que es un formato ligero para el intercambio de datos. JSON es un subconjunto de la notación literal de objetos de JavaScript que no requiere el uso de XML muy simple de usar, por lo que se ha generalizado mucho su uso durante los últimos años. Algunas de las fuentes utilizadas como Facebook o Twitter proveen APIs para la obtención de datos de las mismas en las que los datos son devueltos en este formato, por lo que es necesario el uso de una librería como ésta para su correcto tratamiento.

F.1.6 re

Es una librería para el tratamiento de expresiones regulares, que se ha utilizado en este TFM en algunas ocasiones para extracciones de datos muy finas en lugares donde lxml no podía llegar, como para la extracción de datos incrustados entre etiquetas script de código HTML.

F.1.7 psycopg2

Es uno de los drivers más utilizados en este proyecto, ya que es el intermediario entre el lenguaje Python y el gestor de base de datos escogido por el consorcio para este proyecto: PostgreSQL. Con este módulo se pueden realizar todo tipo de operaciones con la base de datos de forma segura como si los realizáramos de una propia terminal, además de incorporar algunas herramientas muy interesantes de cara a la programación como las transacciones.

F.1.8 subprocess

Se trata de un módulo que permite lanzar nuevos procesos, conectar su estándar de entrada/salida/error, y obtener los códigos de finalización. Mediante esta librería, el demonio lanza en segundo plano los diferentes scrapers conforme está programado desde la interfaz, de manera que su ejecución no afecte al resto del sistema.

F.1.9 traceback

Es un módulo que permite examinar los errores que han podido surgir en la ejecución de un programa en Python, que resultó muy útil para las pruebas y el debugging de los diferentes scrapers creados.

F.1.10 oauth2client

Librería para gestionar la autenticación OAuth2. Algunas de las fuentes como Facebook o Google Analytics necesitan de este tipo de identificación para la extracción de datos, por lo que fue necesaria su utilización.

F.1.11 logging

Es un módulo que contiene una serie de clases y funciones que permiten implementar un sistema de registro de eventos para cualquier tipo de desarrollo en Python. Ha sido muy utilizado también tanto para las pruebas como para el debugging de las diferentes partes desarrolladas para este proyecto.

F.1.12 xlrd

Es una librería para la gestión de documentos en hojas de cálculo, tanto para la lectura como para la escritura. Ha sido necesario su uso ya que algunos de los datos proporcionados por otros socios se encontraban almacenados en este formato.

F.1.13 csv

Se trata de una librería para la gestión de documentos CSV[5], tanto para la lectura como para la escritura. SentiWordNet se encuentra en formato CSV, por lo que es necesario utilizar este módulo para su carga en memoria.

F.1.14 textblob

Es un módulo para procesar textos. Ofrece muchas posibilidades, algunas de ellas han sido muy importantes para el desarrollo de este proyecto, como la traducción, el POS tagging (etiquetado del papel que desempeña una palabra en una frase), singularización y lematización de palabras, y también dispone de la funcionalidad de obtener la polaridad de una frase como veremos en el anexo H, aunque fue desechada por no estar muy completa y no ofrecer tan buenos resultados como la alternativa implementada en este proyecto.

[5] Comma Separated Values

F.1.15 nltk

Natural Language Toolkit es la plataforma más importante para crear programas en Python que utilizan lenguaje humano. Permite utilizar más de 50 corpus y recursos léxicos, así como una gran variedad de funcionalidades y librerías para la clasificación, tokenización, etiquetado, parseo, etc.

Se ha utilizado para el etiquetado POS de las frases a analizar, ya que el etiquetador que se utiliza necesita uno de sus corpus, pero también se usó para probar la alternativa en castellano SentiWordNet-BC, ya que necesitaba un etiquetado en castellano de palabras.

F.2 PostgreSQL

PostgreSQL es un SGBD[6] relacional orientado a objetos y libre, publicado bajo la licencia BSD.

Como muchos otros proyectos de código abierto, el desarrollo de PostgreSQL no es manejado por una empresa y/o persona, sino que es dirigido por una comunidad de desarrolladores que trabajan de forma desinteresada, altruista, libre y/o apoyados por organizaciones comerciales. Dicha comunidad es denominada el PGDG (PostgreSQL Global Development Group).

Entre otros motivos, ha sido utilizado en este proyecto por la experiencia del consorcio en su uso, por ser de código abierto, y por ser uno de los más apropiados para el manejo de datos georreferenciados mediante su extensión PostGIS[7].

F.2.1 Apache

El servidor HTTP Apache es un servidor web HTTP de código abierto, para plataformas Unix (BSD, GNU/Linux, etc.), Microsoft Windows, Macintosh y otras, que implementa el protocolo HTTP/1.12 y la noción de sitio virtual. Cuando comenzó su desarrollo en 1995 se basó inicialmente en código del popular NCSA HTTPd 1.3, pero más tarde fue reescrito por completo. Su nombre se debe a que Behelendorf quería que tuviese la connotación de algo que es firme y enérgico pero no agresivo, y la tribu Apache fue la última en rendirse al que pronto se convertiría en gobierno de EEUU, y en esos momentos la preocupación de su grupo era que llegasen las empresas y "civilizasen" el paisaje que habían creado los primeros ingenieros de Internet. Además Apache consistía solamente en un conjunto de parches a aplicar al servidor de NCSA. En inglés, a patchy server (un servidor "parcheado") suena igual que Apache Server.

El servidor Apache se desarrolla dentro del proyecto HTTP Server (httpd) de la Apache Software Foundation.

[6] Sistema de Gestión de Bases de Datos
[7] Geographic Information System

Apache presenta entre otras características altamente configurables, bases de datos de autenticación y negociado de contenido, pero fue criticado por la falta de una interfaz gráfica que ayude en su configuración.

Apache tiene amplia aceptación en la red: desde 1996, Apache, es el servidor HTTP más usado. Alcanzó su máxima cuota de mercado en 2005 siendo el servidor empleado en el 70% de los sitios web en el mundo, sin embargo ha sufrido un descenso en su cuota de mercado en los últimos años. (Estadísticas históricas y de uso diario proporcionadas por Netcraft3).

La mayoría de las vulnerabilidades de la seguridad descubiertas y resueltas tan sólo pueden ser aprovechadas por usuarios locales y no remotamente. Sin embargo, algunas se pueden accionar remotamente en ciertas situaciones, o explotar por los usuarios locales malévolos en las disposiciones de recibimiento compartidas que utilizan PHP como módulo de Apache.

En este proyecto se ha usado para servir la interfaz web desarrollada en Python mediante el módulo mod_wsgi[8].

