UNIVERSIDAD FRANCISCO DE VITORIA

ESCUELA POLITÉCNICA SUPERIOR

GRADO EN INGENIERÍA INFORMÁTICA

PROYECTO FINAL DE GRADO

ANÁLISIS DE LAS UNIVERSIDADES DE MADRID EN TWITTER UTILIZANDO HERRAMIENTAS DE DATA DISCOVERY

Autor: Álvaro Ullate Sanz

Julio - 2014

UNIVERSIDAD FRANCISCO DE VITORIA

ESCUELA POLITÉCNICA SUPERIOR

GRADO EN INGENIERÍA INFORMÁTICA

PROYECTO FINAL DE GRADO

ANÁLISIS DE LAS UNIVERSIDADES DE MADRID EN TWITTER UTILIZANDO HERRAMIENTAS DE DATA DISCOVERY

AUTOR:

Álvaro Ullate Sanz

TUTOR: Ángel Serrano Sánchez de León

Julio - 2014

VISTO BUENO Y CALIFICACIÓN DEL PROYECTO FINAL DE GRADO

Título: ANÁLISIS DE LAS UNIVERSIDADES DE MADRID EN TWITTER UTILIZANDO HERRAMIENTAS DE DATA DISCOVERY

Autor: ÁLVARO ULLATE SANZ

Tutor: ÁNGEL SERRANO SÁNCHEZ DE LEÓN

VISTO BUENO

VºBº Tutor del PFG:
Fdo.:

Lugar y fecha: _____

CALIFICACIÓN

CUALITATIVA:	
NUMÉRICA:	

Conforme Presidente:	Conforme Secretario:	Conforme Vocal:
Fdo.:	Fdo.:	Fdo.:

Lugar y fecha:_____

Resumen y lista de palabras

Debido a la gran cantidad de datos que generan actualmente las empresas, cada vez es más difícil gestionarlos. Muchas veces no se trata de los datos propios o internos, sino de datos externos, datos de otras fuentes, como de redes sociales. Parte de estos datos proporcionan información útil y de valor para el negocio, proporcionan una ventaja competitiva y nuevas maneras de comprender el negocio y optimizarlo.

La información que se puede obtener de las redes sociales es clave, ya que todas las empresas quieren saber qué se opina de ellas y qué se opina de la competencia. Estos datos pueden ser útiles para lanzar un nuevo producto al mercado, para mejorar los existentes, para mejorar la imagen de marca o para conocer la satisfacción de los clientes.

Para dar respuestas a estas preguntas, se necesitan nuevas herramientas que permitan integrar datos de diferentes sitios, en diferentes formatos y que puedan ser accesibles en el momento.

A lo largo de este proyecto se analizarán las universidades de la Comunidad de Madrid en la red social de Twitter. Se integrarán datos estructurados de usuarios, datos no estructurados sobre comentarios, se enriquecerá el texto obteniendo entidades y se realizará un análisis de sentimiento mediante procesos ETL.

Después del desarrollo se podrán conocer los usuarios que más hablan de las universidades, los que mejor o peor hablan, qué se dice, desde dónde se dice y cuáles son los más influyentes.

Se adquirirán más de 260 mil tuis de las 16 universidades de la Comunidad de Madrid analizadas. De ese conjunto de datos, la Universidad Autónoma de Madrid, la Universidad Complutense de Madrid y la Universidad Nacional de Educación a Distancia serán las universidades que más tuits tengan. La Universidad Francisco de Vitoria se situará en la posición 12ª en cuanto al número de tuits.

Después de realizar el análisis de sentimiento, las universidades mejor valoradas, o que de media tendrán más comentarios positivos, serán la Universidad Antonio de Nebrija seguida de la Universidad Francisco de Vitoria.

Al normalizar los resultados en función del número de alumnos las universidades mejor valoradas o que de media tendrán más comentarios positivos serán la Universidad Internacional Menéndez Pelayo, seguida de la Universidad Camilo José Cela, la Universidad Antonio de Nebrija y la Universidad Francisco de Vitoria.

Analizando los usuarios y focalizándose en la Universidad Francisco de Vitoria, los 3 usuarios que más influencia hacen hablando de la universidad son “Francisco de Vitoria”, “Jane del Tronco” y “Comunicación UFV”. Los 3 usuarios que mejor hablan de la universidad son “Jane del Tronco”, “Nacho Gamma” y “Francisco de Vitoria”. Los 3 usuarios que más comentan sobre la universidad son “Francisco de Vitoria”, “UFV Business” y “Mirada21.es”.

Palabras clave: Big Data, Herramientas de Data Discovery, ETL, Data Mining, Twitter, Universidades de la Comunidad de Madrid, Análisis de sentimiento.

Índice

Índice de figuras

Índice de tablas

Listado de siglas, abreviaturas y acrónimos con sus significados

ADF: Application Development Framework

API o APIs: Application Programming Interface

App: Aplicación

BI: Business Intelligence

CES: Centro de Estudios Superiores

CEU: Universidad CEU San Pablo

CIO: Chief Information Officer

CMS: Content Management System

CPU: Central Processing Unit

CRF: Conditional Random Field

CUNEF: Centro Universitario de Estudios Financieros

DD: Data Domain

EC2: Elastic Cloud Computing

EQL: Endeca Query Language

ESIC: Escuela Superior de Gestión Comercial y Marketing

ESNE: Escuela Universitaria de Diseño e Innovación

ETL: Extract, Transform and Load

GB: Gigabyte

GPL: General Public License

HTTP: Hypertext Transfer Protocol

IAS: Integrator Acquisition System

ICADE: Instituto Católico de Administración y Dirección de Empresas

ICAI: Instituto Católico de Artes e Industrias

IEB: Instituto de Estudios Bursátiles

IEDE: Institute for Executive Development

ISO: International Organization for Standardization

JSON: JavaScript Object Notation

kB: kilobyte

KPI o KPIs: Key Performance Indicator

LDAP: Lightweight Directory Access Protocol

LOPD: Ley Orgánica de Protección de Datos
NLP: Programación de Lenguaje Natural
noSQL: not only SQL
NS/NC: No sabe o no contesta
OEID: Oracle Endeca Information Discovery
OLAP: On-Line Analytical Processing
PB: Petabyte
PHP: Hypertext Pre-processor
RAM: Random Access Memory
RDS: Relational Database Service
SQL: Structured Query Language
SSH: Secure Shell
SSL: Secure Sockets Layer
TASS: Taller de Análisis de Sentimiento de la SEPLN
TI: Tecnologías de la información
TNF: True Negative Fraction
TPF: True Positive Fraction
UAH: Universidad de Alcalá de Henares
UAM: Universidad Autónoma de Madrid
UAX: Universidad Alfonso X el Sabio
UC3M: Universidad Carlos III
UCJC: Universidad Camilo José Cela
UCM: Universidad Complutense de Madrid
UDIMA: Universidad a Distancia de Madrid
UEM: Universidad Europea de Madrid
UFV: Universidad Francisco de Vitoria
UIMP: Universidad Internacional Menéndez Pelayo
UNED: Universidad Nacional de Educación a Distancia
UPCOMILLAS: Universidad Pontificia de Comillas
UPM: Universidad Politécnica de Madrid
URJC: Universidad Rey Juan Carlos
URL: Uniform Resource Locator
U-TAD: Universidad de Tecnología y Arte Digital

VM: Virtual Machine

WSDL: Web Service Description Language

XLS: Microsoft Excel format

XML: eXtensible Markup Language

XPath: XML Path Language

1 Introducción

La mayoría de las organizaciones tienen implementado un sistema de Business Intelligence que permite tomar decisiones, monitorizar KPIs, comprender y mejorar el negocio. Sin embargo, muchos datos no son de fácil acceso ya que se encuentran en sistemas de gestión de contenidos (CMS), correos electrónicos o sistemas de archivos. Hoy en día los datos también se pueden encontrar en empresas ajenas a la organización, blogs o incluso redes sociales.

Además, los sistemas de Business Intelligence tradicionales permiten responder unas pocas preguntas. Una vez que los modelos de datos están definidos y los informes o cuadros de mando creados requieren de un conocimiento y desarrollo específico para modificarlos. Cualquier pregunta adicional requiere cambios en el modelo de datos.

Existen un tipo de herramientas especiales diseñadas para el descubrimiento de información, no requieren de un modelo tan complejo como un sistema tradicional y permiten agregar información de diversas fuentes y diferentes formatos.

A lo largo de este proyecto se desarrollará un sistema de descubrimiento de información con el cual se pretenderá descubrir y detectar respuestas a nuevas preguntas. Para ello el proyecto se centrará en analizar las universidades de la Comunidad de Madrid, tanto privadas como públicas y sus centros asociados.

Mediante este análisis se podrá saber cuáles son las universidades mejores valoradas de la Comunidad de Madrid, se podrá saber qué se dice, quién lo dice y desde dónde lo dice. Además, mediante el análisis de sentimiento, se podrá puntuar automáticamente los comentarios indicando cuáles son positivos o negativos. De esta manera se podrá tener una visión global de cada universidad sin necesidad de leerse los comentarios.

Los datos necesarios se adquirirán de Twitter mediante sus APIs (1), además se diseñará una plataforma de adquisición de datos en la nube mediante los servicios web gratuitos de Amazon (2).

Como herramienta de descubrimiento de información se ha elegido Oracle Endeca Information Discovery (3), ya que Oracle posee todas las capas necesarias para este desarrollo, desde el sistema operativo, base de datos, servidor de aplicaciones, etc. Será necesario realizar diversos procesos ETL para la extracción, transformación y carga, además de un enriquecimiento de texto extrayendo entidades y un análisis de sentimiento valorando los tuits.

Para llevar a cabo este desarrollo se necesitará utilizar una metodología adaptada a las necesidades. Se realizará un estudio previo de algunas metodologías y después de la valoración se optará por un desarrollo en espiral. Uno de los fundamentos de esta metodología es el análisis de riesgos, ya que con los resultados de este análisis se deciden los objetivos de la siguiente fase.

La estructura de esta memoria está divida en los siguientes capítulos.

Capítulo 2: Problema a resolver y objetivos del Proyecto Final de Grado

En este capítulo se describe el problema a resolver, la motivación y el origen, los datos que lo sustentan así como el impacto de la solución del problema. Además se describirán los objetivos y las preguntas de investigación.

Capítulo 3: Antecedentes y estado del arte

En este capítulo se explica la historia de la información y cómo hoy en día nos enfrentamos a un mundo en el que los datos cada vez adquieren más valor. Se posicionarán las herramientas de descubrimiento de información frente a conceptos como Data Science o herramientas especiales para la minería de datos.

Capítulo 4: Limitaciones y condicionantes

En este capítulo se comentarán las limitaciones de las herramientas que se van a utilizar, así como las APIs (1) de Twitter utilizadas, equipos de desarrollo y pruebas, arquitecturas para cada uno de los sistemas y la planificación llevada a cabo.

Capítulo 5: Metodología

Este capítulo comienza explicando algunas metodologías, para más tarde definir unos criterios de elección y poder elegir la más indicada para el proyecto. La metodología usada será el desarrollo en espiral y para ello es necesario definir cómo se van a analizar y supervisar los riesgos de cada fase.

Capítulo 6: Aplicación a la metodología y resultados obtenidos

Este capítulo contiene todo el desarrollo llevado a cabo, desde la adquisición de datos mediante los servicios web de Amazon (2), desarrollo de procesos ETL para el uso de la Rest API (4) y carga al servidor, enriquecimiento de texto y mejora, desarrollo de procesos ETL para la transformación de los datos, integración de todos los componentes necesarios y desarrollos de cuadros de mando.

Capítulo 7: Conclusiones y propuestas

Se exponen conclusiones del trabajo y posibles mejoras.

Capítulo 8: Referencias

Contiene todas las referencias utilizadas y fuentes bibliográficas para el desarrollo del proyecto en formato ISO-690.

Anexos

Se adjuntan todos los anexos necesarios para la comprensión del proyecto.

2 Problema a resolver y objetivos del Proyecto Final de Grado

Descripción del problema a resolver

Las herramientas de descubrimiento de información, combinan las tecnologías involucradas en Big Data para un rápido e intuitivo análisis de la información. Estas plataformas permiten extraer la información de datos no estructurados como redes sociales, webs, sistemas de información, bases de datos y combinarlas con datos no estructurados tradicionales.

Esto proporciona una potencia extra, que tiene resultados inmediatos en ahorro de coste y tiempo, permitiendo mejorar las decisiones de negocio basándose en una mejor comprensión del negocio.

El problema que se quiere resolver mediante estas herramientas es conocer la imagen que las universidades de la Comunidad de Madrid tienen en la red social Twitter. Para ello se analizarán los comentarios y los usuarios que han publicado tuits durante un período de tiempo. Después de este análisis se conocerán los temas sobre los que se hablan, desde dónde se habla y en general qué universidad está mejor valorada por los usuarios que publican en Twitter.

Motivación y origen

Hace 30 años ninguna empresa tenía un sistema de Business Intelligence que permitiese dar una visión global de la empresa. Hoy en día, no hay empresa competitiva que pueda tomar decisiones sin estos sistemas. Sin duda, el uso de las tecnologías que están involucradas en Big Data, será un elemento diferenciador en la próxima década.

A las empresas les interesa saber qué opinan sus clientes, qué opinan los clientes de la competencia y cuáles son los más influyentes. Este análisis permitirá conocer la opinión de los clientes (los alumnos) y estudiar la competencia (las universidades).

Datos que lo sustentan

Existen numerosos estudios que confirman que el aumento de datos será un problema en el futuro, pero podemos recalcar el estudio “Big data: The next frontier for innovation, competition and productivity” de McKinsey & Company en Junio de 2011 (5). Por poner alguna cifra, el estudio concluye que a nivel mundial, se espera un 40% de crecimiento en el volumen de datos generados por año frente a un 5% de crecimiento en el gasto mundial en IT.

Una encuesta a finales de 2013 realizada por Gartner (6) revela que el 30% de las organizaciones ya han invertido en tecnologías Big Data, que el 19% lo hará en 2014, el 15% lo hará en 2016 y tan solo el 8% tiene en realidad un sistema desplegado y funcionando.

Esto quiere decir que los próximos años, las decisiones que se tomen tendrán que tener en cuenta todos estos datos generados por las empresas, además estos datos muchas veces serán la clave de una mejor comprensión del negocio así como la respuesta a su optimización.

Impacto de la solución del problema

Mediante esta solución se pretende por una parte, dar respuesta de manera sencilla a preguntas clásicas sobre la opinión las universidades, sobre los usuarios más activos y sobre la competencia.

El impacto de la solución es muy amplio, desde una mejora en la comunicación, una mejora en la segmentación del mercado, mejora y estudio de los resultados de las estrategias de marketing, detección de personas de influencia e incluso el estudio de la competencia.

Problema de investigación en forma de pregunta

¿Qué información se puede extraer de cientos de miles de comentarios sin tener que leerlos? Y ¿se puede medir la repercusión de las universidades de Madrid en Twitter?

Preguntas de investigación

- ¿Cuáles son las universidades de la Comunidad de Madrid mejor valoradas?
- ¿Quién está hablando de las universidades?
- ¿Desde dónde se habla de las universidades?
- ¿Cuáles son las personas o empresas que mejor o peor hablan de las universidades?
- ¿Sobre qué temas se está hablando cuando se habla de las universidades?
- ¿Qué personas o empresas son las más influentes?
- ¿Cómo combinar datos estructurados y no estructurados?
- ¿Cómo crear un cuadro de mandos?
- ¿Cómo utilizar herramientas ETL?
- ¿Cómo utilizar sistemas de análisis de sentimiento?
- ¿Cómo aprovisionar un sistema de descubrimiento de información?
- ¿Cómo encontrar KPIs adecuados?

Objetivos generales y específicos

- Evaluar la influencia de las universidades de la Comunidad de Madrid en Twitter.
- Detectar personas o empresas de influencia en Twitter.
- Evaluar el sentimiento general de los usuarios de Twitter cuando hablan de las universidades de la Comunidad de Madrid.
- Combinar datos estructurados y no estructurados en una misma base de datos.
- Interactuar con información y visualizarla para obtener valor.
- Evaluar nuevas situaciones de negocio.
- Crear cuadros de mando para visualizar esta información.
- Uso de herramientas ETL para el tratamiento de la información.
- Aprovisionar un sistema de descubrimiento de información.
- Uso de sistemas de análisis de sentimiento.

3 Antecedentes, estado del arte

3.1 Big Data, la electricidad del siglo XXI

Hoy en día la típica conversación sobre Big Data puede ir en muchas direcciones diferentes. Los CIOs con Big Data son un poco como los economistas con la inflación. Si hay cuatro personas en una reunión hay seis opiniones diferentes. Así que Big Data todavía es un poco misterioso. Para aclararlo un poco tenemos que mirar a otro fenómeno que fue igual de misterioso, la electricidad.

En la década de 1700 la electricidad fue descrita como un fluido misterioso. Hubo algunas ideas de cómo moverlo de un lado a otro, pero nadie sabía de qué se componía, era un poco confuso. Y así en 1752, en un famoso experimento, Benjamin Franklin fue a un campo de las afueras de Filadelfia y en medio de una noche húmeda voló una cometa. La cometa tenía la estructura de metal, un hilo de seda y justo en el extremo una llave metálica. Benjamin pudo capturar parte de un rayo en una primitiva batería, una botella de Leyden, para más tarde llevarlo a su laboratorio. Con este experimento demostró que la electricidad del cielo era de un material que podía generar en su laboratorio.

Benjamin no fue impulsado por pura curiosidad, él estaba tratando de resolver un problema. Básicamente estaba tratando de detener la caída de rayos que causaban que los edificios se quemasen. Poco más tarde surgió el pararrayos.

Faraday, Tesla y Edison fueron algunos de los seguidores de los experimentos de Franklin. Convirtieron ese fluido misterioso en una tecnología que trajo un enorme valor para la sociedad y los negocios, pero sobretodo supuso una revolución en la vida de las personas. La electrificación abrió las puertas a la innovación, trayendo nuevos productos, nuevos servicios, nuevas formas de trabajo y nuevas formas de vivir. Era un poder que cambiaba todo lo que tocaba.

Ahora mismo estamos en un momento crucial, con grandes volúmenes de datos. Cuando la datificación de todo es un nuevo tipo de poder que cambia todo lo que toca.

La datificación es la captura y el uso de todo tipo de datos en la vida cotidiana, y en muchos de ellos ya formamos parte. Por ejemplo, los usuarios de Google realizan 2 millones de búsquedas cada minuto (7) y se comparten 350 GB de información en Facebook (7). Antes nuestras opiniones, nuestros pensamientos, los restaurantes favoritos o qué tipo de galletas preferimos desaparecían en un océano de datos. Hoy en día, se utiliza para segmentar campañas de marketing o para producir el diseño de la próxima generación de un producto. Es más, ese aprovechamiento de los datos, hoy en día es lo que diferencia una empresa de éxito frente a una empresa.

Un Boing 747 genera en torno a 2 millones de piezas de información por cada 3 horas de vuelo (8). Pero no sólo son productos técnicos, por ejemplo, en algunas granjas se etiqueta a cada vaca cuando nace, creando así un registro de toda la vida de las cabezas de ganado (9).

Esto es la dataficación de las cosas y estamos justo en el comienzo de lo que podemos hacer con los datos.

Se estima que en 2009 había casi mil millones de dispositivos inteligentes en uso (10). Por dispositivo inteligente se entiende cualquier dispositivo que integra sensores y cierta inteligencia, por ejemplo un smartphone o una lavadora moderna. Para el 2020 se prevé que ese número crezca casi 26 veces, es decir, 26 mil millones de dispositivos inteligentes (10). Sin duda, hay una enorme oportunidad en este sector de los datos, entonces, ¿cuál es el problema? El problema es que la mayor parte de las organizaciones no tienen capacidad para procesar estos datos.

Con la dataficación vienen datos de diferentes tipos, en diferentes formatos y tenemos que saber cómo guardarlos, gestionarlos y sacarles provecho. Es justo ahí donde entran tecnologías como Hadoop, noSQL y las herramientas de descubrimiento de información. Estas tecnologías no precisan tanto de un modelo de datos como las bases de datos relacionales, en las que primero se construye un modelo y luego se llena con los datos. Hoy en día se evoluciona a primero adquirir los datos y luego ver qué información se puede extraer. En estos nuevos modelos no se está muy seguro de las preguntas que vas a resolver o qué partes son más valiosas, incluso no se está seguro de qué combinaciones o transformaciones hay que hacer para que los datos sean más valiosos.

En el caso de la electricidad, hubo una amarga guerra de estándares, dos alternativas: corriente alterna y corriente continua. Tomas Edison fue el líder de la corriente continua, mientras que Tesla y Westinghouse eran los campeones de la corriente alterna. La corriente continua es de baja tensión, así que si la tocas probablemente no pase nada. Por el contrario la corriente alterna, con voltajes más altos, es más peligrosa pero se puede transmitir a larga distancia y de manera más eficiente de una planta de producción a otra.

Esta guerra de estándares no terminó con un bando ganador, sino por una combinación de ambas opciones, adaptando cada una a una situación. Hoy en día generamos y transmitimos energía usando corriente alterna, mientras que la corriente continua es el corazón de las baterías en todos dispositivos móviles.

Lo mismo está sucediendo aquí con el enfoque para obtener el valor de los datos. Por una parte necesitamos un entorno relacional que garantice una velocidad de acceso a los datos estratégicos y por otro un entorno analítico no relacional, perfectamente desplegado para que pueda adquirir datos de diferentes fuentes, tratarlos y enriquecer esos datos relacionales permitiendo así una mejor toma de decisiones.

Esto nos lleva a una analítica, en la que los informes y cuadros de mando se complementan con herramientas de descubrimiento de información, que ayudan a comprender esta gran variedad de datos antes de que sean puestos en un modelo más “limpio”. Este modelo se tendrá que complementar con tecnologías como el aprendizaje automático y el modelado predictivo, ya que se tendrán datos moviéndose rápidamente, en diferentes formatos y de diferentes fuentes.

En resumen, ¿qué es Big Data? Las primeras definiciones de Big Data dicen que este concepto se puede resumir en 3 uves, volumen, variedad y velocidad. Más tarde empezaron a añadir uves y hay estudios que hablan de diez (11) o más uves. Pero hay algo que todos comparten:

- **Volumen**: debido a la gran cantidad de datos y al aumento de los mismos, Big Data tiene que trabajar con grandes volúmenes de información.
- **Variedad**: no solo se trata de datos estructurados, sino datos no estructurados, imágenes, vídeos, audio u otros tipos de datos. Todos estos datos en diferentes formatos y con diferentes estructuras.
- **Velocidad**: el ritmo al que los datos se generan cada vez es más rápido muchas veces incluso más rápido que nuestra capacidad para guardarlos. Big Data tiene que anticiparse y resolver este problema.

Otras definiciones también incluyen una cuarta "v", valor, ya que el objetivo final de Big Data es extraer valor de los datos, ya sea para mejorar la toma de decisiones o para desarrollar nuevas generaciones de productos o servicios.

Volviendo al punto de partida, Chicago (12) (13), 1893, las calles estaban iluminadas con lámparas de gas parpadeante, los hogares utilizaban lámparas de queroseno y ese mismo año, en la exposición universal de Chicago, mediante más de 200.000 bombillas incandescentes (14) se iluminó el edificio de la Administración de Chicago. Nadie había visto nunca nada igual, ya que a las afueras de la exposición estaba totalmente oscuro.

Ahora estamos en ese momento, en esa exposición en Chicago, en el momento en el que sabemos que hay que utilizarlos pero no sabemos cómo, existe alguna luz que brilla en una inmensa oscuridad y estamos a la espera de nuevos productos, nuevos servicios y nuevas formas de sacar valor a esos datos que sin duda revolucionarán nuestra manera de vivir.

3.2 Big Data en números

A medida que más datos están disponibles a partir de una gran cantidad de fuentes, tanto internas como externas, las organizaciones están tratando de utilizar esos recursos para mejorar la innovación, retener a los clientes o aumentar la eficiencia. Al mismo tiempo, las organizaciones se enfrentan al reto de orientarse a los usuarios finales, que exigen mayor capacidad de decisión haciendo que se vean obligadas a integrar su información y analizar nuevas fuentes de información. Está claro que Big Data ofrece nuevas oportunidades para los usuarios de negocio al poder responder a nuevas preguntas.