F.3 Eclipse

Eclipse es un programa informático compuesto por un conjunto de herramientas de programación de código abierto multiplataforma para desarrollar lo que el proyecto llama "Aplicaciones de Cliente Enriquecido", opuesto a las aplicaciones "Cliente-liviano" basadas en navegadores. Esta plataforma, típicamente ha sido usada para desarrollar entornos de desarrollo integrados (del inglés IDE), como el IDE de Java llamado Java Development Toolkit (JDT) y el compilador (ECJ) que se entrega como parte de Eclipse (y que son usados también para desarrollar el mismo Eclipse). Sin embargo, también se puede usar para otros tipos de aplicaciones cliente, como BitTorrent o Azureus.

Este IDE ha sido uno de los utilizados por el autor para desarrollar este proyecto, por su versatilidad para trabajar con distintos lenguajes y su integración con diferentes tecnologías.

F.4 VIM

Vim (del inglés Vi IMproved) es una versión mejorada del editor de texto vi, presente en todos los sistemas UNIX.

Su autor, Bram Moolenaar, presentó la primera versión en 1991, fecha desde la que ha experimentado muchas mejoras. La principal característica tanto de Vim como de Vi consiste en que disponen de diferentes modos entre los que se alterna para realizar ciertas operaciones,

[8]Web Server Gateway Interface

lo que los diferencia de la mayoría de editores comunes, que tienen un sólo modo en el que se introducen las órdenes mediante combinaciones de teclas o interfaces gráficas.

Esta herramienta ha sido utilizada para la edición en remoto de ficheros.

F.5 HTML

HTML, siglas de HyperText Markup Language («lenguaje de marcas de hipertexto»), hace referencia al lenguaje de marcado para la elaboración de páginas web. Es un estándar que sirve de referencia para la elaboración de páginas web en sus diferentes versiones, define una estructura básica y un código (denominado código HTML) para la definición de contenido de una página web, como texto, imágenes, etc. Es un estándar a cargo de la W3C[9], organización dedicada a la estandarización de casi todas las tecnologías ligadas a la web, sobre todo en lo referente a su escritura e interpretación. Es el lenguaje con el que se definen las páginas web.

F.6 CSS

Hojas de Estilo en Cascada (Cascading Style Sheets) es el lenguaje de hojas de estilo utilizado para describir el aspecto y el formato de un documento escrito en un lenguaje de marcas, esto incluye varios lenguajes basados en XML como son XHTML[10] o SVG[11].

La información de estilo puede ser adjuntada como un documento separado o en el mismo documento HTML. En este último caso podrían definirse estilos generales en la cabecera del documento o en cada etiqueta particular mediante el atributo "<style>".

F.7 Javascript

JavaScript (abreviado comúnmente "JS") es un lenguaje de programación interpretado, dialecto del estándar ECMAScript. Se define como orientado a objetos,3 basado en prototipos, imperativo, débilmente tipado y dinámico.

Se utiliza principalmente en su forma del lado del cliente (client-side), implementado como parte de un navegador web permitiendo mejoras en la interfaz de usuario y páginas web dinámicas4 aunque existe una forma de JavaScript del lado del servidor (Server-side JavaScript

[9] World Wide Web Consortium
[10] Extensible HyperText Markup Language
[11] Scalable Vector Graphics

o SSJS). Su uso en aplicaciones externas a la web, por ejemplo en documentos PDF[12], aplicaciones de escritorio (mayoritariamente widgets) es también significativo.

JavaScript se diseñó con una sintaxis similar al C, aunque adopta nombres y convenciones del lenguaje de programación Java. Sin embargo Java y JavaScript no están relacionados y tienen semánticas y propósitos diferentes.

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 DOM[13].

Tradicionalmente se venía utilizando en páginas web HTML para realizar operaciones y únicamente en el marco de la aplicación cliente, sin acceso a funciones del servidor. JavaScript se interpreta en el agente de usuario, al mismo tiempo que las sentencias van descargándose junto con el código HTML.

F.7.1 JSON

JSON, acrónimo de JavaScript Object Notation, es un formato ligero para el intercambio de datos. JSON es un subconjunto de la notación literal de objetos de JavaScript que no requiere el uso de XML.

La simplicidad de JSON ha dado lugar a la generalización de su uso, especialmente como alternativa a XML en AJAX. Una de las supuestas ventajas de JSON sobre XML como formato de intercambio de datos en este contexto es que es mucho más sencillo escribir un analizador sintáctico (parser) de JSON. En JavaScript, un texto JSON se puede analizar fácilmente usando la función eval(), lo cual ha sido fundamental para que JSON haya sido aceptado por parte de la comunidad de desarrolladores AJAX, debido a la ubicuidad de JavaScript en casi cualquier navegador web.

F.7.2 XML

XML, siglas en inglés de eXtensible Markup Language ('lenguaje de marcas extensible'), es un lenguaje de marcas desarrollado por el World Wide Web Consortium (W3C) utilizado para almacenar datos en forma legible. Deriva del lenguaje SGML[14] y permite definir la gramática de lenguajes específicos (de la misma manera que HTML es a su vez un lenguaje definido por SGML) para estructurar documentos grandes. A diferencia de otros lenguajes, XML da soporte a bases de datos, siendo útil cuando varias aplicaciones se deben comunicar entre sí o integrar información. (Bases de datos Silberschatz).

[12] Portable Document Format
[13] Document Object Model
[14] Standard Generalized Markup Language

XML no ha nacido sólo para su aplicación para Internet, sino que se propone como un estándar para el intercambio de información estructurada entre diferentes plataformas. Se puede usar en bases de datos, editores de texto, hojas de cálculo y casi cualquier cosa imaginable.

XML es una tecnología sencilla que tiene a su alrededor otras que la complementan y la hacen mucho más grande y con unas posibilidades mucho mayores. Tiene un papel muy importante en la actualidad ya que permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil.

F.7.3 Google Maps API

Google Maps es un servidor de aplicaciones de mapas en la web que pertenece a Google. Ofrece imágenes de mapas desplazables, así como fotografías por satélite del mundo e incluso la ruta entre diferentes ubicaciones o imágenes a pie de calle Google Street View. Desde el 6 de octubre de 2005, Google Maps es parte de Google Local.

En este proyecto se ha utilizado la API V3 de Javascript para mostrar distintos tipos de mapas en la interfaz.

F.7.4 Google Chart API

Es una herramienta muy simple que permite crear gráficas a partir de ciertos datos e integrarlas en una web. Google crea al vuelo una imagen a partir de los datos que se le envían en una petición HTTP que es mostrada en la web del cliente. Permite dibujar diversos tipos de gráficas de los más conocidos (líneas, barras, de tarta, de radar, etc.), y está disponible para diversos lenguajes, aunque en este proyecto se ha utilizado la de Javascript.

F.7.5 jQuery

jQuery es una biblioteca de JavaScript, creada inicialmente por John Resig, que permite simplificar la manera de interactuar con los documentos HTML, manipular el árbol DOM, manejar eventos, desarrollar animaciones y agregar interacción con la técnica AJAX a páginas web. Fue presentada el 14 de enero de 2006 en el BarCamp NYC. jQuery es la biblioteca de JavaScript más utilizada.1

jQuery es software libre y de código abierto, posee un doble licenciamiento bajo la Licencia MIT y la Licencia Pública General de GNU v2, permitiendo su uso en proyectos libres y privativos.2 jQuery, al igual que otras bibliotecas, ofrece una serie de funcionalidades basadas en JavaScript que de otra manera requerirían de mucho más código, es decir, con las funciones propias de esta biblioteca se logran grandes resultados en menos tiempo y espacio.

En este TFM se utilizó para la realización de peticiones AJAX de datos y para el manejo en general del contenido dinámico de las distintas páginas que se muestran en la interfaz.

F.7.6 Twitter Bootstrap

Twitter Bootstrap es un framework o conjunto de herramientas de software libre para diseño de sitios y aplicaciones web. Contiene plantillas de diseño con tipografía, formularios, botones, cuadros, menús de navegación y otros elementos de diseño basado en HTML y CSS, así como, extensiones de JavaScript opcionales adicionales.

Es el proyecto más popular en GitHub y es usado por la NASA y la MSNBC junto a demás organizaciones.

En este proyecto fue elegido por su facilidad para crear interfaces sencillas y vistosas a la par que funcionales.

F.8 Mercurial

Mercurial es un sistema de control de versiones multiplataforma, para desarrolladores de software. Está implementado principalmente haciendo uso del lenguaje de programación Python, pero incluye una implementación binaria de diff escrita en C. Mercurial fue escrito originalmente para funcionar sobre Linux. Ha sido adaptado para Windows, Mac OS X y la mayoría de otros sistemas tipo Unix. Mercurial es, sobre todo, un programa para la línea de comandos. Todas las operaciones de Mercurial se invocan como opciones dadas a su programa motor, hg (cuyo nombre hace referencia al símbolo químico del mercurio).

Las principales metas de desarrollo de Mercurial incluyen un gran rendimiento y escalabilidad; desarrollo completamente distribuido, sin necesidad de un servidor; gestión robusta de archivos tanto de texto como binarios; y capacidades avanzadas de ramificación e integración, todo ello manteniendo sencillez conceptual.1 Incluye una interfaz web integrada.

El creador y desarrollador principal de Mercurial es Matt Mackall. El código fuente se encuentra disponible bajo los términos de la licencia GNU GPL[15] versión 2, lo que clasifica a Mercurial como software libre.

Durante la realización de este proyecto se utilizó un servidor de versiones remoto para mejorar el desarrollo, además de servir de copia de respaldo.

[15] General Public License

F.9 Teambox

Actualmente Redbooth, es una herramienta collaborativa online para el desarrollo y seguimiento de proyectos. Como ya hemos comentado, debido a que hasta la versión 3 esta herramienta es muy completa y abierta, fue la utilizada para la gestión de la documentación y los avances del proyecto.

F.10 Chrome Developer Tools

Son una serie de herramientas que los navegadores web proveen para facilitar el desarrollo de sitios web permitiendo realizar cambios al vuelo, analizar todas las peticiones realizadas, hacer debugging de código Javascript, realizar llamadas de Javascript en tiempo de ejecución, inspeccionar los elementos del DOM, etc.

F.11 Latex

Es un sistema de composición de textos, orientado especialmente a la creación de libros, documentos científicos y técnicos que contengan fórmulas matemáticas.

LaTeX está formado por un gran conjunto de macros de TeX, escrito por Leslie Lamport en 1984, con la intención de facilitar el uso del lenguaje de composición tipográfica, creado por Donald Knuth. Es muy utilizado para la composición de artículos académicos, tesis y libros técnicos, dado que la calidad tipográfica de los documentos realizados con LaTeX es comparable a la de una editorial científica de primera línea.

LaTeX es software libre bajo licencia LPPL.

Anexo G

Palabras y usuarios clave

G.1 Palabras clave

La elección de unos criterios adecuados para la recogida de datos relevantes para su evaluación en este proyecto ha sido una de las partes clave del mismo. La cantidad de información que se puede llegar a recoger de redes sociales, medios de prensa online, etc., es muy grande, ya que millones de usuarios están aportando contenidos cada instante. El principal problema de la fase de obtención de contenidos es filtrar los que realmente son relevantes, ya que el ruido capturado puede ser muy grande dependiendo de los criterios escogidos, y es algo que hay que tener muy en cuenta para que no afecte a los resultados.

Para ello, por un lado se recogen todos los tweets geoposicionados en España, y de estos se filtran exclusivamente los de la zona de galicia con un pequeño margen de unos 100Km., distancia acordada con los expertos de Unión Fenosa como suficiente para captar toda la región que puede verse afectada por un gran apagón. Los tweets geoposicionados se encuentran en torno al 8% del total, y ya son una muestra suficientemente representativa. En cuanto a noticias, su geoposición es más complicada, ya que normalmente se dispone fácilmente de la ciudad en la que han sucedido, pero hilar más fino y buscar lugares más concretos o direcciones es mucho más complicado. Además, como se trabaja a nivel de subestación como ya hemos explicado, con saber la ciudad es suficiente para ubicar el incidente.

Por otro lado, para la captura de contenidos en base a palabras clave, hubo que realizar un proceso para su elección que llevó cierto tiempo refinar. Inicialmente, se diseñó una lista base conjuntamente con los expertos de Unión Fenosa. Se decidió dejar en marcha el scraper encargado de esta tarea, para posteriormente analizar los resultados. Tras una semana en funcionamiento, las frecuencias de aparición fueron las siguientes:

- **luz:** 901.270 veces, se capturaron demasiados resultados no relacionados con el proyecto, por lo que se decidió capturar combinaciones de otras palabras con ella