A finales de 2012, Unisphere Research, una división de Information Today, Inc., realizó un estudio en el que participaron 298 administradores de datos y profesionales de diferentes organizaciones y diferentes sectores, de los cuales el 36% poseían algún título de administración de bases de datos (15).

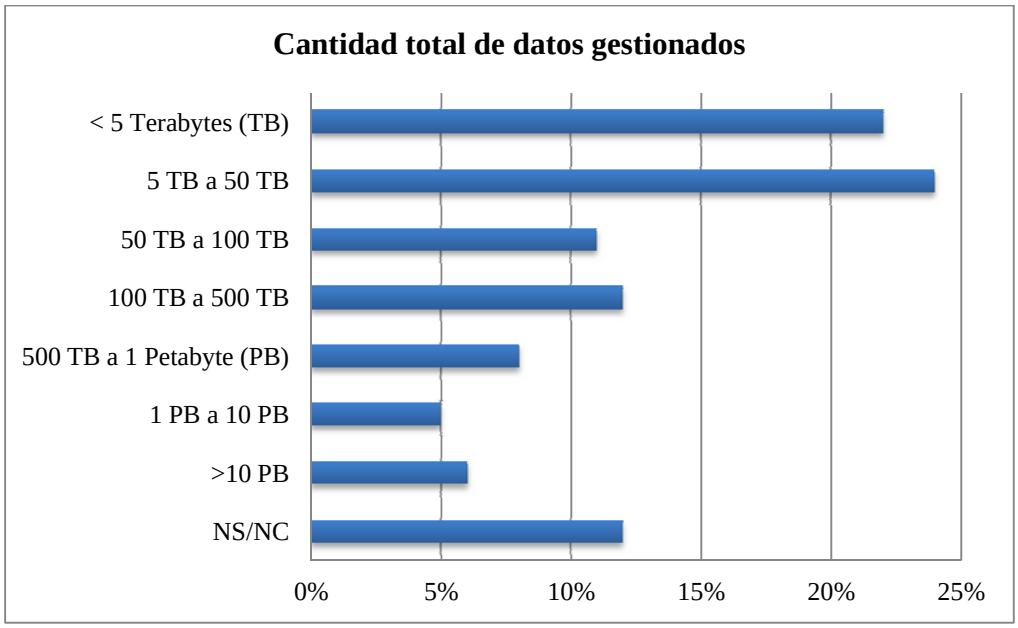


Figura 1: Cantidad de datos gestionados. Fuente: adaptación del informe de Unisphere Research (15)

En términos de volumen, el estudio revela que a finales de 2012 el 55% gestionaban más de 50 TB de información y más de un 10% ya gestionaba más de 1 PB.

Es importante desglosar esta información en función del tamaño de la empresa, como se muestra en la tabla 1. Concretamente el 28% de las empresas con más de 10.000 empleados gestionan más de 1 PB.

Tabla 1. Relación entre el número de empleados en una empresa y la cantidad de datos. Fuente: adaptación del informe de Unisphere Research (15)

	< 1.000 empleados	de 1.000 a 10.000 empleados	> 10.000 empleados
<5 Terabytes (TB)	37%	11%	12%
5 TB a 100 TB	41%	45%	24%
100 TB a 500 TB	10%	18%	9%
500 TB a 1 Petabyte (PB)	2%	7%	14%
< 1 PB	1%	8%	28%

Si investigamos qué porcentaje de esos datos son estructurados y no estructurados, vemos en la figura 2, que de momento solo el 7% tienen más del 50% de datos no estructurados.

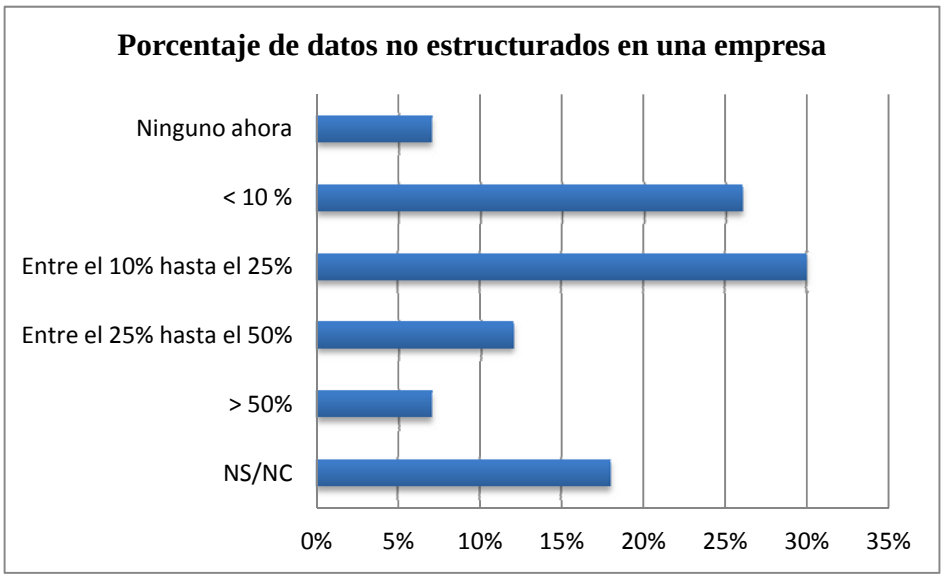


Figura 2. Cantidad de datos no estructurados en una empresa. Fuente: adaptación del informe de Unisphere Research (15)

Cuando se implanta un sistema Big Data, tradicionalmente Hadoop, se necesita una gran cantidad de información, normalmente no estructurada. Como se ha visto en la figura 2, el 20% de las empresas no pueden gestionar más del 25% de la información que poseen sin un sistema Big Data. Esto quiere decir, que el 20% de las empresas toma decisiones sobre su negocio sin tener en cuenta más del 25% de la información. ¿Y si esa información que actualmente no se está utilizando se integrara dentro de los sistemas de toma de decisión empresarial?

Es por ello que solo el 13% de las empresas tienen un sistema planificado o en producción, para aprovechar esa cantidad de información no utilizada actualmente. Además en la figura 3, se observa que más del 20% está considerando implantar un sistema Hadoop.

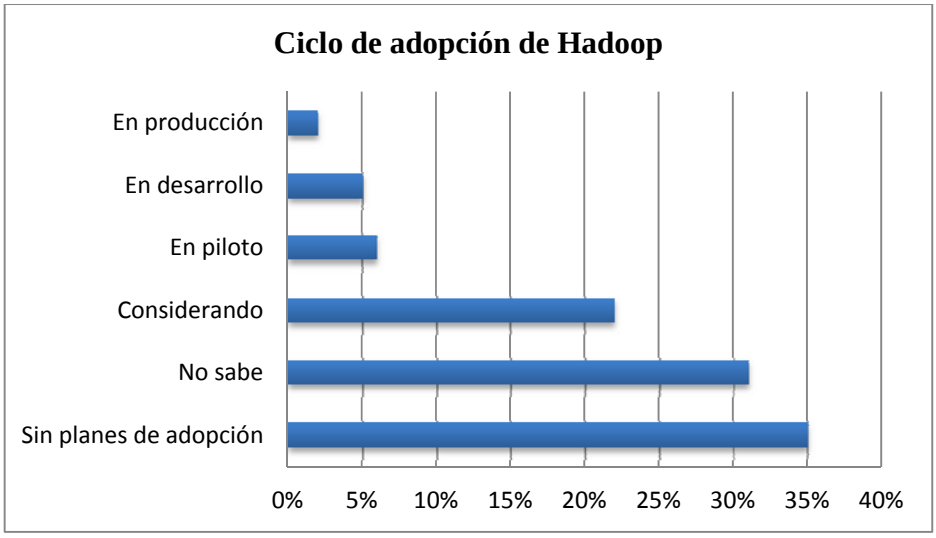


Figura 3. Ciclo de adopción de Hadoop. Fuente: adaptación del informe de Unisphere Research (15)

Otro problema al que nos enfrentamos en un sistema Big Data es la seguridad. Se está exigiendo a las organizaciones aplicar las restricciones de control de acceso y privacidad de estos conjuntos de datos para cumplir con los requisitos normativos como las leyes de proyección de datos LOPD. Los ataques en las redes están en aumento y a menudo se tardan meses en ser detectados, conllevando esto un gasto importante para las empresas.

Según el estudio de Unisphere Research (15), figura 4, más del 55% de las empresas consideran que más del 99% de los datos que poseen son valiosos.

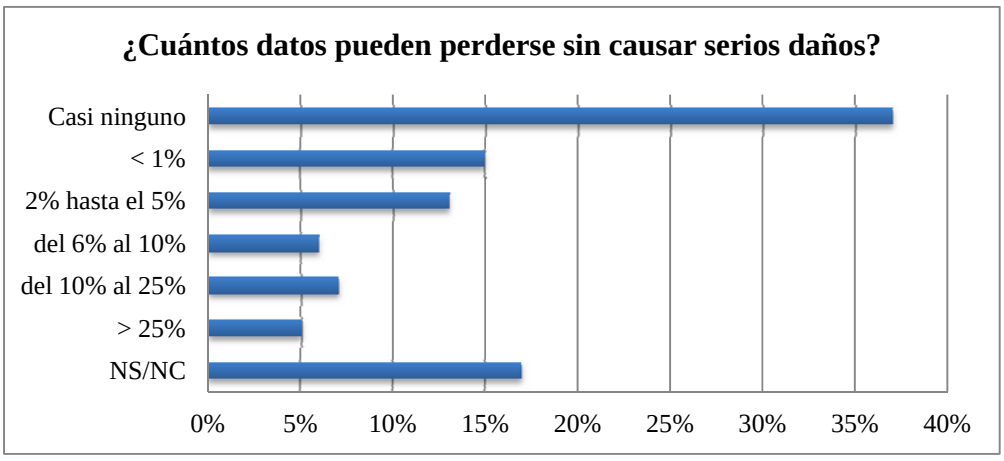


Figura 4. Importancia de los datos en las empresas. Fuente: adaptación del informe de Unisphere Research (15)

3.3 Herramientas de descubrimiento de información

¿Qué pasaría si no tuviese que construir un modelo de datos? ¿Y si pudiera poner un índice a todos esos datos para poder explorar la información? Es lo que hacen las herramientas de descubrimiento de información. Crean un índice de todos los datos diversos y permiten visualizarlos agregándolos o llegando directamente al dato más desagregado posible, es decir, a la propia pieza de información.

Hay dos maneras de utilizar herramientas de descubrimiento de información en el contexto de grandes volúmenes de datos. En primer lugar, el uso de estas herramientas para comprender los datos y su valor, y utilizar los conocimientos adquiridos para diseñar un proyecto de Big Data. Esto permite tomar decisiones sobre qué fuentes se van a utilizar y mediante qué técnicas serán tratados los datos, por ejemplo, enriquecimiento de texto. Además permite mejorar la definición de una propuesta de valor para justificar la inversión. En segundo lugar, es el propio descubrimiento de información un proyecto en sí mismo, permitiendo a los usuarios finales combinar y explorar una gran variedad de datos de forma rápida y sencilla. Si bien estas dos aplicaciones son muy valiosas de diferentes maneras para diferentes partes de cada organización, las capacidades clave subyacentes son compartidas.

Para poder comprender la utilidad lo mejor es poner un ejemplo real. Imaginemos una empresa de bienes de consumo que de repente ve un aumento en las ventas de un producto lácteo, quesos. La primera pregunta que se plantea es ¿por qué ha pasado? Después de investigar los informes de los precios, las promociones y la publicidad, puede ocurrir que ninguno haya tenido cambios recientes. Incluso investigando los precios de la competencia, puede ser que no se encuentre nada inusual. Entonces, ¿qué ha pasado?

Llega el momento de comenzar una lluvia de ideas sobre dónde más buscar, y a alguien se le puede ocurrir la idea de combinarlos con datos de redes sociales. Resulta que lo que no se estaba estudiando son los comentarios de los usuarios y analizándolos se pueden encontrar relaciones. Por ejemplo, puede ser que la campaña de marketing haya tenido un impacto en los clientes porque resulta que es una historia con un final feliz y, claro, la historia rápidamente se difunde en las redes sociales. Esto puede generar un aumento en el número de personas que conocen el producto y por consiguiente un aumento en las ventas de estos quesos.

De esta manera, y solo agregando los datos de sus campañas con los datos de las redes sociales y blogs, se ha detectado un nuevo nicho de mercado. Habrá que generar nuevas métricas para integrarlos en los cuadros de mando tradicionales. Además se puede saber las personas que más influyen en esa campaña e influenciarlas para que sigan comentando positivamente.

Esta información no se puede obtener de informes Business Intelligence, no porque estén equivocados o incompletos, sino porque se basan en comprender el negocio antes de estos eventos. En el caso anterior, se fue directamente a los datos, en particular, a los datos de los medios sociales, y sin saber la pregunta se obtuvo la respuesta.

Pero, ¿para qué sirven todos estos datos?

En primer lugar, como siempre, las empresas quieren mejorar sus procesos internos. Ya sea mejorar campañas de marketing, dirigir la inversión a otros proyectos o mejorar la toma de decisiones.

En segundo lugar, las empresas quieren responder más rápidamente cuando ocurre algo inesperado. Por ejemplo, en un call center, se necesitan tener los datos procesados de diferentes encuestas, medios sociales y datos internos antes que llame el cliente.

En tercer lugar, las empresas ven todas esas nuevas combinaciones de datos como una oportunidad para innovar o incluso para cambiar su modelo de negocio. Es el caso de las empresas de telecomunicaciones que ante la imposibilidad de competir con operadoras virtuales tendrán que cambiar su modelo de negocio a proveedores de datos (16).

3.4 Data Science versus herramientas de descubrimiento de información

A menudo la gente piensa que data discovery o las herramientas de descubrimiento de información son otro término en lugar de data science o ciencia de los datos. Aunque las dos terminologías muchas veces se superponen, existen algunas diferencias.

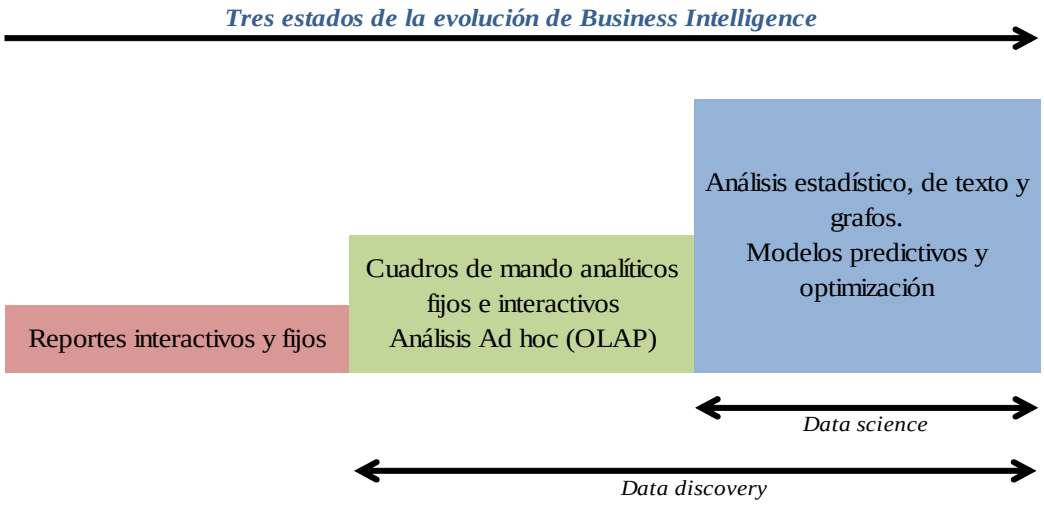


Figura 5. Evolución del Business Intelligence. Fuente: (17)

Data science es una evolución del data mining, la estadística y el aprendizaje automático. Además soporta sofisticadas técnicas y tecnologías de análisis, data science tiene un impacto directo en el negocio con la colaboración de TI identificando claramente los requisitos de negocio. El objetivo de la ciencia de datos es utilizar técnicas analíticas avanzadas procedentes de un entorno más experimental en los principales procesos de negocio de una empresa.

Una de las principales barreras en la implementación de proyectos de data science es la falta de personal cualificado. La ciencia de los datos requiere unos conocimientos avanzados en ingeniería y análisis de datos, análisis de negocios y experiencia en el sector, conocimientos en estadística avanzada y dominio de modelos predictivos. Esto es una barrera importante ya que aunque las universidades hacen un esfuerzo en la educación científica orientada a los datos, estos perfiles requieren preparación de muchos años y el sector los demanda hoy.

Los sistemas de descubrimiento de información o data discovery incluyen tecnologías de la ciencia de los datos, por tanto compatible con esos perfiles, pero además proporcionan herramientas que pueden ser utilizados por los analistas de negocios sin tener unos conocimientos avanzados en modelos predictivos o modelos estadísticos. Las herramientas proporcionan capacidades OLAP tradicionales combinadas con análisis estadísticos, análisis de texto o análisis de grafos.

Estas herramientas se combinan perfectamente con la ciencia de los datos ya que permiten al usuario de negocio, que es el que verdaderamente conoce el negocio, buscar patrones de una manera muy intuitiva, reduciendo la carga de trabajo en los científicos de los datos, para más tarde focalizar los esfuerzos en los descubrimientos que hagan los usuarios de negocio. Además mejora la comunicación entre los perfiles propiamente de TI y los de negocio.

3.5 Data Mining versus Herramientas de descubrimiento de información

Aunque no existe una única definición, la minería de datos es una extracción no evidente a partir de los datos de información implícita, previamente desconocida y potencialmente útil.

La minería de datos recoge ideas de diversas disciplinas, desde el aprendizaje estadístico a la inteligencia artificial que otras técnicas más tradicionales no pueden resolver, ya sea porque el conjunto de datos es excesivo, son heterogéneos o tienen varias dimensiones.

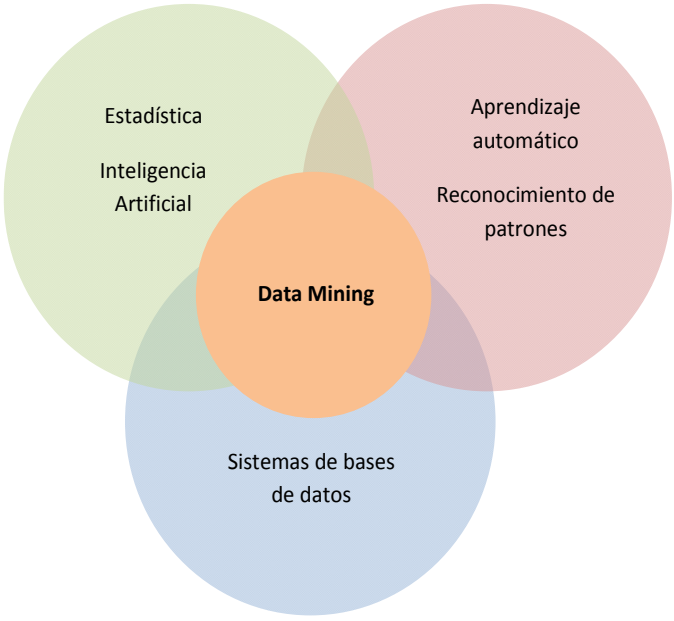


Figura 6. Componentes del Data Mining. Fuente: (17)

Mediante la minería de datos se puede predecir, usando variables cuyos valores son conocidos para determinar esas variables en el futuro, como son la clasificación o el estudio de la desviación. Además se pueden encontrar patrones y relaciones entre los diferentes atributos, mediante técnicas de agrupamiento o reglas de asociación.

La principal diferencia frente a las herramientas de descubrimiento de información son la complejidad y la profundidad del análisis. Una de las tareas más complejas cuando un equipo se enfrenta a un proyecto de minería de datos es el estudio de los datos antes de comenzar su tratamiento, mientras que las herramientas de descubrimiento de información no requieren un modelo, son un paso previo al uso de la minería de datos. Mediante estas herramientas se pueden encontrar relaciones.

Algunas de las principales diferencias son mostradas en la tabla 2.

Tabla 2. Comparativa minería de datos frente a las herramientas de descubrimiento de información. Fuente: elaboración propia.

Característica	Minería de datos	Herramientas de descubrimiento de información
Requiere un estudio previo de los datos	Sí	No
Data Mining y Text Mining en el mismo proceso	No, requieren procesos diferentes	Sí
Tiempo de desarrollo hasta obtener primeros resultados	Alto	Bajo
Profundidad de los resultados	En función del tiempo, se pueden hacer proyectos muy complejos	Proyectos de poca profundidad, principalmente para buscar relaciones
Requiere conocimientos en IA, estadística, aprendizaje automático…	Sí	No

Aunque son tecnologías que se complementan, hay que tener en cuenta que no se sustituyen. Los proyectos de minería de datos son complejos, utilizados para crear nuevas hipótesis que se contrastan contra poblaciones más grandes. Las herramientas de descubrimiento de información obtienen las relaciones entre atributos y pueden ser un paso previo a la minería de datos para entender mejor y ver las relaciones del conjunto de datos.

3.6 Twitter

Con más de 200 millones de usuarios (18) Twitter se posiciona como la mayor plataforma de microblogging del mundo. Esta red conecta a todos los usuarios y permite interactuar con otros enviando pequeños fragmentos de texto, de hasta 140 caracteres, llamados tuits. Esta red genera más de 65 millones de tuits al día y permite agilizar las relaciones entre usuarios. El impacto ha sido tan grande que, no solo conecta usuarios, sino que además permite a las empresas lanzar campañas de marketing y saber qué opinan los usuarios de sus productos o servicios.

Aunque la mayor parte del contenido que se propaga por la red no tiene sentido analizarlo individualmente la agregación de estos pequeños fragmentos de texto ha sido y es estudio por parte de muchas organizaciones en lo referente al impacto en el mercado de una marca. El 40% (19) son comentarios sin sentido, pero el 38% son conversaciones entre personas o entre personas y empresas. Además el 5% son promociones o anuncios y el 9% son conversaciones "retuiteadas", es decir, mensajes repetidos por los usuarios.

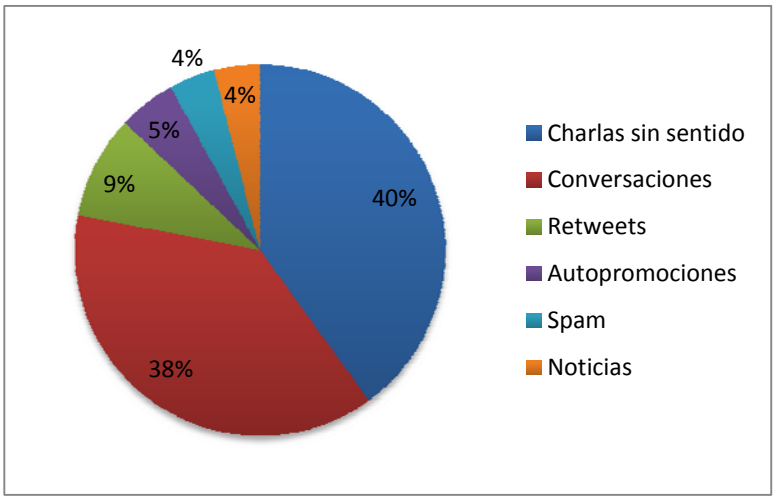


Figura 7. Contenido de los tuits. Fuente (19)

No solo son comentarios, sino que se ha introducido en la cultura de la gente y ante cualquier tipo de evento los usuarios comentan rápidamente a su comunidad. Esto nos proporciona una muestra bastante real y centralizada del movimiento y la opinión de la gente.

El impacto de Twitter es tan grande que la Real Academia Española clasifica tuit y tuitear como neologismos que formarán parte de la 23ª edición del Diccionario, que se publicará en octubre de este año (20).

3.7 Análisis de sentimiento

El análisis de sentimiento, también llamado minería de opinión, es el campo que estudia el análisis de las opiniones, sentimientos, valoraciones, las actitudes de las personas y emociones hacia entidades como productos, servicios, organizaciones, individuos, eventos, temas y sus atributos (21).

Existen numerosos nombres con algunas pequeñas diferencias aunque todos incluidos dentro del análisis de sentimiento, algunos de ellos son: análisis de sentimiento, minería de opinión, extracción de opinión, minería de sentimientos, análisis de emociones, etc.

Aunque el procesamiento del lenguaje natural (NLP) tiene una larga historia, pocas investigaciones han sido anteriores al año 2000 (21). Desde entonces, por diversas razones, este campo se ha convertido en un área de investigación muy activo.