- **apagon:** 6.711 veces, el hashtag que mejor funcionó de todos los iniciales
- **electricas:** 5.463 veces, también funcionó muy bien, aunque se mezcló con temas de facturas, precios, etc.
- **energia:** 2.642 veces, similar al anterior, por lo que se decidió utilizar junto con otras palabras como "sin energía", "corte de energía"
- **fenosa:** 214 veces, palabra clave de obligada captura al formar parte del nombre de la empresa
- **gasnatural:** 66 veces, igual que el anterior
- **vendaval:** 2.880 veces, en Galicia hay muchos temporales, que afectan muchas veces a las instalaciones, por lo que pensamos que sería interesante capturarlo, pero generaba mucho ruido
- **olas:** 85.655 veces, sin lugar a dudas el peor hashtag de todos. Se introdujo pensando en los temporales, pero la gente lo utiliza como saludo, e introduce tanto ruido y aporta tan poco que se decidió eliminarlo
- **temporal:** 104.436 veces, aunque este término funciona mejor que el anterior, se decidió eliminar al ser muy general, e introducir otros más concretos en una segunda fase
- **poste:** 21.971 veces, también muy genérico, pasaban el filtro muchos temas relacionados con el fútbol
- **generador:** 7.340 veces, este se ha mantenido ya que no funcionaba del todo mal
- **tormenta:** 235.769 veces, también demasiado general, se eliminó y se pasaron a usar términos más concretos como rayo o relámpago
- **huracan:** 9.825 veces, se decidió eliminar, ya que hay muy pocos en España y los usuarios lo utilizan con muchos otros fines
- **electricidad:** 70.698 veces, como energía, es demasiado general, y hay que utilizarlo junto con otros términos como "sin electricidad"
- **electrico:** 13.388 veces, igual que en el caso anterior, se decidió usar "corte eléctrico"
- **GNF:** 263 veces, todos los hashtags relacionados con Unión Fenosa es necesario capturarlos
- **gas natural:** 2.507 veces, ídem al anterior, aunque con él se captura algo de ruido, se decide mantenerlo

Como las palabras clave son a nivel general, hubo que pensar en un mecanismo para saber si los incidentes a los que hacen referencia los tweets descargados con este criterio se ubicaban

en Galicia (las noticias no es necesario, ya que se descargan de secciones que pertenecen a Galicia). El método que se empleó finalmente fue la comparación del volumen de tweets recogidos geoposicionados en Galicia con respecto al resto de España, de esta manera, si la proporción de tweets geoposicionados que hablan sobre estos temas en Galicia aumenta con respecto a la proporción de España, sabremos que hacen referencia a sucesos de Galicia.

Después de recoger un volumen tan grande de datos (casi 2 millones en una semana) con mucho ruido, se decidió afinar la lista de términos. En una segunda iteración realizando algunos de los cambios comentados en la lista anterior e introduciendo algunos términos nuevos, los resultados quedaron de la siguiente manera:

- **corte de luz:** 3.122 veces, muy buenos resultados
- **sin luz:** 50.024 veces, buenos resultados
- **ido la luz:** 2.168 veces, muy buenos resultados
- **fue la luz:** 11.876 veces, muy buenos resultados
- **salto la luz:** 1 vez, pocos resultados
- **apagon:** 1.914 veces, buenos resultados
- **electricas:** 1.975 veces, buenos resultados
- **energia:** 47.876 veces, bastante ruido, conviene añadir “sin”
- **fenosa:** 83 veces, muy buenos resultados
- **gasnatural:** 68 veces, muy buenos resultados
- **vendaval:** 1.971 veces, demasiado ruido
- **relámpago:** 110 veces, demasiado ruido
- **trueno:** 16.135 veces, demasiado ruido
- **rayo:** 61.561 veces, demasiado ruido
- **inundación:** 40 veces, demasiado ruido
- **poste de luz:** 1.353 veces, buenos resultados
- **poste de electricidad:** 34 veces, buenos resultados
- **generador:** 5.364 veces, buenos resultados
- **corte de electricidad:** 111 veces, muy buenos resultados
- **ido la electricidad:** 24 veces, muy buenos resultados

- **sin electricidad:** 2.697 veces, muy buenos resultados
- **corte eléctrico:** 0
- **GNF:** 265 veces, muy buenos resultados
- **gas natural:** 8.816 veces, muy buenos resultados
- **transformador:** 6.576 veces, buenos resultados
- **cuadro eléctrico:** 15 veces, buenos resultados
- **fusible:** 659 veces, mucho ruido
- **magnetotérmico:** 0

Tras ver estos resultados, se decidió eliminar las palabras clave relacionadas con tormentas o fenómenos meteorológicos que producían demasiado ruido, y la lista definitiva quedó así:

- **corte de luz**
- **sin luz**
- **ido la luz**
- **fue la luz**
- **saltó la luz**
- **apagon**
- **electricas**
- **fenosa**
- **gasnatural**
- **gasnaturalfenosa**
- **poste de luz**
- **poste de electricidad**
- **generador**
- **corte de electricidad**
- **ido la electricidad**
- **sin electricidad**

- **sin energía**
- **corte electrico**
- **GNF**
- **gas natural**
- **transformador**
- **cuadro eléctrico**

Notar que las palabras clave hay que indicarlas en minúsculas y sin signos de puntuación, y el sistema ya se ocupa de obtener todas las variantes.

Tras estas iteraciones y ciñéndonos a estos términos, el volumen de captura está en torno a 100.000 elementos por semana, bastante limpios de ruido, lo que facilita mucho el almacenamiento y la evaluación de los mismos.

G.2 Usuarios clave

También creímos interesante capturar todo el tráfico de mensajes generado por referencias a las cuentas oficiales de Twitter de Unión Fenosa. Los usuarios a seguir elegidos junto con GNF finalmente fueron:

- UNIONFENOSAnews: cuenta oficial donde Unión Fenosa cuelga todo tipo de noticias
- Gas_Fenosa: cuenta general del grupo Gas Natural Fenosa
- GNF_es: cuenta a nivel nacional de Gas Natural Fenosa
- GNFclientes_es: probablemente la cuenta más importante ya que la usan los clientes para comunicarse con la empresa en cuanto a problemas o incidencias

En todos los casos se trata de cuentas que como mucho reciben unas pocas decenas de tweets al día, y sabemos que todos los que se capturan van a estar directamente relacionados con la empresa, posibles incidentes, y su imagen.

Anexo H

Evaluación de otros sistemas de análisis de sentimiento

En este anexo se presentan y analizan distintas alternativas al sistema implementado en este TFM para la extracción de opinión de textos, justificando con sus resultados el uso de las técnicas y tecnologías utilizadas.

H.1 SentiWordNet-BC

SentiWordNet-BC es una librería en Python para el análisis semántico en castellano, que usa un recurso léxico propio similar a SentiWordNet. Como hemos comentado, SentiWordNet se encuentra únicamente disponible en inglés, por ello una alternativa es elaborar un recurso léxico propio en el idioma deseado. Esto tiene el problema de que es muy costoso, por lo que al probarlo, nos hemos encontrado con que le faltaban muchos términos, lo que llevaba a una evaluación menos correcta en muchas frases.

Como ventaja, está preparado para soportar varios idiomas siguiendo el mismo formato, así como para instalarlo como servicio web.

H.2 Mr. Tuit

Es una de las pocas herramientas online que se pueden probar para analizar sentimientos en textos cortos en castellano. Es comercial, pero ofrece una versión demo que se ha de ejecutar de forma manual mediante una interfaz web muy sencilla. Lo utilizamos en la comparativa, ya que creíamos que era interesante tener la referencia de una herramienta comercial.

H.3 LingPipe

Es un conjunto de herramientas realizadas en Java para el procesamiento de textos. Está basado en técnicas heurísticas, pero según los trabajos estudiados, SentiWordNet mejora los resultados de este tipo de sistemas, al igual que sucede con los basados únicamente en machine learning.

H.4 TextBlob basado en NLTK

Se trata de una librería realizada en Python que funciona de forma similar a la anterior. Fue elegida para la comparativa por basarse en técnicas parecidas a la anterior, pero al estar implementada en el lenguaje de programación utilizado en el proyecto, su integración para realizar la prueba fue mucho más sencilla.

H.5 Conclusiones

Para probar los resultados obtenidos, se comparan el sistema desarrollado en este proyecto con el basado en NLTK, SentiWordNet-BC y Mr. Tuit. Para ello se seleccionaron al azar 101 tweets de los recogidos por nuestro sistema. Se utilizan 101 por resultar un número representativo y a su vez manejable para un humano para ver si los resultados se corresponden con la realidad.

Para realizar este estudio se creó un script en Python que lee las frases de un fichero y las evalúa con las distintas alternativas, salvo para Mr. Tuit, que se hizo manualmente. Para SentiWordNet-BC hubo que realizar una tarea adicional, que consistió en configurar un etiquetador POS específico para el castellano, que es el idioma en que esta herramienta trabaja. Para ello se probaron tres alternativas diferentes, entrenándolos con el 90% de los datos del corpus en español que provee NLTK, y probando los resultados contra el 10% restante. Finalmente usamos el que mejores resultados obtuvo, el BigramTagger (UnigramTagger: 87.6 %, BigramTagger: 89.4 %, TrigramTagger: 89.0 %).

En la tabla H.1 podemos ver los resultados finales de la comparativa, y en la tabla H.2 observamos cuán diferentes son los resultados obtenidos por los distintos analizadores en total.

#	Frase	SWN	NLTK	SWNBC	Mr. Tuit
1	qué calor insomnio y sin luz	0.0425	0.4	0.25	Neg
2	ingenieros diagnostican y corrigen fallas sistema inyeccion electricas	-0.0034	0.0	0.0	Neg

3	rt @fotohorrorosa: si te cae un rayo y sigues con vida esta es la marca que se te queda grabada en el cuerpo	0.0164	0.1	-0.25	Neg
4	nuevas tecnologias para ahorrar energia en smartphones y tablets	0.0848	0.1364	0.0	Pos
5	maría antonia romero: "la redes eléctricas de tenerife y la gomera estarán unidas"	0.0106	0.0	0.0	Neu
6	@tonyalvarez32 si os quedáis sin luz debéis contactar con la distribuidora de la zona. si lo necesitas, podemos ayudarte a saber cuál es	0.0079	0.4	0.25	Pos
7	rt @elmundoinnov: @itenergia desarrolla baterías de litio con huesos de aceituna reciclados	-0.0469	0.0	0.0	Neg
8	¡que sea un lunes cortito y lleno de energia!	0.0657	0.4375	0.0	Neg
9	sin electricidad	0.0	0.0	0.375	Neu
10	lunes y muy feliz, al final siempre hay un rayo de luz que nos alumbra	0.1059	0.7	0.25	Neg
11	#lapazyaviene ahorrador energia electrica disminuye consumo 10% a 30%	0.005	-0.4	0.0	Neu
12	en @elconfidencial: canal+ y el 'apagón tdt' amenazan con dinamitar la tregua mediaset-atresmedia	-0.0554	0.0	0.0	Neg
13	todo el día sin luz en el instituto... estará interesante	0.0693	0.45	0.25	Pos
14	@tiquetaki tu si que vales gracias te mando mucha energia	0.0652	0.25	0.125	Pos
15	sin agua hace 48 horas sin luz desde las 3 am y este gobierno jodiendo a esta hora estudiantes	0.056	-0.6	0.25	Neg
16	rt @bobmarleytweets: "porque en esta vida no hay luz sin oscuridad."	0.0349	-0.2	0.25	Pos
17	el gas licuado por canalización subirá un 3,6% a partir de mañana	-0.0083	0.0	-0.125	Amb
18	comercial gas natural con contrato laboral y alta s.s.	0.0009	0.0867	0.0	Neu

19	rt @gerardotc: felipe lo mismo te monta un gal, que te diseña una joya, que te hace de consejero de gas natural, que te dice quiénes son lo...	0.0117	0.05	0.0	Pos
20	rt @famelica_legion: felipe gonzález, el consejero de gas natural, suelta gases...	-0.0023	0.1	0.0	Neu
21	rt @yoddio: pertenecer al consejo de dirección de gas natural, le ha hecho olvidar muy rápido a felipe gonzález su chaqueta de pana bolivar...	-0.0015	0.18	0.1875	Neg
22	en junio subirán las tarifas eléctricas, dice el ministro para la energía eléctrica	0.0353	0.0	0.0	Pos
23	rt @calvoesperanza: en #garoña están divididos. algunos tienen miedo a la energia nuclear. otros, al paro	-0.1616	-0.3625	0.0	Neg
24	rt @_gustavot_: juré que iba a despertar y vería lluvia, pero lo único que veo es un rayo de sol entrando en mi pieza	0.0331	0.0	0.125	Neg
25	peligro sin...	-0.54	0.0	-0.625	Neg
26	@analuzalberto claaaaro que si, energia positivas para ti mami luz te quiero	0.2201	0.3758	0.25	Pos
27	rt @mario_castro10: utilizar gas natural incrementa la calidad del aire. la #reformaenergética cuida al #medioambiente	0.0383	0.1	0.0	Pos
28	#lareina sin luz @alertachilectra confirma corte por poste chocado en las calles alcala de henares con monseñor edwards.	-0.0238	0.4	0.25	Neg
29	corpoelec: ineficiencia a máxima revolución! 8 horas sin luz	-0.025	0.4	0.25	Neg
30	buenos dias amigos! aqui voy con animo pronto *voy desvelada sin luz toda la noche sin dormir por la calor pero..aqui vamos	0.0494	0.6375	-0.0833	Neg
31	desde la 1:00 pm del día de ayer zona rural de villamaria sin luz, que pasa con el servicio	-0.0017	0.2	0.25	Amb
32	la luz y la ilusión se fue con el adioooos	-0.0429	0.4	0.25	Pos
33	que bajón sin luz!!! :@	-0.0113	0.5	0.25	Neg