3.7.1 Nivel de granularidad

Dependiendo del nivel de granularidad se ha investigado en tres niveles:

- **A nivel de documento**: consiste en puntuar y clasificar un documento entero expresando si es negativo, positivo o neutral. Por ejemplo, para saber la opinión general de un comentario sobre un producto.
- **A nivel de frase**: descomponiendo el documento, este nivel determina la puntuación de cada frase. Es importante recalcar que este nivel es muy complicado de analizar, ya que depende de la persona que esté escribiendo la opinión y del idioma. Por ejemplo, la frase “mira los niños qué tristes están” (y están jugando) es positiva, pero cualquier analizador diría lo contrario. Un problema añadido del español es que existe la ironía como figura retórica.
- **A nivel de aspecto o entidad**: muchas veces los análisis a nivel de documento o de frase no califican correctamente la opinión que se quiere analizar. Este nivel está basado en la idea de que una opinión está compuesta de un sentimiento y un objetivo. Por ejemplo, la frase “aunque en Madrid hace un calor horroroso, las noches de verano están muy animadas” claramente es positiva, aunque no se puede afirmar que es enteramente positiva. En este caso, la entidad es Madrid y hay dos aspectos, “hace un calor horroroso” y “las noches de verano están muy animadas”.

3.7.2 Clasificación del sentimiento de un documento

Dado un documento de opinión existen diferentes maneras de determinar el sentimiento general del documento. Si la valoración del documento toma valores categóricos, por ejemplo, positivo o negativo, entonces es un problema de clasificación. En cambio, si tomas valores numéricos, por ejemplo, dentro de un rango de -5 a 5, el problema cambia a un análisis de la regresión, es decir, existe una relación entre variables que cambia en función de diferentes parámetros.

3.7.2.1 Clasificación de sentimiento usando aprendizaje supervisado

Tradicionalmente la clasificación de texto principalmente clasifica documentos en función de los temas, por ejemplo, ciencia, deportes o política. Palabras como excelente, sorprendente o

maravilloso, horrible o peor indican claramente un sentimiento positivo o negativo. Para resolver este problema los investigadores (21) han utilizado desarrollos basados en algunas características. Algunas de ellas son:

Términos y su frecuencia

Esta característica indica la frecuencia con la que aparece un término en un documento. En algunos casos, también influye el orden de las palabras.

Categoría gramatical

Las palabras con diferente categoría gramatical pueden ser tratadas de manera diferente. Por ejemplo, se ha demostrado (21) que los adjetivos son indicadores importantes de opiniones.

Palabras y frases con sentimiento

Existen palabras como bueno, amar o maravilloso que son palabras con un sentimiento positivo, de la misma manera palabras como pobre, terrible, odiar o basura son negativas. Estas palabras suelen ser adjetivos, adverbios, nombres o verbos que se usan para expresar sentimientos. Además dependiendo del idioma puede haber modismos, es decir, expresiones fijas cuyo significado no se extrae de sus palabras, por ejemplo, “ahogarse en un vaso de agua”.

Palabras que cambian el sentimiento

Palabras como no o nunca pueden cambiar el significado completo de la frase, por ejemplo, “qué gran coche, el primer día no arrancó”.

3.7.2.2 Clasificación de sentimiento usando aprendizaje no supervisado

Existe una técnica descrita por Peter D. Turney en 2002 (22) que además del sentimiento de las palabras, se basa en buscar patrones fijos predefinidos que son usados para expresar opiniones. Por ejemplo, si aparece un adjetivo y un nombre seguido, será siempre una opinión.

Otra aproximación es el método basado en el léxico, el cual usa un diccionario de palabras con sentimiento e incorpora intensificadores y operadores negativos para calcular la puntuación de cada documento. Este método suele usarse en el análisis a nivel de frase (21).

3.8 Otros estudios utilizando Twitter

En el mercado existen numerosas herramientas gratuitas y de pago para analizar los datos de Twitter. Principalmente se dividen en dos grandes grupos: aquellas que analizan los perfiles de los usuarios (publicaciones, retuits, seguidores, amigos, etc.) y las que están especializadas en análisis de sentimiento (extracción de entidades, calificación del texto no estructurado, obtención de temas y resúmenes, etc.).

Herramientas de análisis de perfiles: estas herramientas se basan en analizar los datos estructurados que proporcionan la API de Twitter (4) y obtener resultados sobre los perfiles de los usuarios.

- **Twitter Analytics**
  Twitter ha creado una herramienta online gratuita para sus usuarios, Twitter Analytics (23), que proporciona información sobre la actividad de los tuits publicados, sus retuits y sus respuestas, además de información sobre los seguidores, como sus intereses y su ubicación.
- **Twtrland**
  Es una herramienta (24) de pago que permite el acceso completo a los perfiles sociales de Twitter, monitorizar el progreso de la cuenta (seguidores, amigos, tuits, retuits), análisis de la competencia, monitorización de palabras clave e incluso detectar cuales son las personas más influentes respecto a un tema.
- **Otras herramientas para analizar el perfil de usuario**
  Es el caso de Twitonomy (25), Topsy (26), Simply Measured (26) y Tweetlevel (27) para monitorizar el perfil de usuario, ver el incremento de los seguidores o analizar la influencia de los tuits publicados.

Herramientas de análisis de sentimiento: además hay otro tipo de herramientas orientadas a evaluar el sentimiento general del texto no estructurado (28). Estas aplicaciones ofrecen servicios webs con los que se pueden realizar extracciones de entidades, identificación de idioma, análisis de sentimiento o extracción de temas. Debido a su complejidad suelen ser empresas dedicadas exclusivamente al análisis de sentimiento que se integran en sistemas más complejos.

- **Textalytics** (29) es una herramienta de la empresa Daedalus (30) para el análisis de sentimiento, optimizada en texto no estructurado de medios sociales y puede analizar tuits, comentarios, noticias, etc. Es multilingüe, puede trabajar en la nube y reconoce entidades como personas, marcas o productos, además de clasificar temáticamente los resultados y realizar una valoración del comentario entre otras funcionalidades.
- **Salience Engine** (31) es la herramienta de Lexalytics (32) para analizar texto en varios idiomas. Realiza extracción de entidades, puntuación del comentario, genera resúmenes del comentario e identifica temas en el texto.

El problema de todas estas herramientas es que no incluyen la posibilidad de cruzar los análisis realizados con datos propios del usuario o empresa, por ejemplo, datos o preferencias de los clientes obtenidos mediante encuestas, campañas de marketing realizadas o datos de la competencia.

En este proyecto se desarrollará un sistema basado en obtener relaciones entre los datos. Se utilizarán varios componentes, uno de ellos será el análisis de sentimiento para analizar el impacto de las universidades de la Comunidad de Madrid en la red social de Twitter.

4 Limitaciones y condicionantes

4.1 Oracle Endeca Information Discovery (OEID)

Para el desarrollo del proyecto se ha optado por utilizar una de las herramientas de descubrimiento de información, Oracle Endeca Information Discovery (3). Esta herramienta se compone de diferentes módulos, algunos de ellos sobre un servidor de aplicaciones. En este proyecto se ha optado por Oracle WebLogic 12c (33) ya que proporciona la mejor compatibilidad.

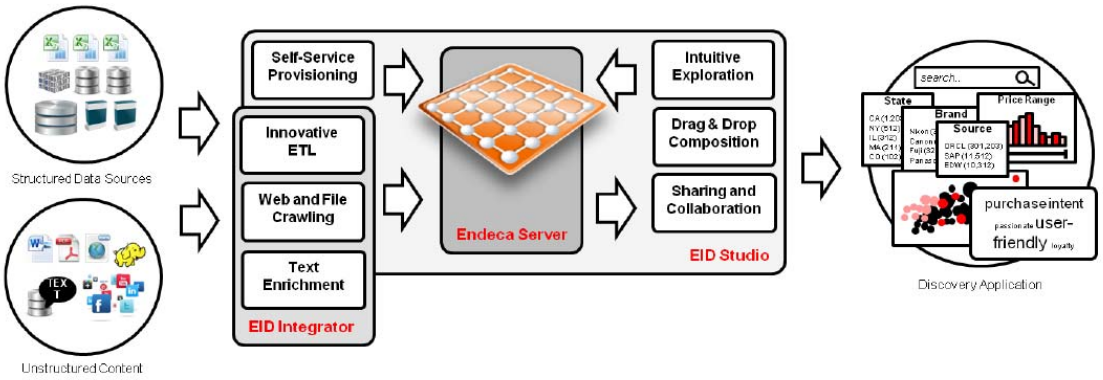


Figura 8. Esquema de los componentes principales de Oracle Endeca. Fuente (34)

Oracle Endeca Server (35)

El núcleo del servidor de Endeca es una base de datos híbrida analítica la cual proporciona una flexibilidad adicional permitiendo combinar diferentes fuentes de datos y obteniendo relaciones entre ellas. Además su arquitectura proporciona un mejor rendimiento que otras herramientas, ya que aprovecha las capacidades de trabajar en memoria guardando los datos menos utilizados en disco y las relaciones entre los datos en memoria.

Oracle Endeca Information Discovery Integrator (34)

Integrator es una herramienta de gestión de datos, una herramienta que permite hacer extracciones, transformaciones y cargas en la base de datos mediante procesos ETL. Además permite combinar las capacidades de sistemas de adquisición de información (IAS) para la recopilación de datos de sistemas de gestión de contenidos, como por ejemplo, sitios web o blogs. Mediante una herramienta gráfica permite entrar en una URL, extraer el contenido, la estructura y combinarla con la base de datos de Endeca.

Oracle Endeca Information Discovery Studio (36)

En el "front end" se encuentra un entorno visual, Oracle Endeca Information Discovery Studio, para generar cuadros de mando. Mediante técnicas de arrastrar y soltar se pueden añadir componentes como tablas, nubes de palabras, mapas, etc. Además permite al usuario conectarse a otras fuentes de datos o agregar sus propios datos y combinarlos con los ya existentes. Los cuadros de mando pueden ser generados por el usuario final o por técnicos de desarrollo de aplicaciones, y mediante el diseño de plantillas pueden compartir vistas o cuadros de mando completos, que más tarde pueden ser modificados.

4.1.1 Oracle Endeca Server

El motor detrás de Oracle Endeca es el servidor, no es una base de datos relacional ni noSQL, es una base de datos híbrida analítica preparada para el descubrimiento de información. Es una base de datos flexible, escalable, orientada a columnas y en memoria. Esto permite una navegación fluida, búsqueda y análisis de cualquier tipo de datos estructurados o no estructurados.

Normalmente antes de realizar una carga a una base de datos hay que realizar un análisis previo en el que se modelan los datos, se estructuran las tablas y las relaciones. Endeca utiliza un modelo único de datos mediante el cual se pueden hacer búsquedas inmediatas obteniendo las relaciones de esos datos.

Endeca organiza los datos en registros. Cada registro es una secuencia de atributo-valor. Por ejemplo, un registro con tres atributos podría ser:

[{ID,1} {Nombre, Alvaro} {Universidad, Universidad Francisco de Vitoria }]

Este modelo de datos significa que cada registro puede ser diferente, no necesita tener el mismo número de atributos o el mismo número de atributo-valor, además pueden tener diferentes valores para los mismos atributos. Por ejemplo, en el mismo conjunto de datos podría haber registros como:

[{ID, 2} {Universidad, Universidad Politécnica de Madrid} {Carrera, Informática} {Edad, 22} {Comentarios, "me gusta la ufv"}]

[{ID, 3} {Deporte, Tenis} {Deporte, Baloncesto} {Universidad, Universidad Francisco de Vitoria}]

Está claro que los registros en Endeca tienen ciertas ventajas frente a los utilizados tradicionalmente en una base de datos relacional. Además, Endeca comprime los datos de manera automática y si un registro no tiene un valor para un determinado atributo, simplemente no lo asocia con ese atributo. Si por el contrario, un registro tiene varios valores para un atributo, almacena todos ellos, sin tener que duplicar el resto del registro.

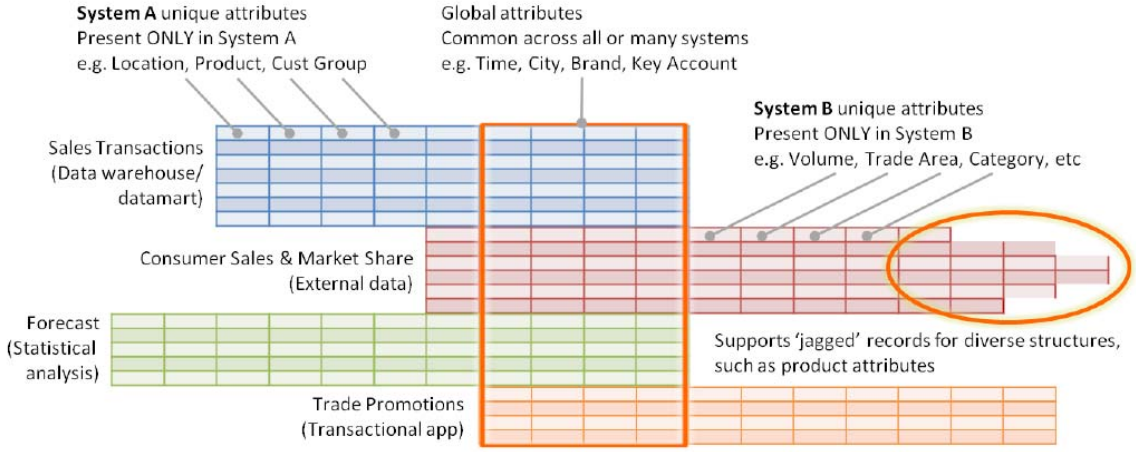


Figura 9. Estructura de la base de datos analítica de Oracle Endeca Server. Fuente: (3)

Endeca Server puede utilizar datos sin crear un modelo por adelantado. Esto reduce el tiempo de desarrollo, tanto para el equipo encargado de desarrollar las aplicaciones como para el usuario final. Por otra parte, si más tarde se quisiesen añadir más fuentes de información Endeca Server volvería a actualizar los datos de manera automática y totalmente transparente tanto para el usuario como para la aplicación.

Este motor tiene las ventajas de un entorno Hadoop, el usuario no tiene por qué conocer el formato o el tipo de ficheros que está guardando y puede acceder a todos los ficheros de manera inmediata. Además añade la ventaja que el desarrollador no tiene por qué escribir código en MapReduce para acceder a los datos ya que la aplicación lo hace automáticamente mediante un leguaje propietario llamado EQL (parecido a SQL).

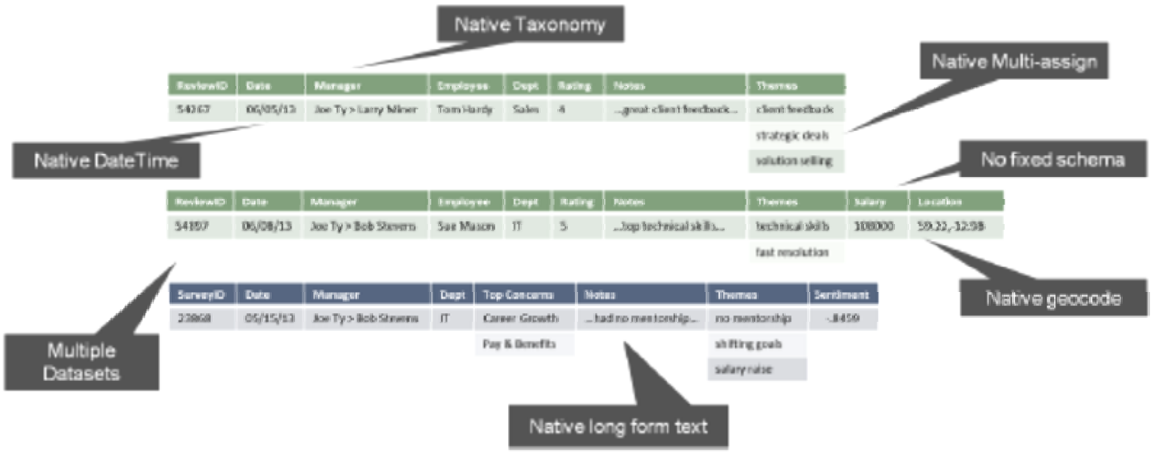


Figura 10. Atributos de la base de datos analítica. Fuente: (3)

Para cada atributo en el conjunto de datos, Endeca Server guarda dos índices, el forward index (índice hacia adelante) y el reverse index (índice hacia atrás). El forward index está ordenado por el ID del registro, permitiendo una rápida búsqueda de los valores asociados con cada registro. Esto es muy útil cuando los usuarios tienen que profundizar en la información y quiere ver la información más desagregada posible, es decir, la pieza de información. El reverse index es un índice ordenado por el valor del atributo, permitiendo así que la base de datos esté más optimizada en los casos en los que el usuario quiera analizar la distribución de los valores en los datos, como agregaciones, filtros o la propia navegación.

Cada uno de los registros, en lugar de almacenar los valores de los atributos por sí solos, son punteros a las posiciones de los índices de los atributos. El conjunto de índices asociados a un atributo se denomina modelo de un atributo, que son guardados en memoria. Cuando un atributo se actualiza con un nuevo valor, la columna entera se actualiza.

Para aprovechar las ventajas de los diferentes criterios de ordenación, cada índice de atributos es prefijado con una estructura de datos en árbol, esto acelera la búsqueda de los registros y valores. Las columnas que son frecuentemente usadas se almacenan en la memoria caché para aumentar la velocidad. De esta manera aunque se carguen todos los datos al inicio, a medida que el usuario vaya restringiendo el conjunto de datos, esas columnas se cargarán en caché aumentando la velocidad en los siguientes cambios que se hagan en la fase de exploración. Existen opciones de cargar todos los datos en memoria aumentando considerablemente la velocidad pero esto limita el tamaño del conjunto de datos.

Cada modelo de atributo es de un tipo específico, permitiendo comprimir los datos en función del tipo de dato. Aunque soporta multitud de tipos de datos numéricos, fechas, booleanos, etc. los tipos de datos que optimiza son:

- **Coordenadas de posicionamiento:** utiliza dos índices, uno para la latitud y otro para la longitud.
- **Jerarquía de valores** (por ejemplo, Persona->Alumno->Estudiante UFV): apunta a la posición en la estructura del árbol de datos, es decir, cuando se solicita a Endeca por un padre, este devuelve el padre y los hijos, aunque estos no pertenezcan al registro del padre. Esta capacidad hace que aumente la velocidad.
- **Cadenas de texto:** cada palabra o valor distinto se guarda solo una vez. En lugar de guardar los valores repetidos, los modelos de atributos guardan las referencias a las posiciones donde se encuentran. Esta práctica hace que determinadas consultas sean más rápidas y que disminuya el tamaño del conjunto de datos.

Enriquecimiento de texto

Endeca no solo permite indexar registros de diferentes fuentes y en diferentes formatos, sino que además mediante un enriquecimiento de texto obtiene palabras clave que facilitan las búsquedas y la navegación por los datos.

- **Whitelist:** mediante una whitelist o lista blanca, se pueden definir los KPIs, propios de cada negocio, para que el usuario decida e integre en el servidor permitiendo buscar relaciones entre el conjunto de datos. Estos KPIs pueden estar en idiomas diferentes ya que Endeca reconoce hasta 35 idiomas.
- **Extracción de términos:** una de las potencias de Endeca es la extracción de términos en texto no estructurado. Este enriquecimiento se realiza en 7 idiomas y permite obtener palabras clave así como sus relaciones de manera automática.
- **Análisis de sentimiento:** Endeca utiliza el motor Salience de Lexalytics. Aunque este análisis de sentimiento se verá con más profundidad en la sección 4.1.4, es importante recalcar algunos puntos. No solo se trata de una extracción de palabras clave, sino de un análisis de sentimiento en el que se interpreta el texto no estructurado y se puntúa si es positivo, negativo o neutral. Además el análisis de sentimiento obtiene temas, citas, nombres de entidades (personas, lugares y empresas) y un resumen del documento.

4.1.2 Oracle Integrator ETL

Es una solución ETL basada en Clover ETL que permite al usuario mediante componentes generar grafos que pueden ser ejecutados desde el propio entorno de desarrollo o desde un servidor de aplicaciones. Permite conectar una amplia gama de fuentes de datos como ficheros XML, JSON, XLS, bases de datos relacionales, sistemas Hadoop y otros muchos. Además se pueden hacer transformaciones como agregar datos, concatenar, hallar intersecciones, eliminar duplicados, filtrar, ordenar, normalizar y unir con otros conjuntos de datos entre otras muchas funcionalidades. Una vez finalizado todo el proceso se carga en el Endeca Server.

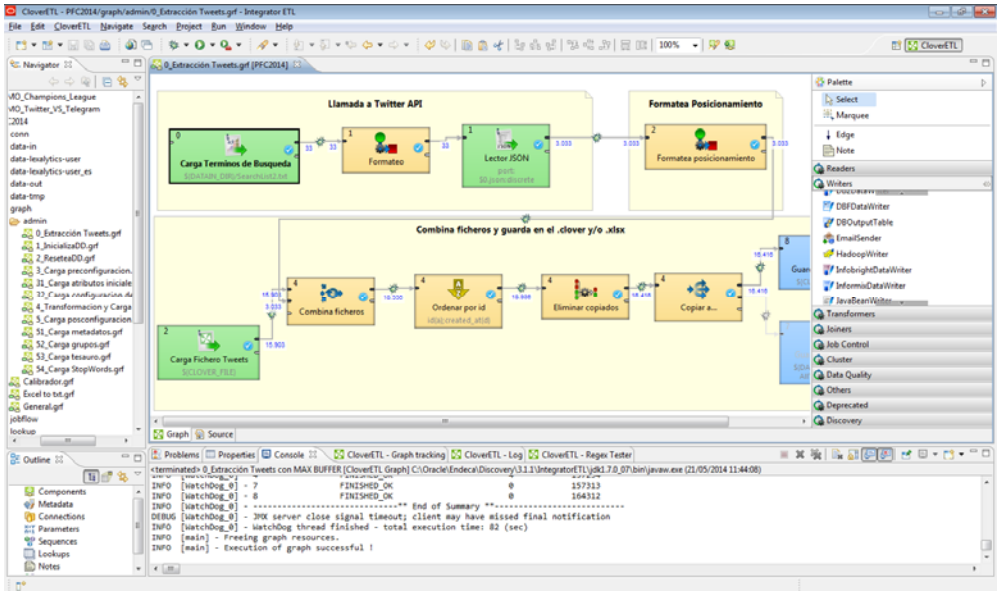


Figura 11. Oracle Endeca Integrator. Fuente: elaboración propia.

Oracle Integrator ETL, permite una programación basada en componentes. En este proyecto se utilizará además de para realizar todas las transformaciones de los datos y carga a Endeca, para llamar a la Rest API (4) de Twitter.

Además estos grafos ETL también pueden ser cargados al Integrator Server, otro componente de Oracle ejecutándose sobre un servidor de aplicaciones, y programar su ejecución así como salidas, entradas y otras modificaciones.

4.1.3 Oracle Endeca Studio

Esta herramienta proporciona la capa de presentación al sistema Endeca, está basado en Liferay (36) que es un portal de gestión de contenidos (CMS). El objetivo principal de esta herramienta es que el usuario final pueda navegar por la información respondiendo a preguntas propias de su negocio y formularse nuevas preguntas que antes no podía hacerse. Además mediante un servicio de aprovisionamiento los usuarios pueden subir sus propias hojas de cálculo o con fuentes de datos que hayan desarrollado los departamentos de IT y combinarlo con los datos existentes en Endeca Server.