34	hoy me levante con mas ganas de verte, tu me das energia	0.0363	0.5	0.0	Pos
35	me cago en la madre se fue la luz	-0.0777	0.1	0.25	Neg
36	y se fue la luz en mi comunidad... wiiii a la mierda todo!!!	0.0202	0.5	0.25	Neg
37	rt @thomasdangel: rt @nievesevelui: @corpoelecinform en el cafetal ayer se fue la luz durante 22 horas!en la urb parque naguanagua se fue 6	0.0079	0.0	0.25	Amb
38	rt @eluniversal: varios sectores de san antonio están sin luz desde anoche	-0.0369	0.0	0.25	Pos
39	anoche les dije que llegaria a los 1000 tweets pero no cumpli porque hubo un apagon anoche.	-0.0574	0.0	0.0	Neu
40	que chevere..! sin luz y sin agua	-0.0487	0.4	0.25	Neg
41	pulso: ultimátum a eléctricas: gobierno formula nuevos cargos por faltas a la norma de emisión	0.0062	0.0682	-0.125	Neg
42	y encima sin luz en el edificio...	0.0284	-0.1	0.25	Neg
43	se fue la luz :(	0.0481	-0.175	0.25	Amb
44	sin luz en el peor momento.	-0.1074	-0.6	0.25	Neg
45	ahora resulta que nuestro secretario de energia esta en el negocio de las gasolineras... pero el dice que no es conflicto de intereses	0.007	0.0	0.0	Pos
46	mi casa sin luz en el peor momento.	-0.0806	-0.6	0.1875	Neg
47	sin luz en las oficinas de mi trabajo. gracias dios!	0.0178	-0.25	0.1875	Pos
48	podemos debería fichar al consejero de gas natural, le hace una propaganda cojonuda.	-0.0675	0.0	0.125	Neg
49	puedo sentir una energia tan intensa entre los dos	0.0044	0.1	-0.625	Neg
50	si comes algo que a sufrido esa energia te contaminara	-0.0377	0.0	0.0	Neg
51	corte de luz y fuerza. todas caras de la barra del loco jajajajajajajaa	-0.0073	-0.6	0.25	Neg
52	las renovables ahorrarán 50.000 millones más que lo recibido en primas en su vida útil	0.0428	0.5	0.0	Pos
53	repsol agrupa en una sociedad su participación de gas natural fenosa	-0.0146	0.1	0.0	Neu

54	como se fue todo el equipo al carajo...ahora tb el lobo..a arrancar de cero en todo sentido...el ultimo q apague la luz	-0.0041	0.2	0.25	Neu
55	rt @bpieroo: hasta cuando el corte de luz este!	0.0316	-0.05	0.25	Neg
56	@reportevln sin luz en la urb la granja	0.0284	0.4	0.25	Amb
57	los corte de luz y lpm, y yo con los batidos de tortas.. se me va a levantar el día del pedo!	-0.0169	0.0	0.1875	Neg
58	rt @kateriine_70: gracias, señor... por mis ojos perfectos,cuando hay tantos sin luz.	0.0387	0.75	0.1875	Pos
59	rt @diario_ecologia: ¿cómo consumir menos energía y ahorrar dinero en casa?	0.0464	-0.1667	0.125	Pos
60	reconociendote que eres mi pasión, mi cruz y que adoro tu silueta sin ropa y con poca luz	0.1112	0.1333	0.25	Neg
61	@1d_ismyworldd_ lo odio por dios !!!! ojalá que le caiga un rayo !!!!	-0.1718	-1.0	-0.375	Neg
62	como es posible que este país pueda durar más de 4,5 horas sin luz, por falta de personal que resuelva los problemas.	-0.0527	0.45	0.0	Neg
63	que necesidad de hacernos estudiar sin luz	0.0253	0.4	0.25	Neg
64	que desgracia con en este "regimen" se fue la luz en la granja valencia!!! @electroprotesta	-0.0166	0.0	-0.0625	
65	tringali admite que el apagón ha beneficiado a mediaset frente a atresmedia pero prepara demanda al gobierno	0.0789	0.0	0.0	Neg
66	los discos ssd podrian ser 300% mas rápidos y reducir en un 60% el uso de energía.	0.0584	0.0	0.125	Pos
67	como odio este tema de se fue la luz , que horrorrrrrrrrrrr	-0.1342	-0.8	-0.0625	Neg
68	tu canción @luztutita ...se fue la luz en todo el barrio (8) justo que apareci no tay....lloraré :´(	-0.0589	-0.25	0.25	Pos

69	rt @sonmicrocuentos: en el momento que apagaron la luz y se quedaron a oscuras fue cuando se vieron realmente por primera vez. bellota.	0.0188	0.14	0.25	Neg
70	@pcmoax atiende puntualmente,derrame de aceite en transformador av.san felipe	-0.0169	0.0	0.0	Pos
71	otro día más sin luz y agua	0.0344	0.4	0.25	Neg
72	@sassfiona y granizos y corte de luz	0.0455	0.0	0.25	Neg
73	#ofertasgasluz luz + gas > endesa energia s.a.u. tarifa luz + gas endesa 3.1 + 2.0a 3 sin calefacción≤	0.0099	0.4	0.25	Neu
74	@jmgomezm que lo diga una web de rusia, 1er suministrador de petróleo y gas natural para españa, me produce una sensación de credibilidad...	0.0204	0.1	0.0	Amb
75	¿agua + energía + cambio climático? http://t.co/hirvf425uw post de @alberto_guijarr en el blog de #nexoaecc	0.021	0.0	0.125	Pos
76	el aburrimiento de felipe gonzález en gas natural le ha llevado a chochear demasiado.	-0.0109	0.1	0.0	Neg
77	y el internado amanecio sin internet, sin agua y con la luz muy baja	-0.0175	0.2	0.25	Neg
78	@olgalosad este hombre es repugnante,se aburre de su trabajo en gas natural!manda cojones,quién lo vería en un andamio!!patético	-0.0107	-0.5859	0.0	Neu
79	@alertachilectra sin electricidad en pasaje dragones de la reina 567!	-0.0073	0.0	0.375	Amb
80	que liquidado llego a los jueves, poca energia	-0.0197	-0.1875	0.0	Neu
81	soy energia oscura k se opone a la norma establecida mal impuesta por la ciencia ekiboca	-0.0641	-0.425	0.0	Neg
82	yo no estoy en barcelona y mira, putazas. mal rayo os parta, de verdad.	-0.1427	-0.25	0.5	Neg
83	de la chaqueta d pana a consejero en gas natural	-0.0296	0.1	0.0	Neu
84	@edenor dejate de joder y veni a arreglar la luz, 2 dias ya, 10 pisos x escalera y ahora sin agua	0.0035	0.4	0.1875	Neg