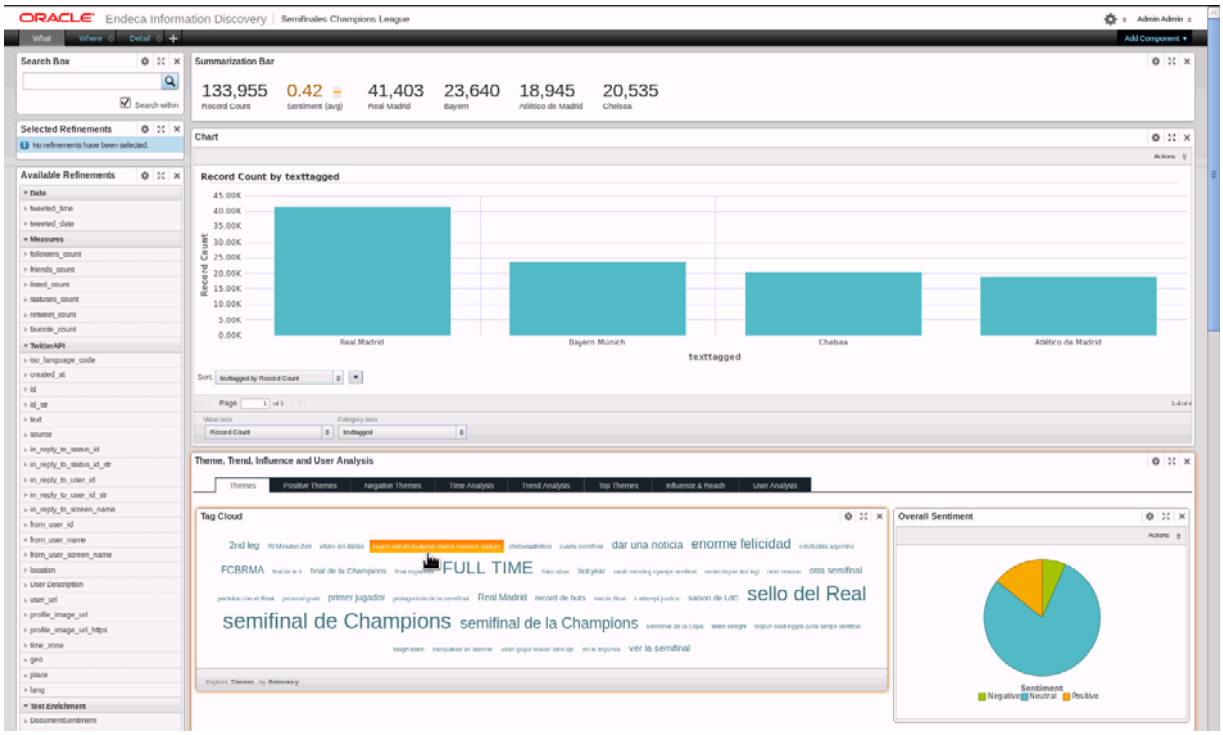


Figura 12. Oracle Endeca Studio. Fuente: elaboración propia.

Mediante un cuadro de mandos interactivo el usuario puede agregar nuevas funcionalidades como gráficos, tablas dinámicas, tablas de resultados, mapas de calor, indicadores, nubes de palabras y otros muchos componentes.

El acceso a la aplicación se puede integrar con sistemas de autentificación como LDAP, Active Directory, etc. facilitando así el acceso y teniendo un control de la aplicación. Además

se puede restringir los datos a los que puede acceder cada usuario en función de su rol, así como auditar la aplicación para saber cómo, cuándo y por quién ha sido usada.

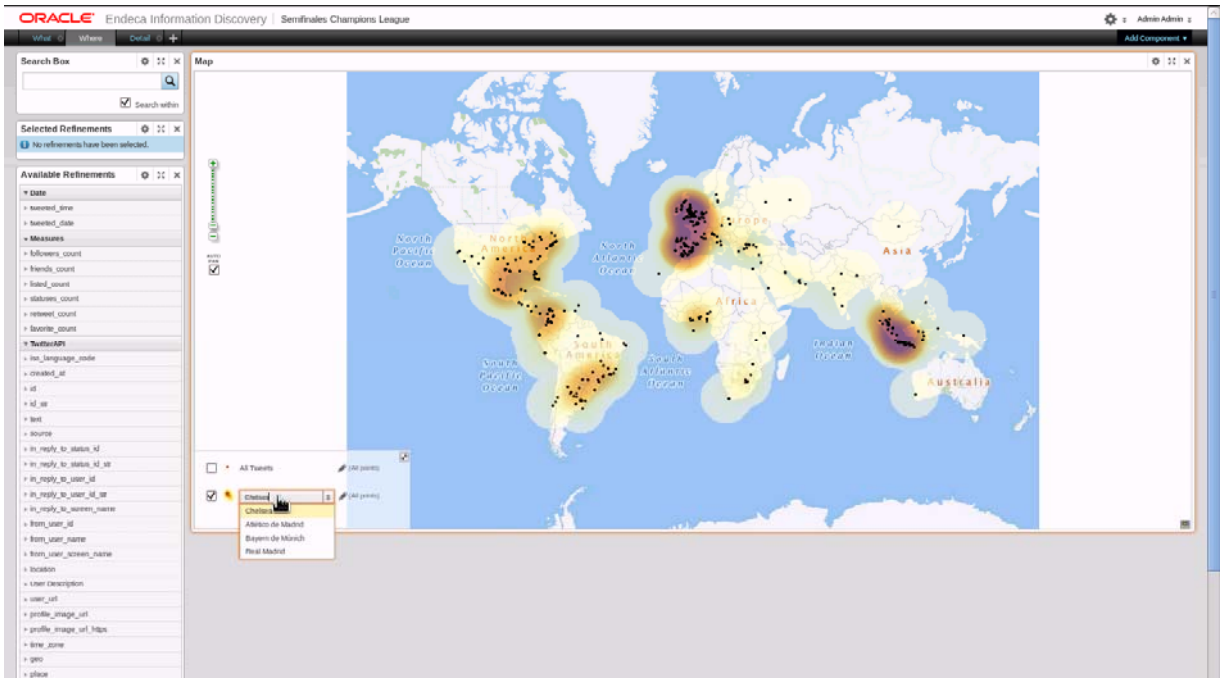


Figura 13. Mapa de Oracle Endeca Studio. Fuente: elaboración propia.

A la hora generar mapas, Endeca sitúa las coordenadas de latitud y longitud de los registros que se indiquen. Una vez se obtengan los resultados, se puede indicar que filtre todos aquellos registros que estén en una determinada ubicación. Por ejemplo, se pueden fijar los registros que estén en Madrid y todos aquellos que estén a menos de 100 kilómetros de Madrid.

4.1.4 Lexalytics Salience Engine

Es un analizador de texto multilingüe escrito en C/C++ que proporciona un enriquecimiento natural del texto no estructurado (32). Procesa el lenguaje natural mediante tecnologías de análisis de texto incluyendo el análisis de sentimiento, la extracción de entidades, extracción de temas, resúmenes y extracción de atributos ente otros.

4.1.4.1 Extracción de entidades

La extracción de entidades es una de las tecnologías más antiguas dentro del análisis de texto, desde finales de 1990 ha sido uno de los pilares en la comprensión del texto no estructurado.

Lo primero que debemos comprender es qué es una entidad. Para este contexto entenderemos por entidad como un valor de una lista, como por ejemplo es “Álvaro” para “Persona” o “iPhone” para “Producto”. Cosas como “energía solar” o “baloncesto” no pueden ser entidades porque son nombres genéricos.

Para encontrar las entidades se utiliza un conjunto de técnicas diferentes:

- **Listas**: mediante una comparación de palabras se puede saber si se trata de una entidad. Utilizado para nombres famosos, nombres de empresa o de productos.
- **Expresiones regulares:** mediante expresiones regulares se puede saber si se trata de un número de teléfono, un email, códigos postales, etc.
- **Modelos CRF:** son modelos predictivos que permiten saber si la palabra se trata de una entidad o no, sin tenerla guardada en una lista. Estos modelos utilizan técnicas de reconocimiento similares a las que utilizamos nosotros. Por ejemplo, normalmente los nombres de las empresas se sitúan después de la frase "trabaja para". Este conocimiento es guardado dentro de estos modelos, por tanto necesitan ser entrenados previamente. Para construir un modelo CRF fiable que detecte personas y nombres de empresas se necesitan entre 3.000 y 5.000 documentos previamente procesados a mano (37). Aun así no son modelos 100% fiables.

El motor de Salience utiliza un modelo híbrido combinando modelos CRF y modelos de listas, en las que el usuario puede definir sus propias listas para asegurarse de que son detectados por el sistema.

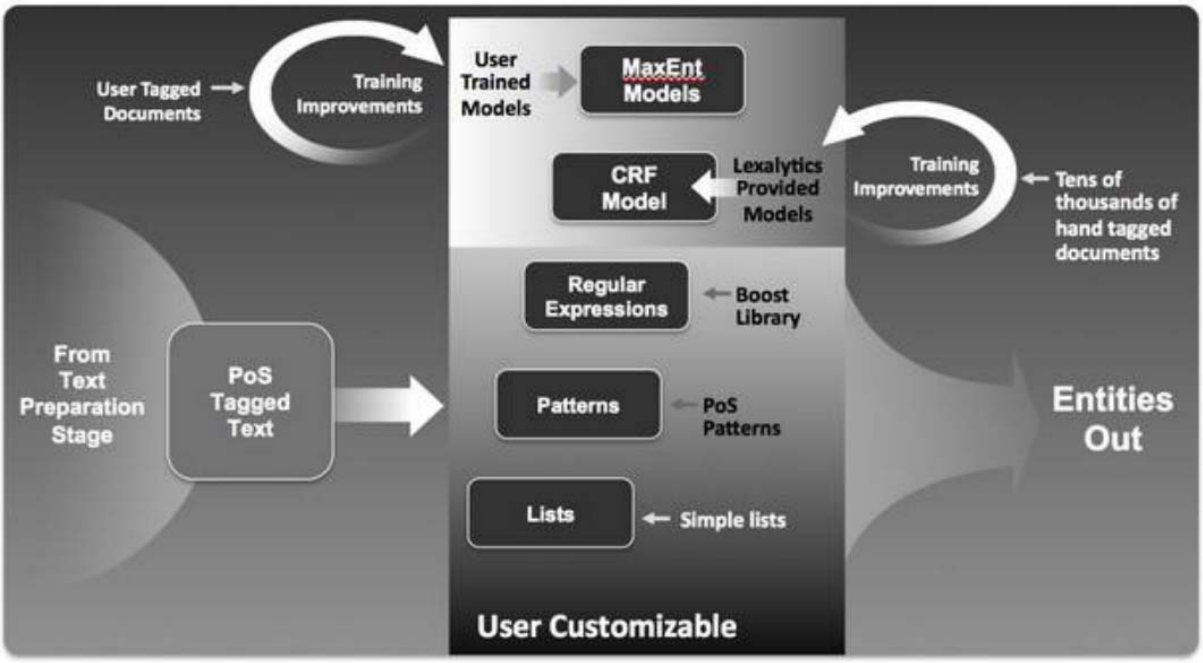


Figura 14. Extracción de entidades de Lexalytics. Fuente (37)

La extracción de entidades detecta personas, lugares, empresas, productos y las entidades que el usuario mediante listas defina para que el motor analítico detecte.

4.1.4.2 Análisis de sentimiento

El análisis de sentimiento permite a un ordenador saber si una frase, fragmento de texto o un documento entero es positivo o negativo. Es muy utilizado para saber la reacción del público en una nueva campaña de marketing o analizar las encuestas de un determinado evento sin tener que leer los documentos.

Uno de los problemas más importantes a la hora de hacer un análisis de sentimiento es la interpretación del texto. El mismo fragmento puede ser para una persona muy positivo, mientras que para otra puede ser neutral o incluso negativo, dependiendo del contexto. El sistema identifica fragmentos emotivos en un documento y las puntúa de -1 a 1, luego combina esas puntuaciones para saber el sentimiento general de la frase y del documento.

El primer paso es descomponer el documento en fragmentos en función de la estructura gramatical ya sean nombres, verbos, adjetivos, adverbios, etc. En la mayoría de los casos las estructuras que denotan sentimiento se tratan de combinaciones de adjetivos con nombres. Normalmente cuando una persona analiza un fragmento de texto no califica palabra por palabra si no que tiene un sentimiento general del fragmento. Estas puntuaciones vienen predeterminadas por la frecuencia de esa palabra o conjunto de palabras aparece cerca del conjunto de palabras calificadas como positivas o del conjunto de palabras calificadas como negativas.

Mediante diccionarios se evalúa la cercanía de palabras positivas o negativas y se obtiene una puntación de cada palabra en caso de detectarla. Este recuento se combina utilizando una operación matemática llamada “razón de oportunidades” para determinar la puntuación de la frase.

Otro problema muy común es que normalmente los documentos no son homogéneos por tanto no son positivos o negativos. El sentimiento está localizado en algún conjunto de palabras más limitado. Para solucionar este aspecto, el motor detecta las entidades de las que se habla en el fragmento de texto y evalúa el sentimiento asociado a la entidad. De esta manera la aproximación es óptima.

4.2 API Twitter

Twitter ofrece sus servicios de manera gratuita no solo a los usuarios sino que a los desarrolladores que necesiten explotar esos datos. Mediante varias APIs da acceso a gran parte del contenido.

4.2.1 Rest API v1.1

Mediante una aplicación proporcionada por Twitter el desarrollador puede hacer consultas bajo demanda. Además la aplicación es parametrizable permitiendo así una búsqueda más exhaustiva de los datos que se quieren obtener.

Cuando el desarrollador hace una consulta, la API de Twitter devuelve un fragmento de texto en formato JSON con los tuits más recientes. Esta búsqueda está limitada en una ventana de

tiempo de 7 días concentrándose la mayor parte en el día más cercano a la consulta y en el anterior.

Limitaciones (38)

Aunque desde Twitter no se especifica un número máximo de tuits que devuelve por consulta, en la práctica podemos hablar aproximadamente 50.000 tuits por hora y por usuario. Este número varía en función del número de tuits que se hayan publicado sobre ese. No es lo mismo buscar "realmadrid" que "ullate".

Respecto al número de consultas, la Rest API v.1.1 de Twitter permite 180 llamadas por usuario o 450 llamadas por aplicación cada 15 minutos, siempre y cuando utilicemos la función search/tuits (4). Además Twitter no asegura que la muestra que puedas obtener mediante numerosas llamadas sea el 100% de los tuits que solicitas.

4.2.2 Streaming API

Esta aplicación permite tener un canal directo abierto con Twitter para que a medida que se produzca el fragmento se envíe a los usuarios que están conectados en tiempo real. De esta manera se obtienen los datos al momento y sin tener que programar consultas periódicas.

Limitaciones

El mayor problema de esta API es la limitación del número total de tuits que se obtienen, aunque no hay fuentes formales, la comunidad de desarrolladores lo estiman en un 1% del total.

4.3 Equipo físico

Para la realización de este proyecto se ha utilizado un portátil con las especificaciones de la Tabla 4, una máquina virtual, especificada en la Tabla 3, y un sistema de desarrollo en la nube basado en servicios webs con el proveedor de Amazon (2).

4.3.1 Equipo de desarrollo y pruebas

Para el desarrollo del sistema se han utilizado diferentes sistemas. Para la adquisición de datos, utilizando la Streaming API de Twitter, se ha utilizado una micro instancia en Amazon EC2 conectada a una base de datos mySQL en Amazon RDS. Por otra parte, para el desarrollo de la aplicación así como todos los sistemas necesarios como el servidor de aplicaciones y componentes necesarios para el funcionamiento de Oracle Endeca se ha utilizado una máquina virtual con las siguientes características.

Tabla 3. Características principales de la máquina virtual. Fuente: elaboración propia.

Máquina Virtual	
Procesador	2 virtuales (Intel Core i5 M520)
Memoria	6 GB
Sistema Operativo	Oracle Linux 6.5
Disco Duro	200 GB de espacio dinámico

Toda la configuración necesaria para el desarrollo se encuentra adjunta al proyecto en el fichero “Manual.pdf”.

Por último, para la construcción del ETL mediante la herramienta Integrator ETL, se ha utilizado las características del ordenador anfitrión instalándolo en el mismo.

Tabla 4. Características principales del portátil. Fuente: elaboración propia.

Ordenador anfitrión	
Procesador	2 físicos, 4 virtuales (Intel Core i5 M520)
Memoria	8 GB
Sistema Operativo	Windows 7
Java	1.6.0_45
Disco Duro	320 GB

A lo largo de la fase de desarrollo y verificación se expondrá el proceso de configuración detalladamente.

4.3.2 Adquisición de datos en la nube

Para poder mantener una conexión abierta con Twitter es necesario mantener un proceso en constante ejecución. Se ha decidido utilizar los servicios web gratuitos proporcionados por Amazon (2). El proceso completo viene documentado a lo largo de la fase de desarrollo y verificación del proyecto.

Figura 15. Esquema conceptual de la arquitectura en Amazon. Fuente: elaboración propia.

4.3.2.1 Amazon EC2

Amazon Elastic Compute Cloud (Amazon EC2) es un servicio web que proporciona un procesamiento en la nube. De manera gratuita se puede utilizar una micro instancia de hasta 613 MB de memoria RAM y un procesador (2). La capacidad de procesamiento de la mirco instancia no es suficiente para instalar el sistema completo, pero sí para instalar un servidor HTTP Apache y ejecutar algunos procesos que se encarguen de mantener una conexión abierta con Twitter y así utilizar la Streaming API.

Aunque el uso de servicios web o sistemas en la nube tiene algunas ventajas importantes como son la escalabilidad y el precio, en este proyecto se ha elegido porque el sistema puede estar ejecutándose de manera gratuita 750 horas al mes (2), es decir, 24 horas al día sin necesidad de pararlo.

4.3.2.2 Amazon RDS

Amazon Relational Database Service (Amazon RDS) es un servicio web de Amazon que proporciona un sistema de base de datos relacional (Oracle, MySQL y PostgreSQL) en la nube. Además facilita la configuración y administración así como la escalabilidad en caso de necesitar aumentar el tamaño de base de datos. La capa gratuita de Amazon incluye 20 GB de almacenamiento y 20 GB para copias de seguridad (2).

4.4 Arquitectura

A lo largo del proyecto se han implementado diferentes arquitecturas de adquisición de datos. La primera arquitectura construida sobre los servicios web gratuitos de Amazon para ejecutar la Streaming API de Twitter. El segundo diseño está construido en local, para utilizar la Rest API de Twitter.

4.4.1 Arquitectura Streaming API

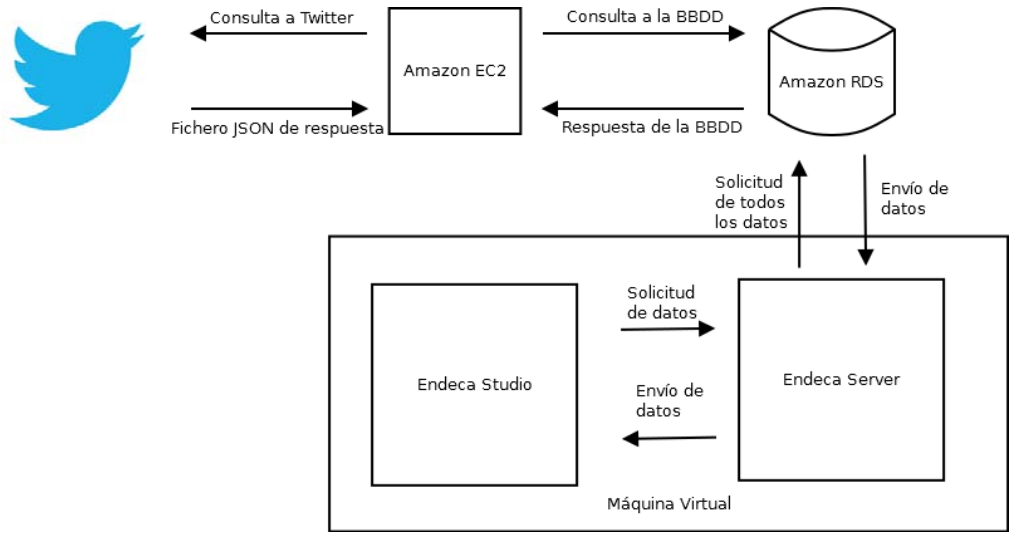


Figura 16. Arquitectura conceptual para la utilización de la Streaming API. Fuente: elaboración propia.

Aunque el proceso viene documentado en la fase de desarrollo y verificación, el proceso completo es el siguiente.

1 **Amazon EC2:** desde la instancia levantada en la nube se ha configurado un proceso que mantiene una conexión abierta y constante con el servidor de Twitter, en espera de nuevos datos. A medida que los datos vayan creándose, el servidor de Twitter le enviará el dato al servidor de Amazon.

2 **Amazon RDS:** el mismo proceso que mantiene la conexión abierta guardará en la base de datos los tuits que vayan llegando.
3 **Oracle Endeca Server:** el servidor de Endeca se conectará con la base de datos en la nube para solicitar los nuevos datos que vayan guardándose. Estos se cargarán en el servidor cuando esté conectado.
4 **Oracle Endeca Studio:** este es componente es la interfaz gráfica desde la que se crearán los cuadros de mando para poder visualizar los datos y obtener los resultados. Realizará las consultas al servidor bajo petición del usuario.

4.4.2 Arquitectura Rest API

Se ha configurado una segunda arquitectura para la utilización de la Rest API de Twitter. El proceso completo viene especificado en la fase de desarrollo y verificación del proyecto.

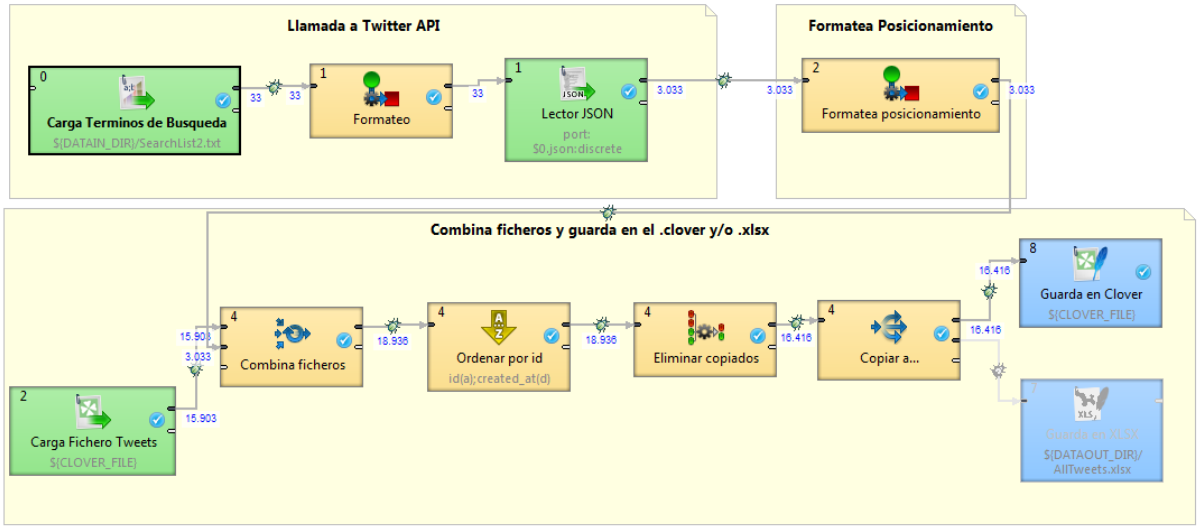


Figura 17. Gráfico ETL para el uso de la Rest API de Twitter. Fuente: elaboración propia.

El procedimiento para llamar a la Rest API de Twitter es el siguiente:

- **Componente "Carga Términos de Búsqueda":** abre y lee un fichero de texto que contiene los términos de búsqueda. Cuando se solicita un conjunto de tuits, en la llamada a la API se incluye el término a buscar. Este término puede estar en el contenido del tuit, en el nombre de usuario o en cualquier otro campo de los tuits.
- **Componente "Formateo":** realiza la llamada, en Java más adelante en el apartado 6.3.3.1 se analizará en detalle.
- **Componente "Lector JSON":** este componente lee el fichero JSON que devuelve la API y lo separa en campos o metadatos por cada tuit recibido. Como se ve en la figura 17, en este caso, la entrada son 33 ficheros y la salida son 3.033 tuits.

- **Componente "Formatea posicionamiento":** la API de Twitter devuelve las coordenadas de posicionamiento en dos metadatos: latitud y longitud. Es conveniente combinarlos en uno con el formato "latitud , longitud"
- **Componente "Carga Fichero Tuits":** este componente carga los tuits que tenemos almacenados procedentes de otras búsquedas.
- **Componente "Combina ficheros":** este componente combina los dos flujos de datos en uno solo, incluyendo campos repetidos.
- **Componente "Ordenar por ID":** se ordenan los datos por el ID único del tuit.
- **Componente "Eliminar duplicados":** se eliminan los tuits que tienen el mismo ID.
- **Componente "Copiar a":** el componente recoge el flujo de datos con sus metadatos y los copia a diversas fuentes.
- **Componente "Guarda en Clover":** este componente guarda los datos y metadatos en un fichero ".clover" que servirá de contenedor de los tuits antes de cargarlos a Endeca Server. A la hora de guardar los tuits se podría haber utilizado un componente para guardarlos en una base de datos relacional siendo esta más rápida, ahora bien, si se va a trabajar con pocos datos podemos utilizar ficheros de datos facilitando los backups y sin tener que arrancar en el equipo de desarrollo una base de datos.
- **Componente "Guarda en XLSX":** este componente se ha utilizado durante las pruebas de adquisición de datos para comprobar que todos los tuits quedan formateados correctamente.