85	@cfe_vallemex seguimos sin luz el numero de reporte es 8324937305	0.0051	0.4	0.25	Neu
86	hasta cuando el corte de luz este!	0.0051	0.0	0.25	Neg
87	#recomendación sobre #gasnatural, atento a estos síntomas! @defensacivilmlp	0.1793	0.5	0.5	Pos
88	pleno siglo xxi y viviendo en la era arcaica... sin luz y agua	-0.0291	0.0	0.25	Neg
89	@daanifernnandez con la suerte que tengo seguro que me parte un rayo :(jajaja	0.0489	0.2208	0.0	Neg
90	rt @_disidente: escuchar al consejero de #gasnatural intentar dar lecciones de democracia provocan vómitos instantáneos #felipegonzalez	-0.0024	0.0	-0.1875	Neg
91	@prdoming esta es de energia renovable y la otra tb habiamos salido como el pais con menos contaminacion ambiental	0.0023	-0.0972	0.0	Neg
92	varias calles sin luz publica por favor solucionar el asusto	-0.0612	0.0	0.1875	Pos
93	segimos sin electricidad :'(	0.0682	-1.0	0.375	Neu
94	curso online técnico en electricidad y electrónica del automóvil. sistemas de encendido. inyección	0.0122	0.0	0.375	Neu
95	12 días sin ver la luz pero bueno tendré mi recompensa!!	0.0566	0.5	0.4583	Pos
96	y cuando se me pudra toda la comida q estaba en el refri despues de las 8hrs sin luz,a cual sucursal de #cfe paso por mis vales de despensa?	0.001	0.4	0.25	Neg
97	puso sin agua ni luz creo que de confundió de tt	-0.0246	-0.4	0.25	Neg
98	"quien de joven no es de izquierdas no tiene corazón. quien de mayor no se sienta en el consejo de gas natural no tiene cabeza".	-0.0776	0.075	0.125	Neg
99	susurrándote un poco más lento, puedo estar años sin agua, sin luz ni viento	-0.0591	-0.1875	0.25	Neg
100	buen dia.. energia positiva	0.262	0.4636	0.0	Pos

101	puta luz! que coño pasa? me va a joder la play o la tv y mato al de fenosa mañana o el que sea. el microondas y eso me la pela....	-0.0018	0.05	0.25	Neg

Cuadro H.1: Comparativa de resultados de diferentes analizadores de sentimiento

	Positivas	**Negativas**
SWN	54	44
NLTK	49	23
SentiWordNetBC	59	10
Mr. Tuit	25	53

Cuadro H.2: Resultados totales de la comparativa de diferentes analizadores de sentimiento

Tras observar los resultados, la única conclusión que podemos extraer es que todos obtienen valores bastante diferentes para una misma frase. De forma totalmente subjetiva, y en base a los criterios del autor, parece que la herramienta desarrollada en este proyecto obtiene unos resultados razonables. No se ha incluido una columna con el análisis de un humano, ya que al consultar el autor con otras personas sobre diferentes frases, el resultado también variaba bastante, sobre todo a la hora de diferenciar si algo era neutro o no.

Para intentar realizar una evaluación más contrastada, hemos probado una muestra aleatoria de 1000 elementos del conjunto de opiniones sobre películas de Cornell Movie Review Data en el que hay 5331 frases positivas y 5331 negativas, comparando el analizador de sentimiento desarrollado en el proyecto con el basado en NLTK, ya que según los resultados obtenidos, nos pareció el mejor de los evaluados. Además, de todas las reviews disponibles, sólo evaluamos aquellas que están en inglés, con el fin de no sesgar el resultado con esta variable.

Los resultados del test con frases positivas se pueden encontrar en la tabla H.3.

	Positivas	**Negativas**
SWN	74%	17%
NLTK	69%	14%

Cuadro H.3: Comparativa de resultados utilizando el conjunto de datos de Cornell sobre frases positivas

Con estos resultados podemos concluir que aunque para frases positivas nuestro sistema obtiene algunos falsos negativos más que el basado en NLTK, tiene un porcentaje mayor de acierto.

Los resultados del test con opiniones negativas se pueden encontrar en la tabla H.4.

	Positivas	**Negativas**
SWN	40%	48%
NLTK	43%	31%

Cuadro H.4: Comparativa de resultados utilizando el conjunto de datos de Cornell sobre frases negativas

En cambio, las frases negativas parece que funcionan peor en los dos sistemas, dando ambos una tasa alta en falsos positivos, aunque un poco menor en el nuestro. Sin embargo, nuestro sistema consigue un porcentaje de acierto mucho mayor catalogando como negativas las frases que realmente lo son.

Notar que analizándolas como humano, tanto el autor de este TFM como otras personas a las que se ha consultado, se han encontrado algunas de las frases en ambos corpus difíciles de clasificar en alguno de los dos grupos, ya que el resultado es muy subjetivo.

Acrónimos

TFM Trabajo Fin de Máster

BIFI Instituto de Biocomputación y Física de Sistemas Complejos

SEPS Sistema Experto de Probabilidad y Severidad en Red

HW Hardware

SW Software

IDE Integrated Development Environment

LTS Long Term Support

HTML HyperText Markup Language

CSS Cascading Style Sheets

JSON JavaScript Object Notation

SQL Structured Query Language

API Application Programming Interface

HTTP Hypertext Transfer Protocol

XPATH XML Path Language

URL Uniform Resource Locator

AJAX Asynchronous JavaScript And XML

POS Part Of Speech

UTF8 8-bit Unicode Transformation Format

GNU GNU's Not Unix

BSD Berkeley Software Distribution

XML eXtensible Markup Language

XSLT Extensible Stylesheet Language Transformations

CSV Comma Separated Values

SGBD Sistema de Gestión de Bases de Datos

GIS Geographic Information System

WSGI Web Server Gateway Interface

W3C World Wide Web Consortium

XHTML Extensible HyperText Markup Language

SVG Scalable Vector Graphics

PDF Portable Document Format

DOM Document Object Model

SGML Standard Generalized Markup Language

GPL General Public License

Bibliografía

[1] Madsen, Edith and Mulalic, Ismir and Pilegaard, Ninette: A model for estimation of the demand for on-street parking. Munich Personal RePEc Archive, 2013.

[2] GEOK K. KUAH. Estimating Parking Demand for Mixed-Use Developments Subject to TSM Ordinances. ITE Journal. Febrero de 1991.

[3] Choy Peng NG, Dadang Mohamad MA'SOEM. The Development Of Model Estimation To Determine Paring Needs At LRT Stations In Suburban Area. Proceedings of the Eastern Asia Society for Transportation Studies, Vol. 5, pp. 877 - 890, 2005.

[4] Nicholas J. Garber, Hua Wang. Demand for Commercial Heavy Truck Parking on Interstate Highways: A Case Study of I-81 in Virginia. TRB 2003 Annual Meeting.

[5] Christopher F. Dumas, John C. Whitehead, James H. Herstine, Robert B. Buerger, Jeffery M. Hill. Estimating Peak Demand for Beach Parking Spaces Under Capacity Constraints. 2005, working paper University of North Carolina-Wilmington.

[6] Saif M. Mohammady and Tony (Wenda) Yang, Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. arXiv:1309.6347v1 [cs.CL] 24 Sep 2013.

[7] Javier Toret, @Datanalysis15m, Antonio Calleja, Óscar Marín Miró, Pablo Aragón, Miguel Aguilera. Tecnopolítica: la potencia de las multitudes conectadas. El sistema red 15M[1], un nuevo paradigma de la política distribuida. IN3 Working Paper Series, Internet Interdisciplinary Institute. 2013.

[8] Javier Borge-Holthoefer, Alejandro Rivero, Iñigo García, Elisa Cauhé, Alfredo Ferrer, Darío Ferrer, David Francos, David Iñiguez, María Pilar Pérez, Gonzalo Ruiz, Francisco Sanz, Fermín Serrano, Cristina Viñas, Yamir Moreno. Structural and Dynamical Patterns on Online Social Networks: The Spanish May 15th Movement as a Case Study. PLoS ONE 6(8): e23883, 2011.

[9] González Bedia, M., y García Carrasco, J. (2006). Arquitecturas emocionales en Inteligencia Artificial : una propuesta unificadora. [Versión electrónica]. "Teoría de la Educación : educación y cultura en la sociedad de la información", 7 (2), 156-168.

[1] Denominación que recibió el movimiento social de protesta del 15 de mayo de 2011

[10] Yi-jie Tang, Hsin-Hsi Chen, Mining Sentiment Words from Microblogs for Predicting Writer-Reader Emotion Transition. LREC 2012: 1226-1229.

[11] Rui Fan, Jichang Zhao, Yan Chen and Ke Xu, Anger is More Influential Than Joy: Sentiment Correlation in Weibo. arXiv:1309.2402. Septiembre 2013.

[12] Proceedings of the Workshop on Semantic Analysis in Social Media, 13th Conference of the European Chapter of the Association for Computational Linguistics.

[13] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, Sentiment Strength Detection in Short Informal Text. Journal of the American Society for Information Science and Technology. Vol 61 Issue 12.

[14] Bo Pang and Lillian Lee, Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. Vol. 2, No 1-2 (2008) 1–135.

[15] Andea Esuli and Fabrizio Sebastiani, SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06).

[16] Marco Guerini, Lorenzo Gatti, Marco Turchi. Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. arXiv:1309.5843v1 [cs.CL] 23 Sep 2013.

[17] Monalisa Ghosh, Animesh Kar. Unsupervised Linguistic Approach for Sentiment Classification from Online Reviews Using Sentiwordnet 3.0. International Journal of Engineering Research & Technology. Vol.2 - Issue 9 (September - 2013).

[18] Aurangzeb khan, Baharum Baharudin, Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs. International Journal of Computer Science & Emerging Technologies. Vol 2, No 4 (2011).

[19] Kerstin Denecke, Using SentiWordNet for Multilingual Sentiment Analysis. Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference.

[20] Alaa Hamouda, Mohamed Rohaim. Reviews Classification Using SentiWordNet Lexicon. The Online Journal on Computer Science and Information Technology (OJCSIT) Vol. (2) - No. (1).

[21] SentiWordNet : `http://sentiwordnet.isti.cnr.it/`

[22] Mr. Tuit : `http://www.mrtuit.com/`

[23] GRIAL Corpus SenSem : `http://grial.uab.es/sensem/corpus?idioma=es`

[24] Petra Tag - Spanish POS Tagger: `http://sourceforge.net/projects/petrapostagger/`

[25] SentiWordNet-BC: `https://github.com/rmaestre/Sentiwordnet-BC`

[26] Microsoft Bing Translation API: http://www.microsoft.com/web/post/using-the-free-bing-translation-apis

[27] A good POS tagger in about 200 lines of Python: http://honnibal.wordpress.com/2013/09/11/a-good-part-of-speechpos-tagger-in-about-200-lines-of-python/

[28] Historical Attendances Spain matches: http://www.european-football-statistics.co.uk/attn/2000/aveesp.htm

[29] Histórico de asistencias a Balaidos : http://foro.delcelta.com/

[30] What are the most powerful open-source sentiment-analysis tools?: http://breakthroughanalysis.com/2012/01/08/what-are-the-most-powerful-open-source-sentiment-analysis-tools/

[31] 10 Tips for Sentiment Analysis projects: http://blog.datumbox.com/10-tips-for-sentiment-analysis-projects/

[32] Export google analytics data via API with Python: http://zonca.github.io/2013/08/export-google-analytics-data-via-api.html

[33] Cornell Datasets: http://www.cs.cornell.edu/people/pabo/movie-review-data/

[34] TEXT CLASSIFICATION FOR SENTIMENT ANALYSIS – NAIVE BAYES CLASSIFIER: http://streamhacker.com/2010/05/10/text-classification-sentiment-analysis-naive-bayes-classifier/

[35] Non-english POS tagger: http://moliware.com/non-english_pos-tagger_with_nltk.html

[36] Marie-Claire Jenkins: How Sentiment Analysis works in machines

[37] Sentiment analysis of movie reviews using Support Vector Machines

[38] Wikipedia: http://www.wikipedia.org

[39] Python: https://www.python.org/

[40] Django: https://www.djangoproject.com/

[41] LXML: http://lxml.de/

[42] PostgreSQL: http://www.postgresql.org

[43] Psycopg2: http://initd.org/psycopg/

[44] Textblob: http://textblob.readthedocs.org/

[45] NLTK: http://www.nltk.org/

[46] Mercurial: http://mercurial.selenic.com/

[47] Apache: http://httpd.apache.org/

[48] StackOverflow: http://stackoverflow.com/

[49] Google Charts: ttps://developers.google.com/chart

[50] CSS: http://www.w3schools.com/Css/

[51] Google Maps: https://developers.google.com/maps

[52] jQuery: http://jquery.com

[53] Bootstrap: http://getbootstrap.com/

[54] Teambox v3: https://github.com/teambox/teambox

[55] Chrome Developer Tools: https://developer.chrome.com/devtools/

[56] Latex Stack Exchange: http://tex.stackexchange.com/

[57] Websays : http://websays.com/