4.5 Planificación

Para la realización de este proyecto, siguiendo la metodología de desarrollo en espiral, se ha comenzado con la especificación y diseño del proyecto, seguido cuatro fases del ciclo en espiral y la documentación del proyecto (Figura 22).

- Especificación y diseño del proyecto
- Desarrollo y verificación
  - Fase 1: Adquisición de datos en la nube e integración de componentes
  - Fase 2: Proceso ETL para la adquisición de datos
  - Fase 3: Análisis de sentimiento
  - Fase 4: Visualización
- Documentación

A continuación se detalla la planificación llevada en cada fase.

4.5.1 Planificación fase 1

Para la primera fase se ha seguido la siguiente planificación (figura 18).

Fase 1: Adquisición de datos en la nube e integración de componentes	30 days?	**6/01/14 8:00**	**14/02/14 17:00**
Determinar objetivos	2 days	**6/01/14 8:00**	**7/01/14 17:00**
Adquisición de datos	2 days	6/01/14 8:00	7/01/14 17:00
Integración de componentes	2 days	6/01/14 8:00	7/01/14 17:00
Análisis de riesgo	3 days?	**8/01/14 8:00**	**10/01/14 17:00**
Identificación de riesgos	1 day?	8/01/14 8:00	8/01/14 17:00
Análisis de riesgos	1 day?	8/01/14 8:00	8/01/14 17:00
Planificación de riesgos	1 day?	9/01/14 8:00	9/01/14 17:00
Supervisión de riesgos	1 day?	10/01/14 8:00	10/01/14 17:00
Desarrollo y verificación del producto	24 days?	**13/01/14 8:00**	**13/02/14 17:00**
Adquisición de datos	10 days?	**13/01/14 8:00**	**24/01/14 17:00**
Preparación del entorno	9 days?	**13/01/14 8:00**	**23/01/14 17:00**
Ejecución y verificación	1 day?	24/01/14 8:00	24/01/14 17:00
Integración componentes Oracle Endeca	10 days?	**27/01/14 8:00**	**7/02/14 17:00**
Integración Oracle Enterprise Linux	2 days?	27/01/14 8:00	28/01/14 17:00
Integración Oracle WebLogic	1 day	29/01/14 8:00	29/01/14 17:00
Oracle Database	2 days?	30/01/14 8:00	31/01/14 17:00
Oracle Endeca Server	2 days	3/02/14 8:00	4/02/14 17:00
Oracle Endeca Studio	2 days	5/02/14 8:00	6/02/14 17:00
Oracle Endeca Provisioning	1 day	7/02/14 8:00	7/02/14 17:00
Resultados y verificación	4 days?	**10/02/14 8:00**	**13/02/14 17:00**
Resultados adquisición de datos	2 days	10/02/14 8:00	11/02/14 17:00
Resultados integración de componentes	3 days	11/02/14 8:00	13/02/14 17:00
Resultados visualización	2 days	12/02/14 8:00	13/02/14 17:00
Resultado análisis de riesgos	1 day?	13/02/14 8:00	13/02/14 17:00
Planificación	1 day?	14/02/14 8:00	14/02/14 17:00

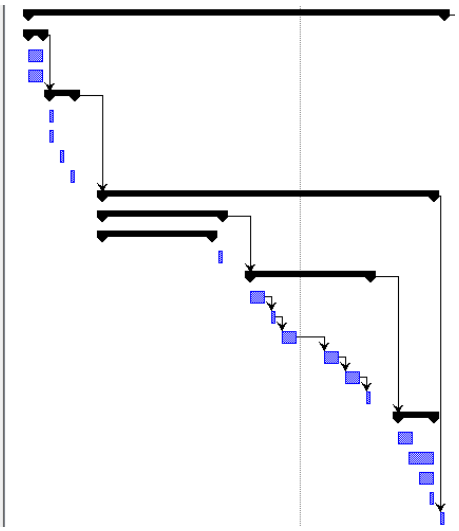


Figura 18. Planificación fase 1. Fuente: elaboración propia.

4.5.2 Planificación fase 2

En la segunda fase, la figura 19, muestra su planificación.

Fase 2: Proceso ETL para la adquisición de datos	35 days?	**17/02/14 8:00**	**4/04/14 17:00**
Determinar objetivos	4 days	**17/02/14 8:00**	**20/02/14 17:00**
Proceso ETL: extracción	4 days	17/02/14 8:00	20/02/14 17:00
Proceso ETL: transformación	4 days	17/02/14 8:00	20/02/14 17:00
Proceso ETL: carga	4 days	17/02/14 8:00	20/02/14 17:00
Análisis de riesgos	1 day?	**21/02/14 8:00**	**21/02/14 17:00**
Identificación de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Análisis de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Planificación de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Supervisión de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Desarrollo y verificación	29 days?	**24/02/14 8:00**	**3/04/14 17:00**
Proceso ETL: extracción y transformación	25 days	24/02/14 8:00	28/03/14 17:00
Proceso ETL: carga en el servidor	20 days	3/03/14 8:00	28/03/14 17:00
Reultados	3 days	31/03/14 8:00	2/04/14 17:00
Resultados análisis de riesgos	1 day?	3/04/14 8:00	3/04/14 17:00
Planificación	1 day?	4/04/14 8:00	4/04/14 17:00

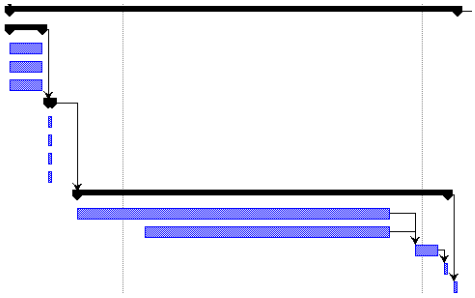


Figura 19. Planificación fase 2. Fuente: elaboración propia.

4.5.3 Planificación fase 3

La figura 20 muestra la planificación llevada a cabo en la tercera fase.

Fase 3: Análisis de sentimiento	19 days?	**8/04/14 8:00**	**2/05/14 17:00**
Determinar objetivos	3 days	**8/04/14 8:00**	**10/04/14 17:00**
Análisis de sentimiento	2 days	8/04/14 8:00	9/04/14 17:00
Modificación proceso ETL	2 days	9/04/14 8:00	10/04/14 17:00
Análisis de riesgos	1 day?	**11/04/14 8:00**	**11/04/14 17:00**
Identificación de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Análisis de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Planificación de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Supervisión de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Desarrollo y verificación	14 days?	**14/04/14 8:00**	**1/05/14 17:00**
Modificar ETL	10 days	14/04/14 8:00	25/04/14 17:00
Análisis de sentimiento	12 days	16/04/14 8:00	1/05/14 17:00
Resultados análisis de riesgo	1 day?	1/05/14 8:00	1/05/14 17:00
Planificación	1 day?	2/05/14 8:00	2/05/14 17:00

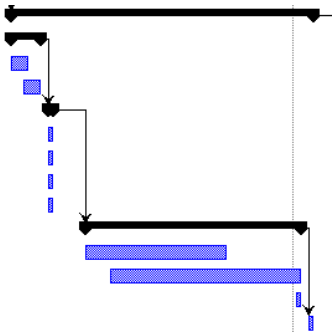


Figura 20. Planificación fase 3. Fuente: elaboración propia.

4.5.4 Planificación fase 4

La figura 21, muestra la planificación de la cuarta fase.

Fase 4: Visualización	20 days?	**5/05/14 8:00**	**30/05/14 17:00**
Determinar objetivos	4 days?	**5/05/14 8:00**	**8/05/14 17:00**
Integración datos universidades y alumnos	2 days	5/05/14 8:00	6/05/14 17:00
Diseño de los cuadros de mando	2 days	6/05/14 8:00	7/05/14 17:00
Análisis de sentimiento	1 day?	8/05/14 8:00	8/05/14 17:00
Análisis de riesgos	1 day?	**9/05/14 8:00**	**9/05/14 17:00**
Identificación de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Análisis de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Planificación de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Supervisión de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Desarrollo y verificación	15 days?	**12/05/14 8:00**	**30/05/14 17:00**
Análisis de sentimiento con datos de las universidades y pruebas	10 days	12/05/14 8:00	23/05/14 17:00
Desarrollo cuadro de mandos: unviersides	5 days	16/05/14 8:00	22/05/14 17:00
Desarrollo cuadro de mandos: Análisis de sentimiento	5 days	19/05/14 8:00	23/05/14 17:00
Desarrollo cuadro de mandos: usuarios	5 days	22/05/14 8:00	28/05/14 17:00
Desarrollo cuadro de mandos: mapas	4 days	26/05/14 8:00	29/05/14 17:00
Verificación y pruebas	9 days	19/05/14 8:00	29/05/14 17:00
Resultados análisis de riesgos	1 day?	30/05/14 8:00	30/05/14 17:00

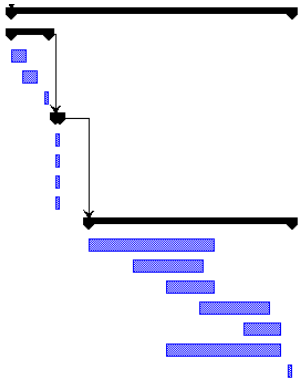
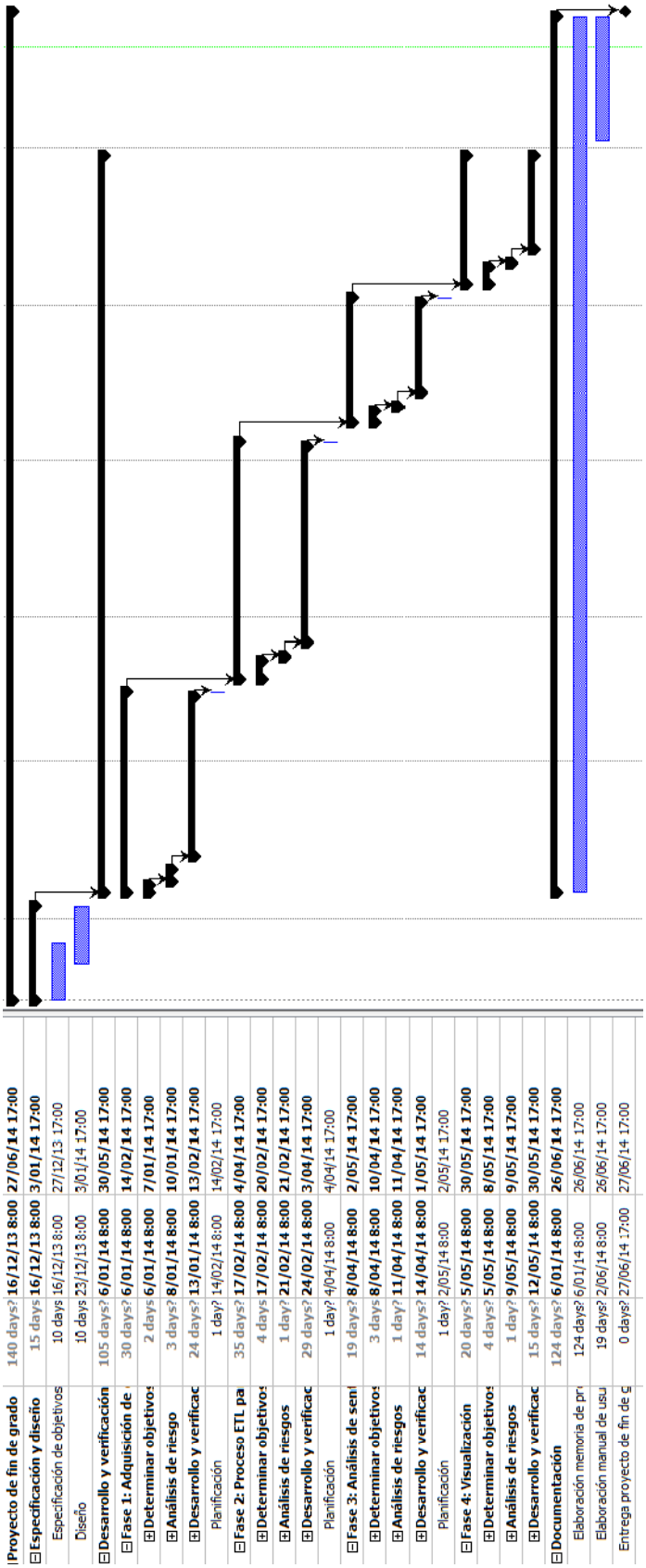


Figura 21. Planificación fase 4. Fuente: elaboración propia.

La siguiente figura, figura 22, contiene la planificación global llevada a cabo en el proyecto



Proyecto de fin de grado	140 days?	16/12/13 8:00	27/06/14 17:00	
Especificación y diseño	15 days	16/12/13 8:00	3/01/14 17:00	
Especificación de objetivos	10 days	16/12/13 8:00	27/12/13 17:00	
Diseño	10 days	23/12/13 8:00	3/01/14 17:00	
Desarrollo y verificación	105 days?	6/01/14 8:00	30/05/14 17:00	
Fase 1: Adquisición de	30 days?	6/01/14 8:00	14/02/14 17:00	
Determinar objetivos	2 days	6/01/14 8:00	7/01/14 17:00	
Análisis de riesgo	3 days?	8/01/14 8:00	10/01/14 17:00	
Desarrollo y verificac	24 days?	13/01/14 8:00	13/02/14 17:00	
Planificación	1 day?	14/02/14 8:00	14/02/14 17:00	
Fase 2: Proceso ETL pa	35 days?	17/02/14 8:00	4/04/14 17:00	
Determinar objetivos	4 days	17/02/14 8:00	20/02/14 17:00	
Análisis de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00	
Desarrollo y verificac	29 days?	24/02/14 8:00	3/04/14 17:00	
Planificación	1 day?	4/04/14 8:00	4/04/14 17:00	
Fase 3: Análisis de sent	19 days?	8/04/14 8:00	2/05/14 17:00	
Determinar objetivos	3 days	8/04/14 8:00	10/04/14 17:00	
Análisis de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00	
Desarrollo y verificac	14 days?	14/04/14 8:00	1/05/14 17:00	
Planificación	1 day?	2/05/14 8:00	2/05/14 17:00	
Fase 4: Visualización	20 days?	5/05/14 8:00	30/05/14 17:00	
Determinar objetivos	4 days?	5/05/14 8:00	8/05/14 17:00	
Análisis de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00	
Desarrollo y verificac	15 days?	12/05/14 8:00	30/05/14 17:00	
Documentación	124 days?	6/01/14 8:00	26/06/14 17:00	
Elaboración memoria de pr	124 days?	6/01/14 8:00	26/06/14 17:00	
Elaboración manual de usu	19 days?	2/06/14 8:00	26/06/14 17:00	
Entrega proyecto de fin de g	0 days?	27/06/14 17:00	27/06/14 17:00	

Figura 22

5 Metodología

Para la realización del proyecto se ha utilizado la metodología en de desarrollo en espiral. Para ello se analizarán a los largo del capítulo diferentes modelos de proceso, se valorarán y al final se seleccionará el más adecuado.

5.1 Elección de la metodología

El desarrollo de cualquier sistema informático normalmente se divide en diferentes procesos de software, también llamado ciclo de vida del desarrollo de software, aunque existen multitud de ellos hay algunos que son comunes para la mayoría:

- **Especificación del software:** se definen las funcionalidades del software así como las restricciones del mismo.
- **Diseño e implementación del software:** en este proceso se diseña y desarrolla el software que cumple las especificaciones del proceso anterior.
- **Validación del software:** cuando el software es desarrollado se valida para asegurar que cumple los requisitos especificados.
- **Evolución del software:** una vez que el software es validado requiere una evolución para adaptarse a los cambios que vayan surgiendo para cubrir las necesidades del cliente.

5.1.1 Metodologías consideradas

Para este proyecto se han considerado diferentes metodologías clásicas, algunas de ellas han sido las siguientes:

- **Modelo en cascada:** se presenta como fases separadas del proceso. Cada fase tiene como resultado documentos que debe ser aprobados por el usuario y una fase no comienza hasta que termine la anterior, normalmente se incluyen la corrección de los problemas encontrados en las fases anteriores.
- **Desarrollo evolutivo:** comienza con un desarrollo inicial para más tarde revisarla con el cliente. Estas revisiones se iteran hasta que se desarrolle el sistema que el cliente quiere. Las actividades son concurrentes.
- **Desarrollo exploratorio:** mediante este desarrollo se exploran las necesidades del cliente y se empieza por las partes que se tienen más claras. A medida que se sigue explorando se añaden al diseño del sistema hasta llegar al sistema final.
- **Evolutivo prototipado:** este desarrollo comienza con la definición de los requisitos, con ello se diseña un prototipo que sirve de ayuda para concretar los requisitos con el cliente. Los prototipos se adaptan hasta llegar a la versión final.
- **Desarrollo incremental:** se parte de una definición de requisitos, con ellos se divide el proyecto en incrementos y a medida que el proyecto avanza se iteran las fases de

desarrollo, validación de incrementos, integración de estos incrementos y velicación del sistema. En cada iteración se añaden nuevos requisitos hasta que después de sucesivas iteraciones se llega al sistema final.

5.1.2 Criterios de elección

Para la elección de la metodología más apropiada a este proyecto se han evaluado las metodologías anteriormente descritas en función de los objetivos del proyecto.

- **Funcionamiento con requisitos y arquitectura no predefinidos:** debido a que el proyecto depende de factores externos, como la adquisición de datos o el tiempo de desarrollo, es importante utilizar una metodología flexible a la hora de adaptar requisitos y arquitectura.
- **Producción de software altamente fiable:** aún teniendo factores externos, la fiabilidad del software es importante ya que es producto final del proyecto.
- **Gestión de riesgos:** es necesario medir, tener controlado y planificar un plan de actuación en los riesgos identificados.
- **Correcciones sobre la marcha:** la metodología tiene que permitir cambios rápidos sin tener que especificar desde el principio y teniendo en cuenta la gestión de riesgos.
- **Visión general del progreso:** aunque no sea un factor crítico, hay que tener en cuenta que el cliente debe poder tener visión del avance del proyecto.

Mediante la tabla 5 se comparan las metodologías propuestas para este proyecto y su valoración.

Tabla 5. Comparativa de los diferentes modelos de proceso. Fuente (39)

Modelo de proceso	Funciona con requisitos y arquitectura no predefinidos	Produce software altamente fiable	Gestión de riesgos	Permite correcciones sobre la marcha	Visión del progreso del cliente
Cascada	Bajo	Alto	Bajo	Bajo	Bajo
Evolutivo exploratorio	Medio	Medio	Medio	Medio	Medio
Evolutivo prototipado	Alto	Medio	Medio	Alto	Alto
Desarrollo incremental	Bajo	Alto	Medio	Bajo	Bajo
Desarrollo en espiral	Alto	Alto	Alto	Medio	Medio

Como no todos los criterios tienen la misma importancia necesitamos una matriz de pesos, en este caso una tabla indicando que criterios van a ser más importantes para el proyecto.

Tabla 6. Matriz de pesos para la elección de la metodología. Fuente: elaboración propia.

Funciona con requisitos y arquitectura no predefinidos	Produce software altamente fiable	Gestión de riesgos	Permite correcciones sobre la marcha	Visión del progreso del cliente
Alto	Alto	Medio	Alto	Bajo

Como salida de tabla 5 por la tabla 6 se obtiene la tabla 7. Se ha tenido en cuenta los siguientes valores:

- **Alto**= 3
- **Medio** = 2
- **Bajo** = 1

Tabla 7. Puntuación de los modelos de proceso. Fuente: elaboración propia.

Modelo de proceso	Funciona con requisitos y arquitectura no predefinidos	Produce software altamente fiable	Gestión de riesgos	Permite correcciones sobre la marcha	Visión del progreso del cliente	Puntuación Total
Cascada	3	9	2	3	1	18
Evolutivo exploratorio	6	6	4	6	2	24
Evolutivo prototipado	9	6	4	9	3	31
Desarrollo incremental	3	9	4	3	1	20
Desarrollo en espiral	9	9	6	6	2	32

Según el criterio establecido metodología que mejor se adapta a este proyecto es el desarrollo en espiral seguido por el evolutivito prototipado (39).

5.2 Desarrollo en espiral

Para la realización de este proyecto se ha utilizado la metodología tradicional de desarrollo en espiral. Esta metodología fue definida por Barry Boehm (40) en 1988 y se basa en un enfoque que entrelaza los procesos de especificación, desarrollo y validación.

Las principales ventajas de este sistema es que se desarrolla rápidamente a partir de especificaciones abstractas u objetivos y a medida que se van iterando los procesos junto con el cliente, se refina el sistema hasta su finalización. Otra ventaja es que la especificación se puede desarrollar de manera creciente. Esto es muy útil para el desarrollo de la interfaz de usuario, ya que es difícil especificarla por adelantado.

Además mediante un desarrollo exploratorio, el proyecto empieza con las partes que mejor se comprenden y el sistema evoluciona mediante una rápida retroalimentación entre los procesos.

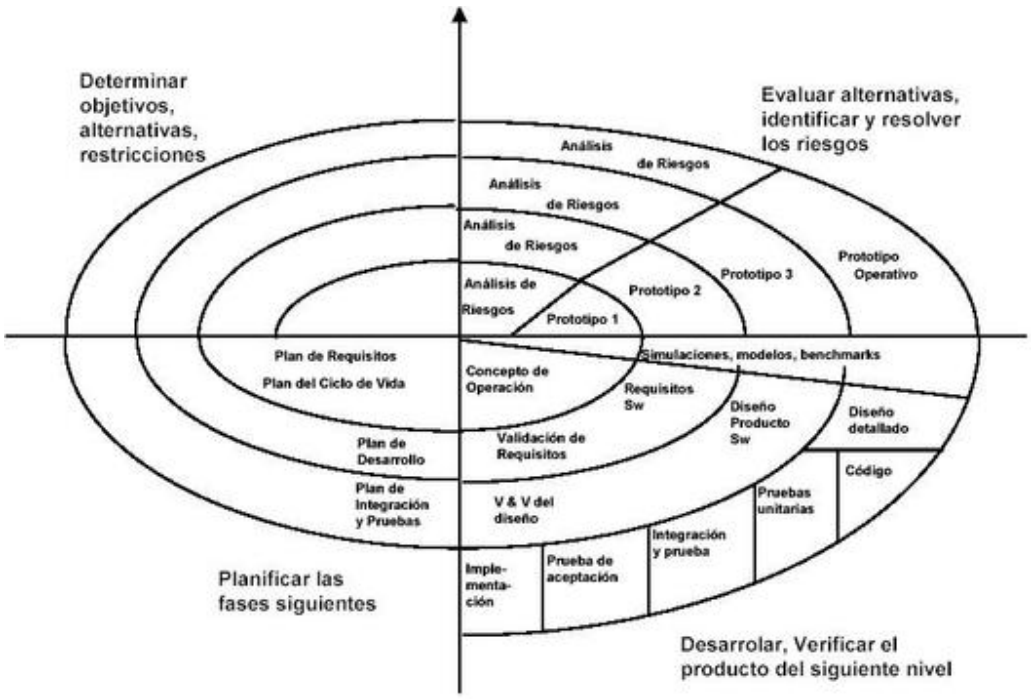


Figura 23. Iteraciones de la metodología en espiral. Fuente (39)

5.2.1 Fases del desarrollo en espiral

Cada ciclo del desarrollo en espiral tiene las siguientes etapas:

1. **Determinar o fijar los objetivos:** cada ciclo de la espiral comienza con :
   - La identificación de los objetivos de la parte del producto que está siendo desarrollado (características, funcionalidades, etc.).

- La manera de desarrollar esta porción, ya sea desarrollarlo, comprarlo, reutilizar código de otra parte, etc.
- Las restricciones impuestas por la fase anterior, como tiempo, coste, etc.

2. **Análisis de riesgo:** el siguiente paso es evaluar los riesgos de los objetivos y las limitaciones, en otras palabras, realizar un análisis de riesgo. Comienza con una identificación de las áreas que desconocemos ya que serán las que tengamos que tener controladas, medidas y tener un plan de acción para su contingencia.
3. **Desarrollo y verificación del producto:** En este paso se desarrolla la parte del producto planificada y se realizan las pruebas. Dependiendo de la fase anterior se decide cómo se va a realizar el desarrollo, ya sea en cascada, evolutivo, basado en componentes, etc.
4. **Planificación:** una característica importante del modelo en espiral es que cada ciclo se completa con una revisión que incluye al personal principal que participa en el proyecto, es decir, al cliente, organizaciones involucradas y el jefe de proyecto. Esta revisión cubre todos los productos desarrollados durante el ciclo anterior, incluyendo los planes para el próximo ciclo y los recursos que se van a necesitar.

5.3 Análisis de riesgos

La gestión de riesgos se concibe de alguna manera como la probabilidad para que una circunstancia ocurra. Estas circunstancias afectan al proyecto siempre de manera negativa ya sea amentando costes, incrementando los plazos de entrega o disminuyendo la calidad del proyecto.

En este proyecto se han identificado las siguientes categorías de riesgos:

- **Riesgos del proyecto:** afectan a los plazos de entrega o a los recursos del proyecto.
- **Riesgos del producto:** estos riesgos afectan a la calidad o al rendimiento del software.

Utilizando la metodología de desarrollo en espiral se tienen que tener en cuenta los riesgos que afectan al proyecto en cada fase, este proceso comprende varias etapas:

1. **Identificación de riesgos:** consiste en identificar los riesgos que afectan a la iteración que se va a realizar.
2. **Análisis de riesgos**: de los riesgos identificados en la etapa anterior se valoran las probabilidades de que ocurran y las posibles consecuencias de los mismos.
3. **Planificación de riesgos:** para cada riesgo es necesario planificar un plan de actuación o medida de contingencia. Esto se realiza para poder prevenir el riesgo y en caso de ocurrir tener previsto un plan de acción.
4. **Supervisión de riesgos:** una vez estén planificados los riesgos conviene tenerlos supervisados y medidos para poder actuar a la mayor brevedad posible.

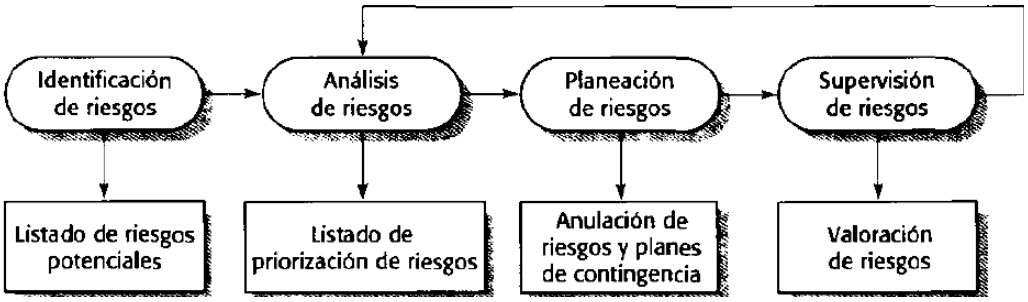


Figura 24. Proceso de gestión de riesgos. Fuente (39)

Aunque el análisis de riesgo sea un proceso complejo y que requiere tiempo, es necesario cuando el proyecto tiene restricciones críticas.

5.3.1 Identificación de riesgos

Consiste en la identificación de los riesgos. En este proyecto se identificarán los siguientes tipos de riesgos:

1. **Riesgo de tecnología:** ya sea de hardware o software a utilizar en el sistema.
2. **Riesgos de requerimientos:** estos riesgos vienen dados por los cambios de los requerimientos del sistema.

Aunque existen otros tipos de riesgos, como por ejemplo el riesgo de personal, no se ha considerado ya que más que un riesgo es uno de los objetivos del proyecto.

Otro riesgo muy común en los proyectos de software es el riesgo de estimación, es decir, aquel que se deriva de la estimación de los costes de las herramientas de software o de los recursos necesarios para el proyecto. Este riesgo tampoco ha sido considerado ya que todas las herramientas han sido obtenidas mediante un acuerdo de licencia de prueba, y se pueden utilizar siempre y cuando no se obtenga un beneficio económico con ellas.

5.3.2 Análisis de riesgos

Una vez estén los riesgos identificados se evalúa la probabilidad y la seriedad del riesgo. En este proyecto se evalúan los riesgos en función de la experiencia que se ha obtenido diseñando sistemas similares. El criterio para evaluarlos especifica en la siguiente tabla:

Tabla 8. Probabilidad del análisis de riesgo. Fuente (39)

Probabilidad	Probabilidad (%)
Muy bajo: sería excepcional	<10%
Bajo: es raro que suceda	10-25%
Moderado: es posible	25-50%
Alto: muy probable	50-75%
Muy alto: casi seguro que sucede	>75%

Además de la probabilidad de cada riesgo, se analizará la consecuencia del mismo, pudiendo ser: catastrófico, serio, tolerable o insignificante.

El resultado de este proceso es una tabla con los riesgos identificados en la fase de identificación de riesgos, la probabilidad que ocurra y las consecuencias. A medida que el proyecto madure los riesgos pueden ir cambiando, por tanto en cada iteración se analizará la probabilidad de los riesgos identificados.

5.3.3 Planificación de riesgos

Una vez se obtengan los riesgos analizados, se procede a planificar una estrategia para su gestión. Para este proyecto se han identificado las siguientes estrategias:

- **Estrategia de prevención:** para cada riesgo se planificará siempre y cuando se pueda planificar, una estrategia para disminuir la probabilidad que ocurra.
- **Plan de contingencia:** en caso que la estrategia de prevención no tenga éxito, es necesario planificar una estrategia de acción para así cumplir los objetivos del proyecto.

5.3.4 Supervisión de riesgos

Mediante técnicas gráficas como los indicadores, gráficos de barras y gráficos lineales, se supervisarán los riesgos, especialmente los que la probabilidad y el impacto sean mayores.

En esta fase se utilizará entre otras técnicas, la matriz de riesgos, que determina de manera visual la relación entre el impacto y la probabilidad de ocurrencia de cada riesgo identificado. Además el modelo

Tabla 9. Matriz de riesgos. Fuente (41)

	IMPACTO				
PROBABILIDAD	**1.- Insignificante**	**2.- Pequeño**	**3.- Moderado**	**4.- Grande**	**5.- Catástrofe**
5.- Casi seguro que sucede	Medio (5)	Alto (10)	Alto (15)	Muy alto (20)	Muy alto (25)
4.- Muy probable	Medio(4)	Medio(6)	Alto (12)	Alto (16)	Muy alto (20)
3.- Es posible	Bajo (3)	Medio (5)	Medio (9)	Alto (12)	Alto (15)
2.- Es raro que suceda	Bajo (2)	Bajo (4)	Medio (6)	Medio (8)	Alto (10)
1.- Sería excepcional	Bajo (1)	Bajo (2)	Bajo (3)	Bajo (4)	Medio (5)

Una vez visto la metodología que se va a utilizar, en el siguiente capítulo se realizará el desarrollo y la verificación del proyecto. Para ello en cada fase se tendrán que identificar, analizar, planificar y supervisar los riesgos.

6 Aplicación de la metodología y resultados obtenidos

A lo largo de este capítulo se desarrollará la metodología elegida, desarrollo en espiral, y se aplicarán las actividades correspondientes a lo largo de cuatro iteraciones.

6.1 Introducción

Este proyecto consistirá en analizar datos de la red social de Twitter sobre las universidades y centros asociados de la Comunidad de Madrid. Se analizará de qué hablan las universidades, de qué hablan los usuarios cuando hablan de las universidades, desde dónde se habla de las universidades, qué usuarios tienen más influencia y qué universidad está mejor valorada.

Para llevar a cabo el proyecto se ha dividido principalmente en dos partes principales. La primera referente a la adquisición de datos, y la segunda a una herramienta analítica preparada para el descubrimiento de información.

En la parte de adquisición de datos se utilizará las APIs de Twitter, tanto Streaming API como Rest API y se desarrollará una plataforma de desarrollo, con diferentes arquitecturas, necesaria para ejecutar cada API.

Para desarrollar la parte analítica se ha elegido la herramienta Oracle Endeca Information Discovery. Aunque no existen muchas herramientas en el mercado y menos OpenSource con semejantes características, uno de los principales motivos de elegir esta herramienta es que Oracle proporciona todos los componentes necesarios para el estudio, desde el sistema operativo, base de datos, servidor de aplicaciones, herramienta analítica, enriquecimiento de texto, etc.

A lo largo de este capítulo se irá siguiendo la metodología de desarrollo en espiral, explicada en el capítulo 5. Las principales actividades de cada fase serán:

- Determinar los objetivos
- Análisis de riesgos
- Desarrollo y verificación
- Planificación

6.2 Fase 1: Adquisición de datos en la nube e integración de componentes

Para diseñar esta solución, lo primero que necesitamos, es ver qué datos se pueden adquirir, la calidad de los mismos así como su cantidad. Si se quiere realizar un estudio de las diferentes universidades de Madrid, se necesitará fijar unos términos de búsqueda, descargar datos y contrastar los resultados.

Por otra parte, en la primera fase o iteración del proyecto se va a integrar todos los componentes y preparar un entorno operativo con el que se pueda empezar a visualizar estos datos adquiridos.

Esta fase comenzó el 6 de enero de 2014 y finalizó el 16 de febrero de 2014.

6.2.1 Determinar objetivos

La primera actividad según el desarrollo en espiral, es determinar los objetivos para esta primera fase.

6.2.1.1 Adquisición de datos

Para adquirir los datos que se van a necesitar se ha utilizado la Streaming API de Twitter, esta API permite tener una conexión abierta con Twitter y recibir parte de estos datos en nuestro sistema.

Figura 25. Arquitectura Streaming API de Twitter. Fuente: elaboración propia.

Se puede observar en la figura 25, que desde una instancia siempre activa se mantendrá la comunicación con el servidor de Twitter, para que cada vez que un usuario publique un tuit este llegue directamente al servidor en formato JSON.

Este fichero JSON se tratará y se guardará en una base de datos desde la que más tarde se accederá y se recuperará la información guardada.

6.2.1.2 Integración de componentes

Se ha optado por utilizar Oracle Endeca como plataforma de búsqueda de relaciones debido a sus buenos resultados integrando datos no estructurados con datos estructurados.

Figura 26. Esquema de integración de Oracle Endeca con Amazon RDS. Fuente: elaboración propia.

El sistema requerirá principalmente de dos componentes:

- **Oracle Endeca Server:** es la base de datos de búsqueda analítica que integra todos los datos del dominio de datos. Este componente se tendrá que comunicar con la base de datos que contiene los datos semiestructurados.
- **Oracle Endeca Studio:** permite acceder y explorar los datos almacenados en el servidor.

6.2.2 Análisis de riesgos

Una vez estén fijados los objetivos de esta fase se procede al estudio de los riesgos implicados. Este estudio se divide en cuatro fases.

6.2.2.1 Identificación de riesgos

Se han identificado los siguientes riesgos:

Tabla 10. Tabla de identificación de riesgos de la fase 1. Fuente: elaboración propia.

Identificador	Riesgo	Tipo	Descripción
RIESGO-F01-01	Pérdida de datos	Proyecto	Pérdida de datos, ya sea por corrupción de los mismos o por no poder acceder al sistema remoto
RIESGO-F01-02	Exceso de transferencia de datos	Producto y proyecto	Al utilizar un sistema gratuito, este se encuentra limitado en cuanto a la transferencia de datos
RIESGO-F01-03	Cantidad de datos adquiridos	Proyecto	No tener suficientes datos para realizar el estudio
RIESGO-F01-04	Rendimiento del sistema integrado	Producto	Capacidades de hardware

RIESGO-F01-01: Pérdida de datos

Como los datos estarán guardados en un servicio gratuito proporcionado por Amazon, se identifica un riesgo (RIESGO-F01-01) relacionado con la pérdida de datos, ya sea por la corrupción de los datos o por no poder acceder al sistema remoto.

RIESGO-F01-02: Exceso de transferencia de datos

Al utilizar Amazon EC2, la transferencia de datos está limitada a 15 GB (2) de ancho de banda saliente. Además al utilizar un servicio en la nube podría ocurrir un cambio de las condiciones de servicio por lo que se perdería el acceso al sistema que mantiene la conexión abierta con Twitter lo que supondría un retraso en el proyecto. Este riesgo se identifica como RIESGO-F01-02.

RIESGO-F01-03: Cantidad de datos adquiridos

Para poder realizar el estudio es importante tener una cantidad de datos suficientemente grande y variada, es decir, poder tener acceso a la información de las diferentes universidades de Madrid. Por tanto han de tener suficiente repercusión en Twitter. Existen webs como Topsy (42) que dado un término de búsqueda, analizan cuántos tuits hay. Se van a analizar 16 universidades (43), mediante 94 términos de búsqueda. Esto generará aproximadamente 1,25 millones tuits en 6 meses (ver Anexo I), teniendo en cuenta que las APIs de Twitter están limitadas (aunque no siempre documentada) o por el número de consultas o por la cantidad de datos enviados al cliente, un objetivo de un 20% de todos los datos es suficiente para realizar el estudio, por tanto el objetivo mediante la Streaming API será adquirir 250 mil tuits en 6 meses.

RIESGO-F01-04: Rendimiento del sistema integrado

Por último, uno de los objetivos principales de la primera fase es la integración de todos los componentes. Hay que tener en cuenta que Oracle Endeca es una herramienta empresarial que normalmente se instala en potentes ordenadores en los centros de procesamiento de datos empresariales. Al querer instalar estos componentes en una máquina virtual, se identifica un riesgo (RIESGO-F01-04) de capacidades de hardware.

Para cada uno de estos datos se analizará la probabilidad de que ocurra.

6.2.2.2 Análisis de riesgos

Una vez se han identificado los posibles riesgos de esta fase del proyecto se procede al análisis de los mismos.

Tabla 11. Análisis de riesgos fase 1. Fuente: elaboración propia.

Identificador	Nombre	Probabilidad
RIESGO-F01-01	Pérdida de datos	Muy bajo
RIESGO-F01-02	Exceso de transferencia de datos	Moderado
RIESGO-F01-03	Cantidad de datos adquiridos	Bajo
RIESGO-F01-04	Rendimiento del sistema integrado	Muy bajo

RIESGO-F01-01: Pérdida de datos

Se ha considerado que la probabilidad de que ocurra la pérdida de datos es muy baja ya que, realizando backups, en el peor de los casos habría una pérdida de los datos de 1 o 2 días.

RIESGO-F01-02: Exceso de transferencia de datos

Realizar un estudio previo de la transferencia de datos generada entre la instancia de Amazon y Twitter es muy difícil. Cuando se utiliza la Streaming API de Twitter existe un límite de 15 GB al mes (2), que equivale a 6 kB al segundo, es decir, unos 6.000 caracteres al segundo de tráfico saliente.

Además siempre que se trabaja con un sistema no propietario existe un riesgo de perder esa relación con el proveedor por un cambio en las condiciones de uso. Se ha calificado el riesgo como moderado.

RIESGO-F01-03: Cantidad de datos adquiridos

Aunque la fase 1 dure 6 semanas, las primeras 3 se dedicarán a la preparación del entorno, por tanto en 3 semanas se espera adquirir la cantidad de datos objetivo. Esta cantidad será el objetivo de 250 mil tuits entre el número de semanas que falten para terminar, es decir, 11.363 a la semana.

RIESGO-F01-04: Rendimiento del sistema integrado

La integración de los componentes formará la estructura del sistema y aunque es un riesgo muy importante la probabilidad de que ocurra es muy baja debido a que, en caso de verse limitadas las capacidades hardware, se puede proceder a una ampliación de las mismas.

6.2.2.3 Planificación de riesgos

Una vez analizados los riesgos procedemos a la planificación para su supervisión.

RIESGO-F01-01: Pérdida de datos

- **Estrategia de prevención:** se realizarán backups semanales de la base de datos en el equipo local.
- **Plan de contingencia:** en caso que ocurra el riesgo identificado se procederá a la restauración del último backup guardado.

RIESGO-F01-02: Exceso de transferencia de datos

- **Estrategia de prevención:** se procederá a una monitorización del tráfico saliente generado.
- **Plan de contingencia:** en caso de no poder utilizar el sistema de Amazon, no se podrá mantener una conexión constante con Twitter. Por tanto se tendrá que cambiar el modo operando para la adquisición de datos. Se procederá a utilizar la Rest API para adquirir datos bajo demanda.

RIESGO-F01-03: Cantidad de datos adquiridos

- **Estrategia de prevención:** se monitorizará la cantidad de datos adquiridos y se estudiará un valor semanal de datos a adquirir.
- **Plan de contingencia:** en caso de no poder conseguir suficientes datos se estudiará el caso y en caso de necesitar más mediante la Rest API se podrá adquirir más datos, siempre y cuando existan suficientes datos.

RIESGO-F01-04: Rendimiento del sistema integrado

- **Estrategia de prevención:** mediante el gestor de tareas de Windows se podrá monitorizar las capacidades consumidas por el sistema. Se pueden eliminar procesos que no sean esenciales para la ejecución de la máquina.
- **Plan de contingencia:** en caso de verse el hardware totalmente insuficiente se procederá a una ampliación del mismo.

6.2.2.4 Supervisión de riesgos

Una vez planificados los riesgos se procede a explicar los métodos de supervisión.

RIESGO-F01-01: Pérdida de datos

Mediante la tabla 12 se mantendrá un control de los backups realizados semanalmente (-1 backup no realizado, 1 backup realizado, 0 backup pendiente de realizar).

Tabla 12. Supervisión de riesgos fase 1. Fuente: elaboración propia.

RIESGO-F01-01	
Semana	Backup realizado
4	0
5	0
6	0
7	0
8	0
9	0

RIESGO-F01-02: Exceso de transferencia de datos

Mediante la tabla 13, será monitorizado el tráfico generado, el tráfico acumulado y el límite semanal teórico de tráfico saliente. Esta monitorización se realizará semanalmente. El límite de tráfico mensual se ha repartido entre las semanas del mes y se ha establecido en 3,8 GB a la semana.

Tabla 13. Exceso de transferencia de datos. Fuente: elaboración propia.

RIESGO-F01-02			
Semana	Tráfico generado (MB)	Tráfico acumulado (MB)	Límite de tráfico (MB)
4	0	0	3.840
5	0	0	3.840
6	0	0	3.840
7	0	0	3.840
8	0	0	3.840
9	0	0	3.840

La figura 27 muestra mediante un gráfico los parámetros de la tabla 13.

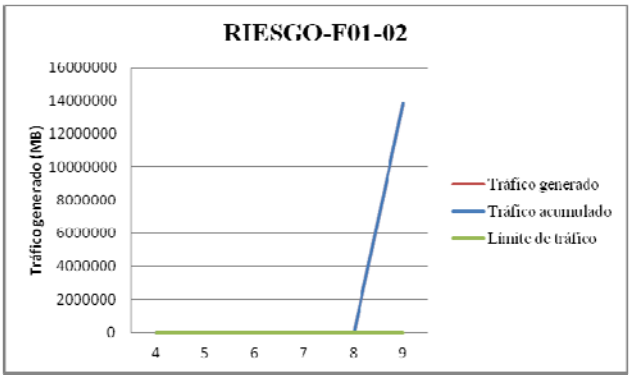


Figura 27. RIESGO-F01-02. Fuente: elaboración propia.

RIESGO-F01-03: Cantidad de datos adquiridos

Mediante una tabla se mantendrá monitorizado la cantidad de datos adquiridos semanalmente, así como el objetivo semanal, la cantidad de datos adquiridos en total y el objetivo teórico total de cada semana. Además se mostrará el porcentaje total real, es decir, el porcentaje de datos adquiridos del objetivo de 250 mil tuits frente al porcentaje total teórico, es decir, el porcentaje esperado de tuits adquiridos.

Tabla 14. RIESGO-F01-03. Fuente: elaboración propia.

RIESGO-F01-03						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
4	-	-	-	-	0%	0%
5	-	-	-	-	0%	0%
6	-	-	-	-	0%	0%
7	-	11.363	-	11.363	0%	5%
8	-	11.363	-	22.726	0%	9%
9	-	11.363	-	34.089	0%	14%
TOTAL	-	53.571	-	250.000		

La figura 28 muestra mediante un gráfico de barras la cantidad de tuits adquiridos frente al objetivo semanal. En el eje X se muestra el número de semana de la fase y en el eje Y la cantidad de tuits.

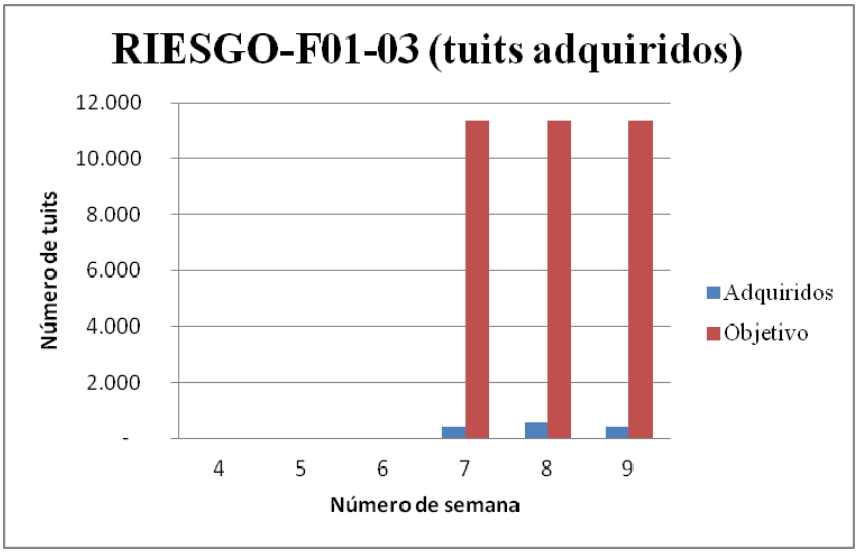


Figura 28. RIESGO-F01-03 (tuits adquiridos). Fuente: elaboración propia.

En la figura 29 se muestra la relación entre el porcentaje real de datos adquiridos frente al ideal de datos que se debería tener.

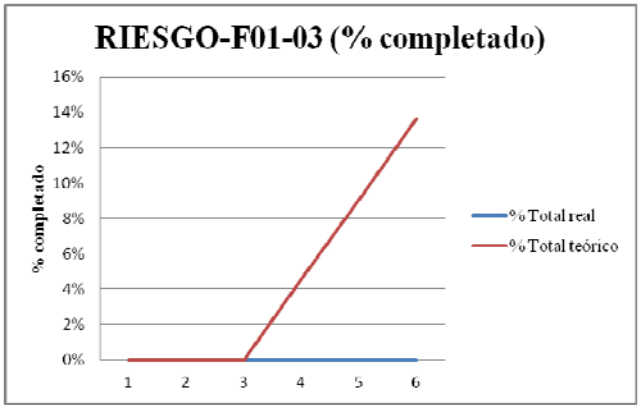


Figura 29. RIESGO-F01-03 (% completado). Fuente: elaboración propia.

RIESGO-F01-04: Rendimiento del sistema integrado

Para supervisar este riesgo procederemos, mediante el administrador de tareas, a observar el uso de la CPU, el uso memoria física así como la paginación del sistema. Se sabrá que estamos llegando al límite de las capacidades del ordenador cuando la interactuación con la máquina virtual sea demasiado lenta, se bloquee o no se pueda trabajar.

Una vez se ha desarrollado el análisis de riesgos se procede al desarrollo y verificación de los objetivos de esta fase.

6.2.3 Desarrollo y verificación

Esta fase se divide en dos desarrollos principales. La primera es la adquisición de datos, con actividades como la preparación del entorno y configuración del framework entre otras, y la segunda la preparación del entorno de desarrollo.

6.2.3.1 Adquisición de datos

6.2.3.1.1 Preparación del entorno

Como se ha mencionado anteriormente se ha utilizado la capa gratuita Amazon Web Services (2) que ofrece una plataforma en la nube flexible y gratuita. Se ha utilizado una microinstancia de Amazon Elastic Compute Cloud para desplegar un servidor HTTP, en el que se ejecutará código php para mantener la conexión abierta con la Streaming API de Twitter. Por otra parte también se ha utilizado el servicio Amazon Relational Database Service que permite alojar una base de datos relacional de hasta 20 GB de almacenamiento (2), en donde se guardarán los tuits descargados.

6.2.3.1.1.1 Amazon Elastic Compute Cloud (Amazon EC2)

La microinstancia proporcionada por Amazon es de 1 procesador y 613 MB de memoria (2), sobre la que se ha instalado:

- Sistema Operativo: Amazon Linux 2013.09 (44)

  Se ha optado por esta distribución ya que está desarrollada y optimizada por Amazon para trabajar con microinstancias con poca memoria RAM y con poca capacidad de

procesamiento. Esta distribución viene sin entorno gráfico por tanto se ha operado siempre por consola remota mediante la aplicación putty 0.63 (45).
Para aumentar la seguridad a la hora de conectarse a la instancia levantada, Amazon proporciona unas claves de seguridad únicas que se generan desde el panel de control. Estas claves utilizan un certificado digital X.509 v3 con codificación en Base64 (46). Para acceder a la instancia desde cualquier cliente SSH, Amazon solicitará estas claves además de las propias del sistema operativo.

- Servidor HTTP: Apache HTTP Server 2.4.7 (47)

  Tanto para mantener una conexión abierta con Twitter mediante la Streaming API como para poder acceder a la base de datos, se necesita un servidor http en el que se puedan ejecutar ficheros php.
  - Administración de la base de datos: phpMyAdmin 4.0.10 (48)

    PhpMyAdmin es un software gratuito muy utilizado, escrito en PHP y con interfaz gráfica, que permite de una manera sencilla administrar una base de datos MySQL.
  - Conexión Streaming API: se ha utilizado un framework gratuito proporcionado por 140dev (49) bajo la licencia GPL. Está escrito en PHP y permite una conexión a una base de datos MySQL. El componente principal es Twitter Database Server que se explicará un poco más en profundidad en el apartado Framework 140dev de este capítulo.

6.2.3.1.1.2 Amazon Relational Database Service (Amazon RDS)

Es un servicio web que facilita configurar una base de datos ya sea MySQL, PostgreSQL, Oracle o SQL Server, administrarla y realizar operaciones en ella. Mediante el panel de control de Amazon se pueden programar backups y restaurarlos o crear nuevas bases de datos. Esto permite centrarse en la aplicación a desarrollar y no en la gestión de la base de datos.

Se ha optado por utilizar una base de datos MySQL 5.6.13, con 5 GB de capacidad, ya que es ligera y compatible con el resto de los componentes que se van a utilizar.

6.2.3.1.1.3 Streaming API Twitter v1.1

Para poder tener acceso a esta API solo hay que crearse una cuenta de desarrolladores (1) y crear una aplicación. Esto generará unas claves de acceso con las que se podrá autentificarse en la aplicación creada. Los datos necesarios para conectarse a la aplicación serán los siguientes:

- Configuración de la aplicación
  - API key y API secret (1): pareja que identifica la aplicación
- Token de acceso
  - Access token y Access token secret (1): pareja que identifica al usuario

Estos cuatro parámetros permiten tener una conexión segura y proporcionar a los usuarios un acceso a sus datos al mismo tiempo que protegen los credenciales de su cuenta mediante la identificación OAuth[1].

Figura 30. Identificación OAuth. Fuente: elaboración propia.

6.2.3.1.1.4 Framework 140dev

El propósito de este framework es recoger los tuits de la Streaming API de Twitter y distribuirlos en las tablas de la base de datos (49).

6.2.3.1.1.4.1 Arquitectura

A la hora de diseñar la arquitectura es importante tener en cuenta que la respuesta de la Streaming API es en formato JSON por tanto, la extracción de tuits tendrá que realizarse en dos pasos. El primero mantendrá la conexión abierta con el servidor recibiendo los tuits en formato JSON y los almacenará en una tabla. El segundo paso será leer de esa tabla en formato JSON y estructurarlos y distribuirlos por las demás tablas según un modelo de datos definido.

Además es una buena práctica separar los procesos en dos ya que si todo se hiciese con un proceso, podría darse el caso de estar distribuyendo el JSON en las tablas y llegase un tuit desde el servidor, ocurriría que no sería recibido y se perdería esa información.

[1] OAuth es un protocolo abierto que permite la identificación segura de una API

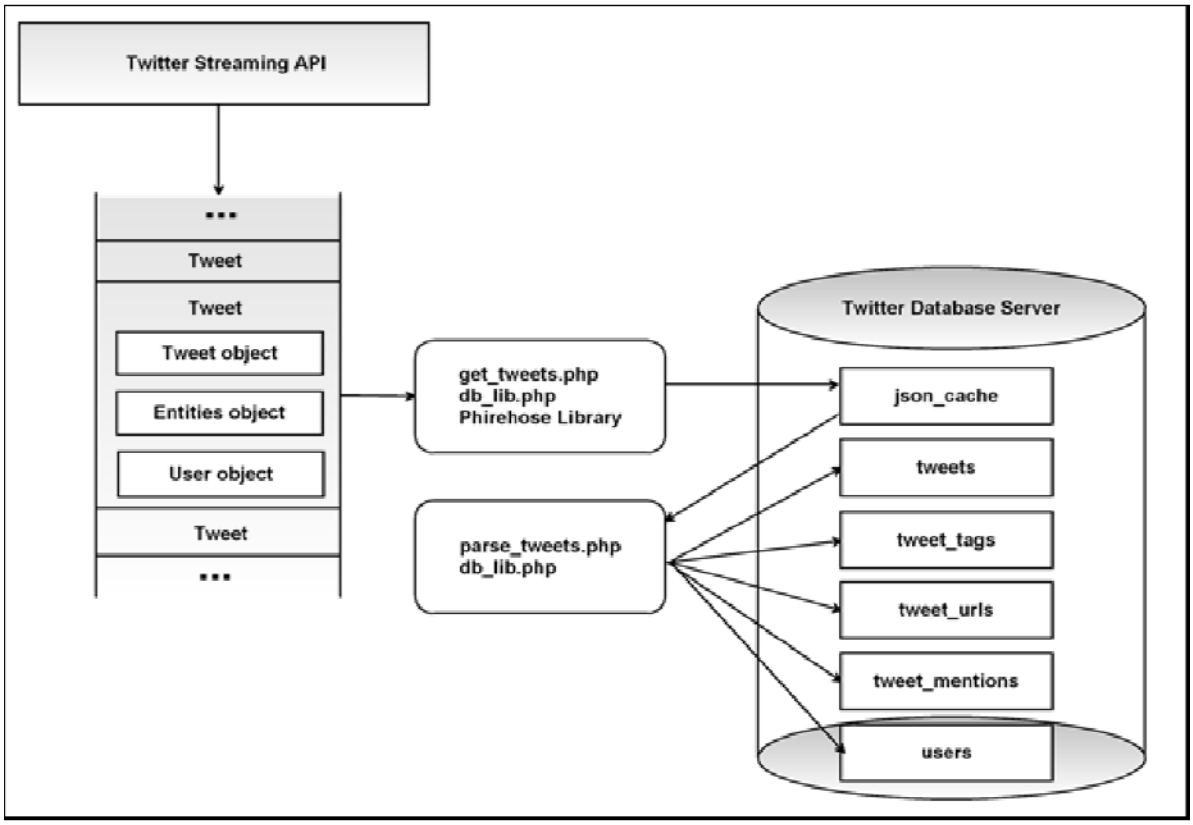


Figura 31. Arquitectura funcionamiento servidor de base de datos Streaming API. Fuente: (50)

Para ello el primer proceso se hace mediante el proceso "get_tweets.php" el cual está ejecutándose continuamente a la espera de nuevos tuits. Cuando un tuit es recibido se guarda en la tabla "json_cache". Más tarde el proceso "parse_tweets.php" recoge ese tuit y lo divide en las tablas correspondientes, es decir, tabla de usuarios, tags, urls, tuits y menciones.

6.2.3.1.1.4.2 Esquema base de datos

Como se ha visto anteriormente, el proceso "parse_tweets.php" es el encargado de leer de la tabla "json_cache" y distribuirlo en las tablas. En total se utilizan 6 tablas siguiendo el esquema de la figura 32.

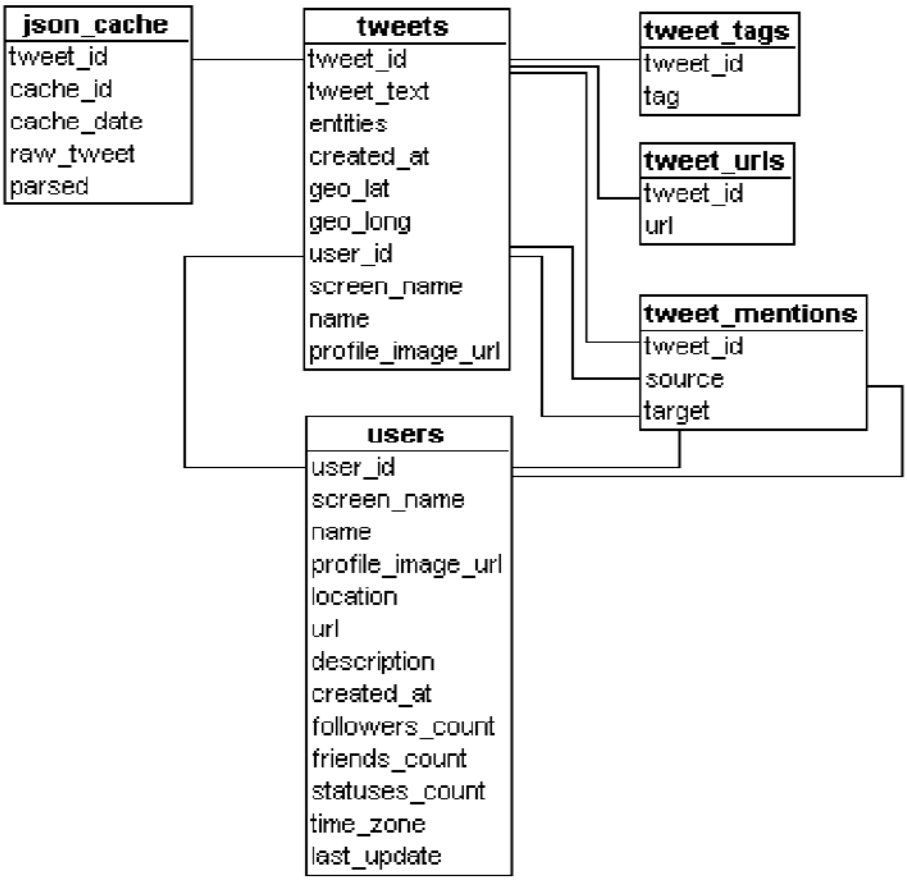


Figura 32. Esquema tablas base de datos Streaming API. Fuente: (51)

La descripción de las tablas es la siguiente:

- **Tabla "json_cache":** guarda los json que recibe el proceso "get_tweets.php"
- **Tabla "tweets"**: es la tabla principal y guarda desde el propio comentario, su localización, id de usuario, la fecha de creación, nombre y entidades entre otros atributos.
- **Tabla "tweets_tags"**: esta tabla mantiene una relación entre el id del tuit y los tags relacionados.
- **Tabla "users":** la tabla usuarios guarda desde el nombre del usuario, url de la imagen de perfil, número de seguidores, de amigos y la localización entre otros.
- **Tabla "tweets_urls":** mantiene una relación entre el id del tuit y la url del mismo.
- **Tabla "tweets_mentions":** esta tabla guarda la guarda la fuente y el destino cuando en tuit se habla de otro usuario.

6.2.3.1.1.5 Configuración

Es necesario configurar el framework para que trabaje con la API que está creada y la base de datos de Amazon RDS.

```
$db_host = 'dbinputdata.cpfmehotvp9p.eu-west-1.rds.amazonaws.com';
$db_user = 'root';
$db_password = 'oracle123';
$db_name = 'dbtwitter_es';
```

Configuración de acceso a la base de datos

```
define('TWITTER_CONSUMER_KEY','KIlp6T7N4qxDSOhajGv0g');
define('TWITTER_CONSUMER_SECRET','btM3BoupmTD78wM9fFHKZBsCP6ue9rayBgDEYYIg');
define('OAUTH_TOKEN','714812628-LKvjPNpJ6lGI1CbRnBbaAGcWZtNdCiEQIdZ4HsDD');
define('OAUTH_SECRET','eR6DmnmndRDHaVOVz4RZpQwWyDxOG92AnVSN5ACCdmNMV');
```

Otro parámetro necesario es definir los términos de búsqueda en el proceso “get_tuits.php”, estos términos vienen especificados en el Anexo I.

6.2.3.1.2 Ejecución y verificación

Una vez se tienen todos los componentes preparados para la adquisición de datos se accede por putty (45) a la consola del servidor y se ejecuta los procesos por orden.

1. “get_tweets.php”
2. “parse_tweets.php”

Si se accede a la aplicación phpMyAdmin de la instancia en EC2 desde el navegador se puede observar que los resultados son inmediatos. En el apartado “Resultados” de esta actividad se mostrarán las conclusiones obtenidas.

6.2.3.2 Integración componentes Oracle Endeca

En esta fase se preparará el entorno de desarrollo para integrar los componentes necesarios del sistema. Para ello se configurará una máquina virtual en Oracle VM Virtual Box 4.3.6 con 2 CPUs y 6 GB de memoria RAM.

6.2.3.2.1 Oracle Enterprise Linux 6.5

El sistema operativo elegido es Oracle Enterprise Linux debido a su capacidad de integración con los demás componentes de Oracle. A la hora de instalarse, se ha tenido en cuenta activar el modo escritorio para poder tener una interfaz gráfica desde la que se pueda interactuar fácilmente con el sistema. En el Anexo III se detalla la información para su instalación así como los datos de acceso.

6.2.3.2.2 Oracle Weblogic Application Server 10.3.6

Todos los componentes de Oracle Endeca se instalan sobre un servidor de aplicaciones, que en este caso es Oracle Weblogic. Se ha elegido este servidor de aplicaciones por la facilidad de

integración con otros componentes de Oracle. Se ha creado un único servidor llamado AdminServer que contendrá todos los dominios necesarios de Oracle Endeca.

Además es necesario instalar Oracle Application Development Framework (Oracle ADF) que permite simplificar el desarrollo de aplicaciones proporcionando una interfaz visual.

En el Anexo III se detalla la información para la instalación de Oracle Weblogic y Oracle ADF, así como los datos de acceso.

6.2.3.2.3 Oracle Database 11gR2

Los dominios necesarios para desplegar el sistema Oracle Endeca necesitan una base de datos para guardar datos de configuración y temporales. Por defecto, Oracle Endeca utiliza Hypersonic Database, una base de datos relacional. Como se necesitará acceder a la base de datos para guardar datos de los usuarios, se ha procedido a cambiarla por Oracle Database 11gR2. Esta base de datos es completamente compatible con el resto de componentes de Oracle además que mejora el rendimiento de la máquina de virtual.

En el Anexo III se detalla la información para la instalación de Oracle Database, así como los datos de acceso y configuración.

6.2.3.2.4 Oracle Endeca Server 7.6.1

Es el núcleo de Oracle Endeca, una base de datos analítica, instalada en un dominio de Weblogic.

En el Anexo III se detalla la información para la instalación de Oracle Endeca Server, así como los datos de acceso y configuración.

6.2.3.2.5 Oracle Endeca Studio 3.1

Es la aplicación que conecta a Oracle Endeca Server, por la cual el usuario accede al sistema. Instalada en otro dominio de aplicaciones de Oracle Weblogic.

En el Anexo III se detalla la información para la instalación de Oracle Endeca Studio, así como los datos de acceso y configuración.

6.2.3.2.6 Oracle Endeca Provisioning Service 3.1

Este servicio permite a los usuarios subir sus propios ficheros Excel al servidor de Oracle Endeca, para poder analizar sus datos con los del servidor. Está instalado en otro dominio del servidor de aplicaciones Oracle Weblogic.

En el Anexo III se detalla la información para la instalación de Oracle Endeca Provisioning Service, así como los datos de acceso y configuración.

6.2.3.2.7 Ejecución

El tiempo de instalación y configuración de los componentes para un usuario experto, con conocimientos en software Oracle, es de 1-2 días y para un usuario intermedio 1-2 semanas.

Una vez se tienen todos componentes instalados y configurados se puede proceder a levantar los servicios. El orden de ejecución es el siguiente:

1. Oracle Enterprise Linux 6.5
2. Oracle Database 11gR2
3. Oracle Endeca Server 7.6.1
4. Oracle Endeca Studio 3.1
5. Oracle Endeca Provisioning Service 3.1

Para una rápida ejecución se han creado unos scripts de arranque automáticos que vienen documentados en el Anexo III al igual que los detalles de acceso e instalación del sistema.

El tiempo necesario para arrancar todos los componentes es de 30 minutos aproximadamente.

6.2.3.3 Resultados

Los objetivos definidos para esta fase eran la adquisición de datos de Twitter y una primera aproximación a su visualización. En la primera parte de este desarrollo se ha conseguido, mediante la Streaming API adquirir datos y guardarlos en una base de datos (Amazon RDS). En la segunda parte del desarrollo se han integrado los principales componentes de Oracle Endeca.

6.2.3.3.1 Adquisición de datos

Al poco tiempo de ejecutar los procesos de conexión y tratamiento de tuits, el sistema empieza a descargar datos. La tabla "json_cache" es cargada con los datos en bruto y vaciada periódicamente por el proceso "parse_tweets.php".

Después de 3 semanas descargando tuits, se consiguen 1.426 tuits, como se puede ver en la Figura 33. Este número parece poco pero hay que tener en cuenta que la Streaming API limita el envío de tuits.

Estructura | SQL | Buscar | Generar una consulta | Exportar | Importar | Operaciones | Privilegios

	Tabla	Acción	Filas	Tipo	Cotejamiento	Tamaño	R
☐	json_cache	Examinar Estructura Buscar Insertar Vaciar Eliminar	0	MyISAM	utf8_general_ci	1 KB	
☐	tweets	Examinar Estructura Buscar Insertar Vaciar Eliminar	1,426	MyISAM	utf8_general_ci	736.6 KB	
☐	tweet_mentions	Examinar Estructura Buscar Insertar Vaciar Eliminar	465	MyISAM	latin1_swedish_ci	36.4 KB	
☐	tweet_tags	Examinar Estructura Buscar Insertar Vaciar Eliminar	730	MyISAM	utf8_general_ci	39.1 KB	
☐	tweet_urls	Examinar Estructura Buscar Insertar Vaciar Eliminar	560	MyISAM	utf8_general_ci	46.9 KB	
☐	users	Examinar Estructura Buscar Insertar Vaciar Eliminar	791	MyISAM	utf8_general_ci	363.8 KB	
	6 tablas	Número de filas	3,972	InnoDB	latin1_swedish_ci	1.2 MB	

Figura 33. Resultados fase 1 en phpMyAdmin. Fuente: elaboración propia.

Aunque esa limitación no está documentada, normalmente se habla de en torno a un 1%, es decir, solo llegan al sistema un 1% de todos los tuits que se generan.

Para esta fase se han configurado 98 términos de búsqueda, especificados en el Anexo I, en la llamada a la API (mediante el proceso "get_tweets.php"). Estos términos han generado según Topsy (42) durante el 14 de enero y 14 de febrero de 2014 un total de 192.861 tuits.

Según estos datos, tan solo el 0,7% de los tuis generados por los usuarios con los términos de búsqueda definidos en el Anexo I han sido enviados por la Streaming API de Twitter y guardados en el almacenamiento en la nube.

6.2.3.3.2 Integración de componentes

Después de instalar y configurar todo el sistema, se procede a arrancarlo mediante el script creado para este propósito "startAll.sh". Posteriormente se abre el navegador y se accede a Oracle Studio mediante el enlace http://oeid:31001/eid/web/home.

Se configuran los accesos a la base de datos MySQL, de Amazon RDS, mediante la cadena de conexión jdbc.

jdbc::mysql://dbinputdata.cpfmehotvp9p.eu-west-1.rds.amazonaws.com:3306/dbtwitter

Y se crean tantas conexiones como tablas en la base de datos.

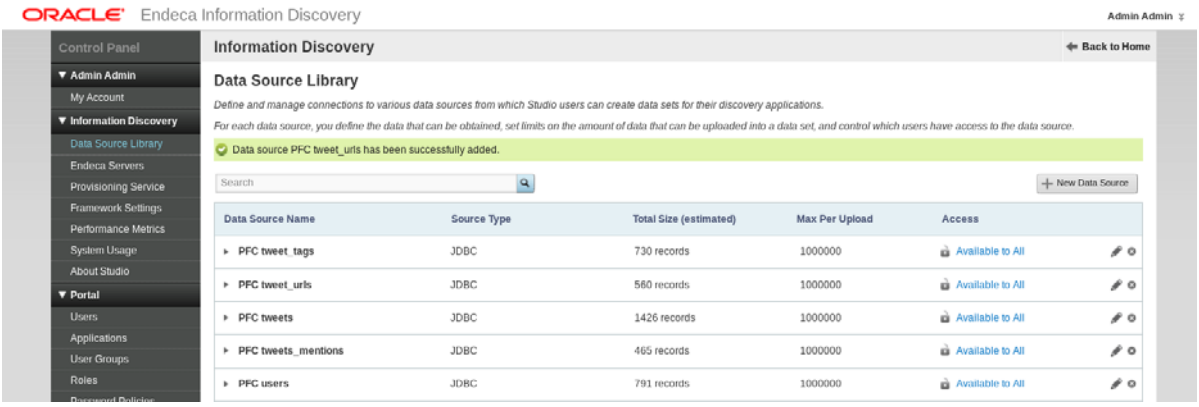


Figura 34. Fuentes de datos en Oracle Endeca Studio. Fuente: elaboración propia.

Una vez creadas las conexiones se procede a crear la aplicación y se integran las fuentes de datos (figura 34).

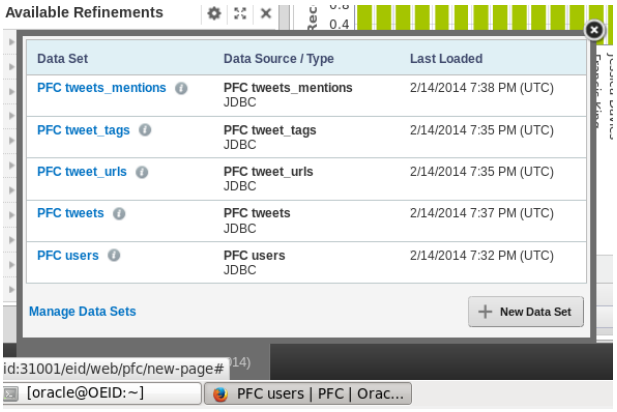


Figura 35. Tablas cargadas en Oracle Endeca Studio en la fase 1. Fuente: elaboración propia.

Se crearán paneles sencillos para visualizar los datos cargados.

6.2.3.3.3 Visualización de los datos

El objetivo de esta fase es mostrar los primeros resultados, cargándolos en Endeca Server. Para ello los cuadros de mando generados son los siguientes.

Pestaña usuarios (PFC users)

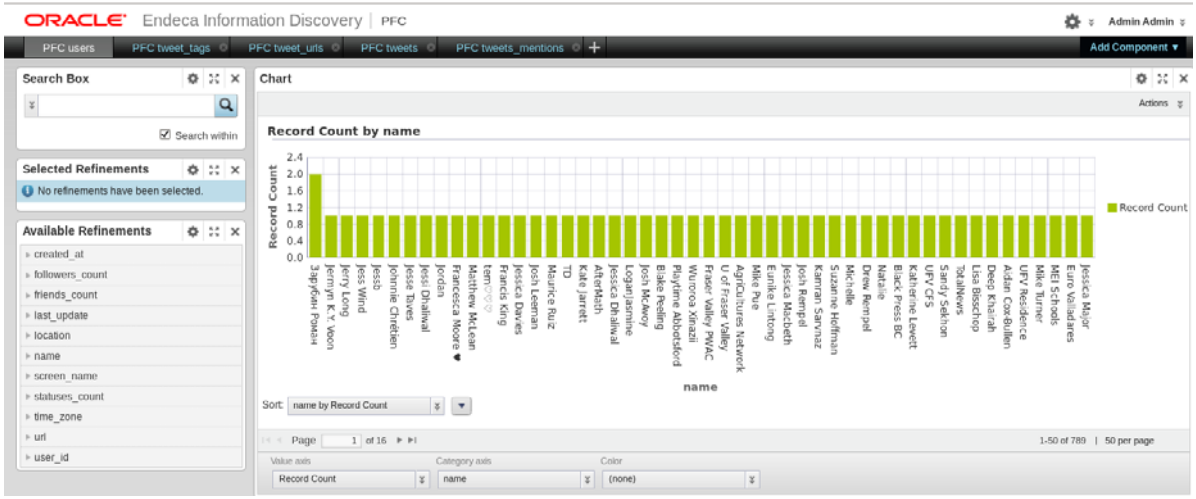


Figura 36. Pestaña PFC users de la fase 1. Fuente: elaboración propia.

Mediante una gráfica se muestran los usuarios de la tabla usuarios ordenados descendentemente por el número de tuits que han generado. En la figura 36 se puede observar que la mayoría de los usuarios solo tienen un registro, esto quiere decir que la tabla usuarios solo tiene una entrada por usuario, es decir, el framework de acceso a la API de Twitter comprueba si existe un usuario en la tabla y en caso de existir lo actualiza, manteniendo, en la mayoría de los casos, la tabla de usuarios como registros únicos sin duplicidad.

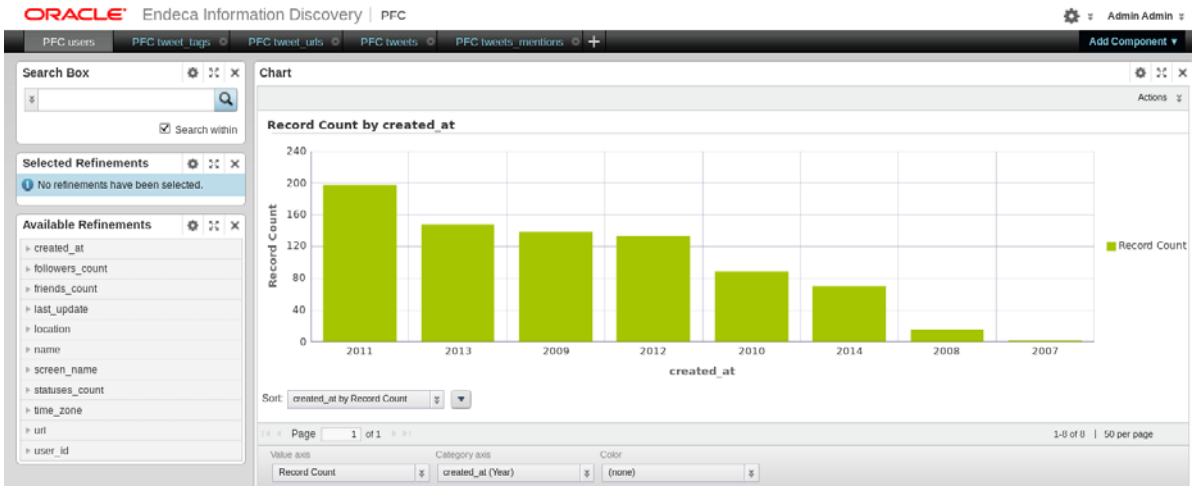


Figura 37. Gráfico creación tuits en Oracle Endeca Studio. Fuente: elaboración propia.

En la figura 37 se ha mostrado el año de creación de la cuenta de Twitter de los usuarios. Se puede observar que de los 789 usuarios cargados, casi 200 cuentas están creadas en el año 2011.

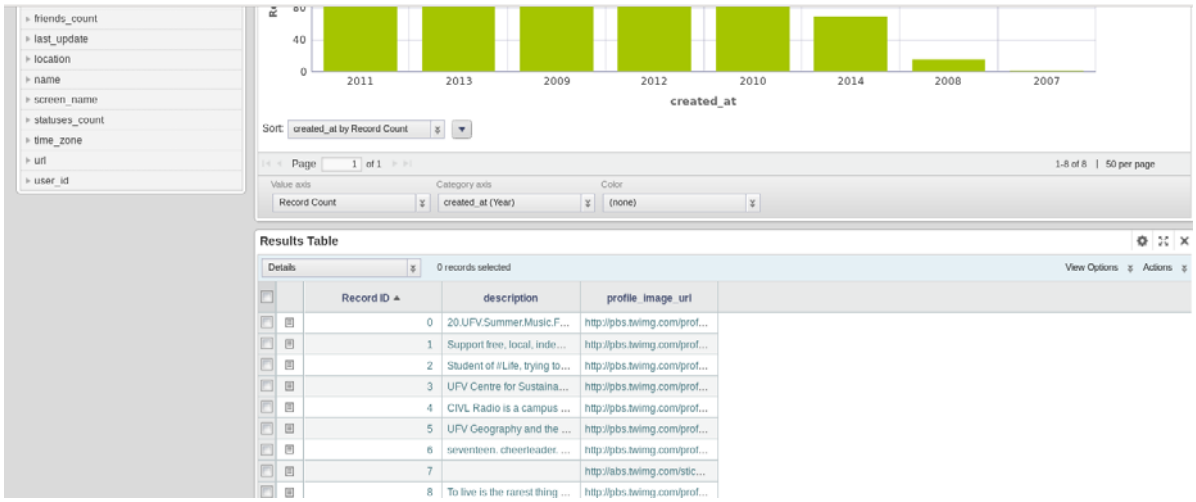


Figura 38. Tabla detalles en Oracle Endeca Studio. Fuente: elaboración propia.

Por último, en la figura 38 se muestra tabla con la descripción de los usuarios y el enlace a la imagen del perfil.

Pestaña Tag (PFC tweet_tags)

En esta pestaña se muestra la tabla "tweet_tags" de la base de datos. Se puede observar en la Figura 39 que la mayoría de los tuits tienen como etiqueta UFV. En la siguiente fase habrá que mejorar estos resultados ya que resulta extraño que sea esa la etiqueta más frecuente.

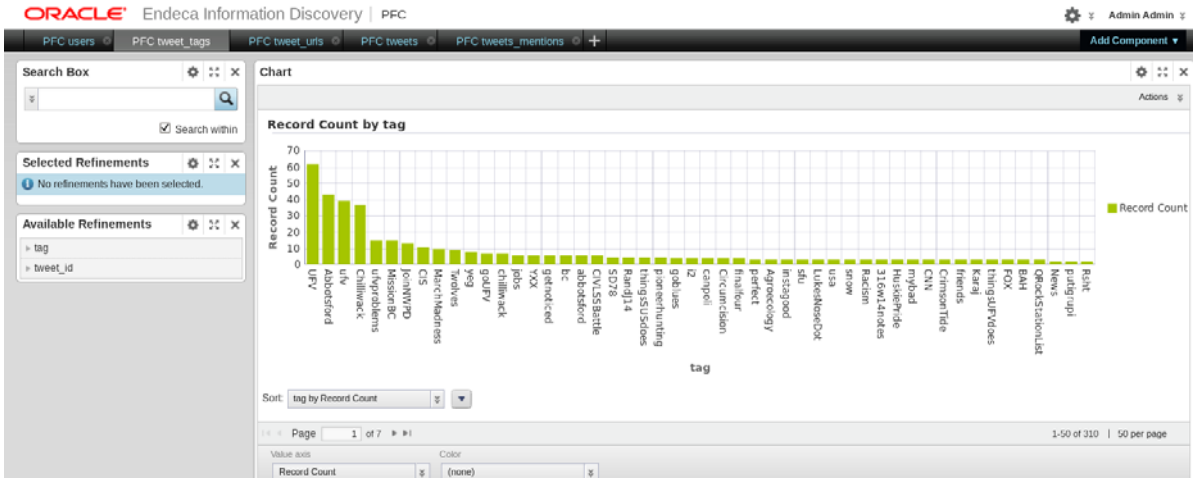


Figura 39. Pestaña PFC tweet_tags en Oracle Endeca Studio. Fuente: elaboración propia.

Pestaña tuits URL (PFC tweet_urls)

Esta pestaña muestra la tabla "tweet_urls" de la base de datos. A primera vista no aporta mucha información excepto que hay 90 tuits que vienen de una misma URL (figura 40). En la siguiente fase se mejorarán estos resultados.

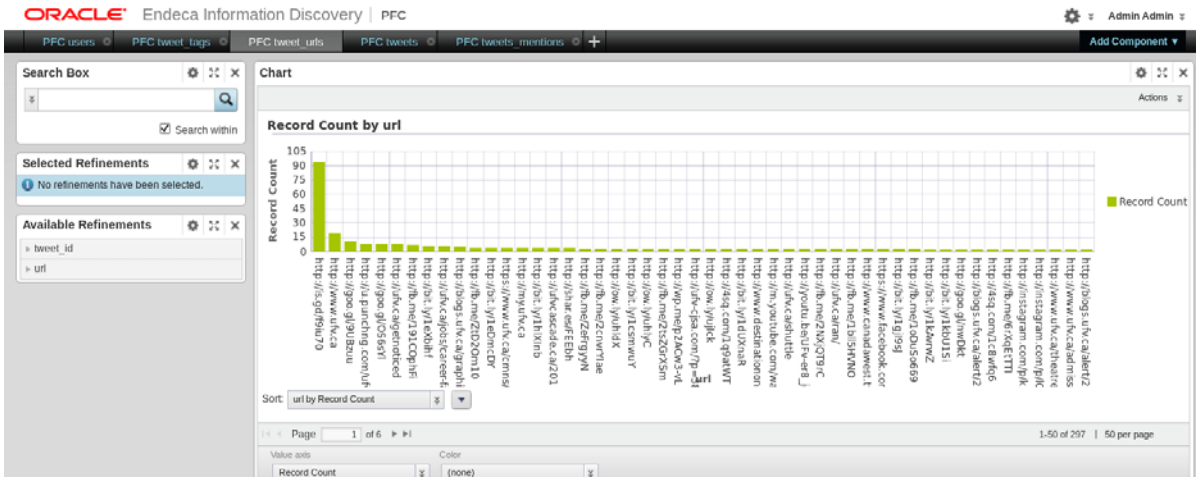


Figura 40. Pestaña PFC tweet_urls en Oracle Endeca Studio. Fuente: elaboración propia.

Pestaña Tuit (PFC tuits)

Esta pestaña muestra el contenido de la tabla "tweets". En esta pestaña sí que se puede observar que hay 50 tuits aproximadamente procedentes del mismo usuario (figura 41). En la siguiente fase se mejorarán estos resultados.

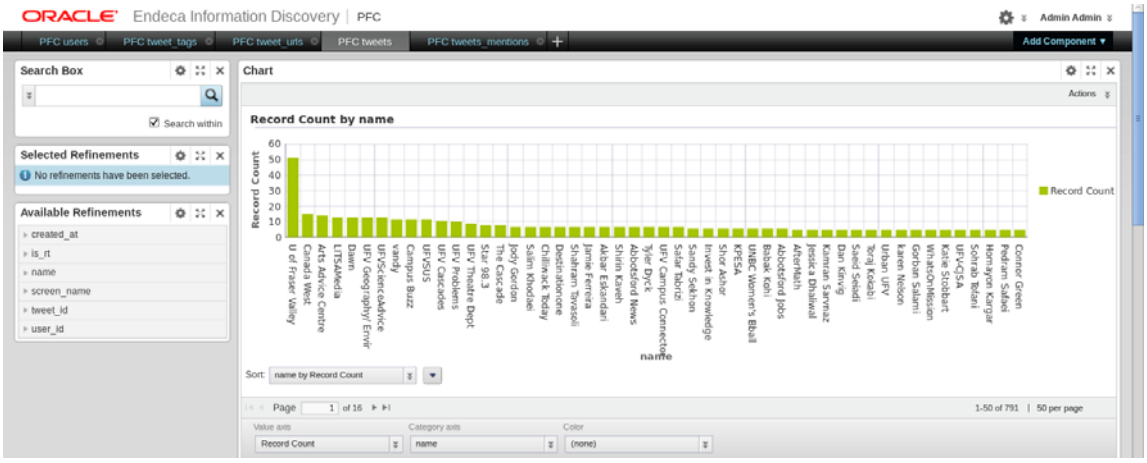
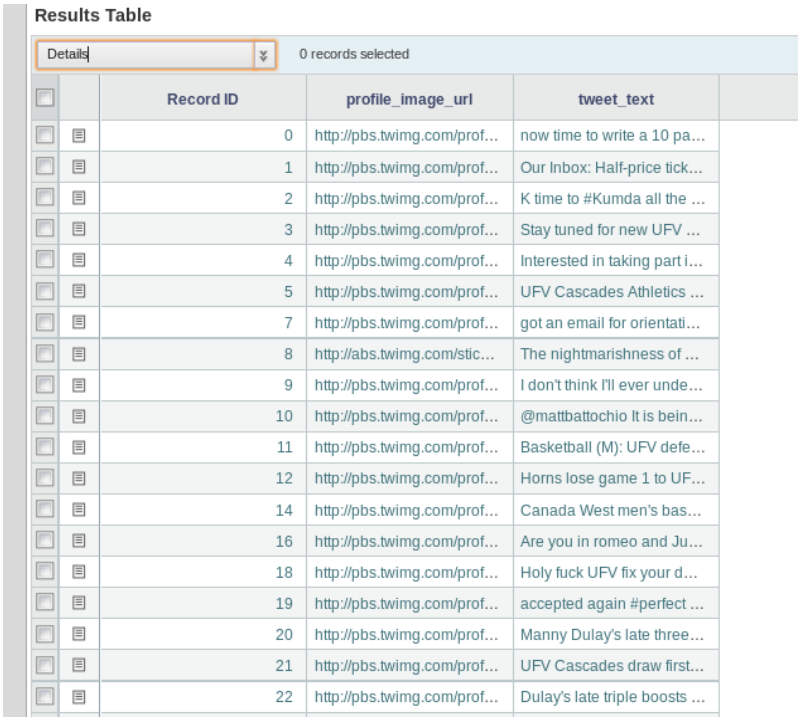


Figura 41. Pestaña PFC tuits. Fuente: elaboración propia.

En la Figura 42 se observa que en el campo "tweet_text" (el contenido del tuit) que la mayoría de los mismos se encuentran en inglés. En la siguiente fase se tendrá que configurar la API para que solo traiga los tuits en español, ya que el estudio es solo de las universidades de Madrid.



Record ID	profile_image_url	tweet_text
0	http://pbs.twimg.com/prof...	now time to write a 10 pa...
1	http://pbs.twimg.com/prof...	Our Inbox: Half-price tick...
2	http://pbs.twimg.com/prof...	K time to #Kumda all the ...
3	http://pbs.twimg.com/prof...	Stay tuned for new UFV ...
4	http://pbs.twimg.com/prof...	Interested in taking part i...
5	http://pbs.twimg.com/prof...	UFV Cascades Athletics ...
7	http://pbs.twimg.com/prof...	got an email for orientati...
8	http://abs.twimg.com/stic...	The nightmarishness of ...
9	http://pbs.twimg.com/prof...	I don't think I'll ever unde...
10	http://pbs.twimg.com/prof...	@mattbattochio It is bein...
11	http://pbs.twimg.com/prof...	Basketball (M): UFV defe...
12	http://pbs.twimg.com/prof...	Horns lose game 1 to UF...
14	http://pbs.twimg.com/prof...	Canada West men's bas...
16	http://pbs.twimg.com/prof...	Are you in romeo and Ju...
18	http://pbs.twimg.com/prof...	Holy fuck UFV fix your d...
19	http://pbs.twimg.com/prof...	accepted again #perfect ...
20	http://pbs.twimg.com/prof...	Manny Dulay's late three...
21	http://pbs.twimg.com/prof...	UFV Cascades draw first...
22	http://pbs.twimg.com/prof...	Dulay's late triple boosts ...

Figura 42. Tabla detalles en Oracle Endeca Studio. Fuente: elaboración propia.

Pestaña Tuits menciones (PFC tweets_mentions)

Esta tabla (figura 43) muestra el contenido de la tabla "tweets_mentions" y como se puede observar a primera vista no muestra información relevante. Muestra la relación entre el usuario que emite el tuit (source_user_id) y el usuarios que lo recibe (target_user_id).

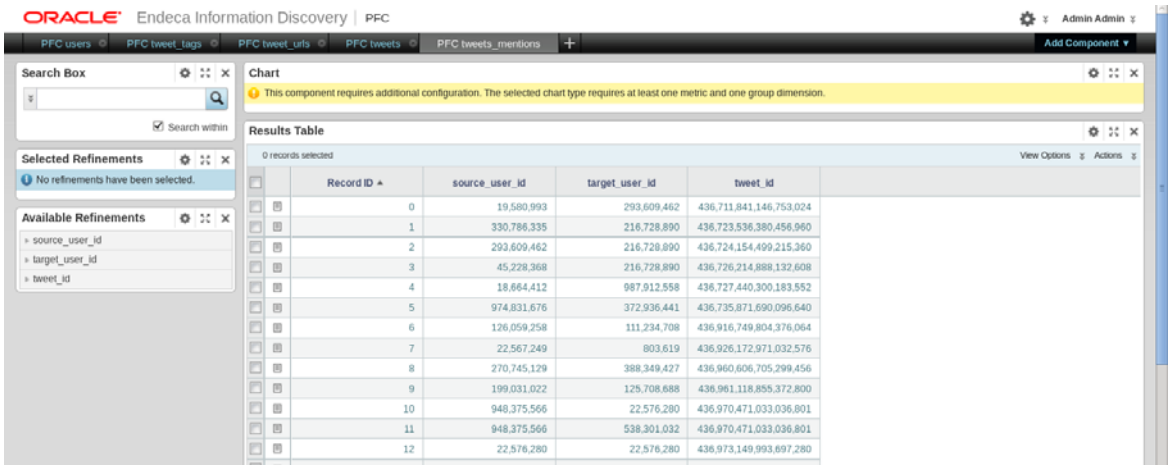


Figura 43. Pestaña PFC tweets_mentions. Fuente: elaboración propia.

6.2.3.4 Resultado análisis de riesgos

En esta sección se expondrán los resultados del análisis de riesgos que se ha llevado a cabo.

RIESGO-F01-01: Pérdida de datos

Como plan preventivo ante una pérdida de datos se ha planificado una estrategia de backups semanal. Durante las primeras 3 semanas no se ha realizado un backup ya que todavía no estaba desarrollado el sistema de adquisición de datos.

Tabla 15. RIESGO-F01-01. Fuente: elaboración propia.

RIESGO-F01-01	
Semana	Backup realizado
4	-1
5	-1
6	-1
7	1
8	1
9	1

Aunque se posee backup de las últimas semanas estos datos se descartarán ya que están en inglés y el idioma debe de ser español.

RIESGO-F01-02: Exceso de transferencia de datos

Durante el domingo 16 de febrero y el lunes 17 de febrero de 2014 el sistema de adquisición de datos en la nube, alojado en Amazon EC2, sufre un ataque de seguridad y se generan casi 14 TB de tráfico saliente hacia Asia. Casi dos meses más tarde se descubre un fallo de seguridad en el protocolo SSL conocido como “HeartBleed” (52). Además ese mismo día se recibe un informe de abuso, procedente de Amazon, informando que se están realizando escaneos de puertos a desde la instancia levantada.

Este fallo afecta al sistema operativo de instalado AMI (Amazon Linux) en todas las versiones (53) (54). Días más tarde, al haber superado el tráfico saliente gratuito, Amazon se pone en contacto indicando que hay una factura pendiente de 1.646,85$.

Como se puede ver en la tabla 16, el tráfico generado supera el límite de tráfico gratuito semanal.

Tabla 16. RIESGO-F01-02. Fuente: elaboración propia.

RIESGO-F01-02			
Semana	Tráfico generado (MB)	Tráfico acumulado (MB)	Límite de tráfico (MB)
4	12	12	3.840
5	15	27	3.840
6	32	59	3.840
7	52	111	3.840
8	58	169	3.840
9	13.854.041	13.854.210	3.840

En la figura 44 se muestra la evolución del tráfico generado (eje y en MB) a lo largo de las semanas de la fase 1.

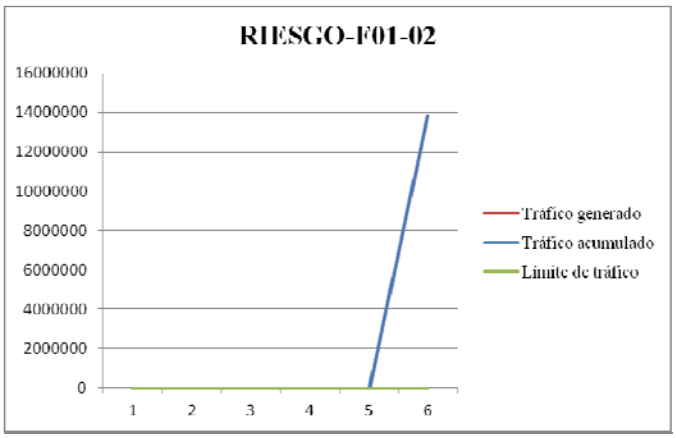


Figura 44. RIESGO-F01-02 gráfico. Fuente: elaboración propia.

Aunque el riesgo estaba supervisado, no se había contemplado su severidad. Se procede al plan de contingencia del RIESGO-F01-02, cambio de la Streaming API de Twitter por la Rest API (4).

Durante la Fase 2, se desarrollará una reestructuración del sistema de adquisición de datos y se prescindirá de los servicios web de Amazon. Este riesgo supone una carga extra en el proyecto y podría provocar retrasos.

RIESGO-F01-03: Cantidad de datos adquiridos

El uso de la Streaming API de Twitter no ha dado los resultados esperados en cuanto a la cantidad de datos adquiridos. Como se observa en la tabla 17, el objetivo de esta fase era obtener el 14% de los tuits y se ha conseguido un 1%.

Esto es debido a que la Streaming API parece estar limitada a un 1% de los tuits generados.

Tabla 17. RIESGO-F01-03. Fuente: elaboración propia.

RIESGO-F01-03						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
4	-	-	-	-	0%	0%
5	-	-	-	-	0%	0%
6	-	-	-	-	0%	0%
7	430	11.363	430	11.363	0%	5%
8	575	11.363	1.005	22.726	0%	9%
9	421	11.363	1.426	34.089	1%	14%
TOTAL	1.426	53.571	1.426	250.000	1%	

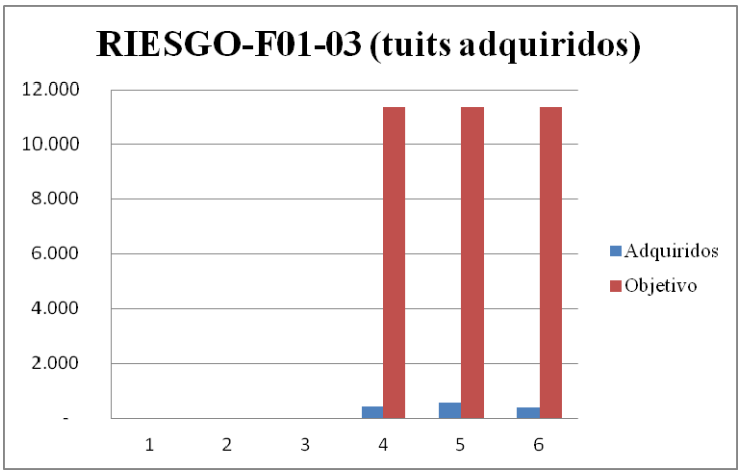


Figura 45. RIESGO-F01-03 (tuits adquiridos). Fuente: elaboración propia.

El número de tuits adquiridos no ha sido el esperado (figura 45) y el crecimiento (figura 46) esperado no se acerca al teórico.

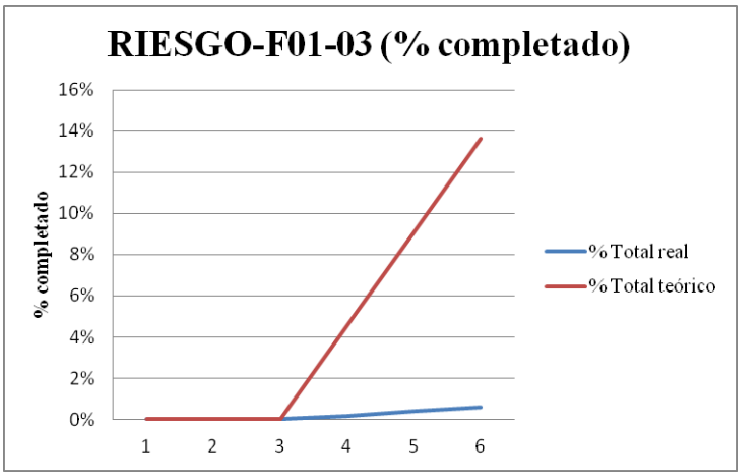


Figura 46. RIESGO-F01-03 (%completado). Fuente: elaboración propia.

RIESGO-F01-04: Rendimiento del sistema integrado

Con todos los componentes de Oracle Endeca cargados y arrancados, el sistema es estable y relativamente fluido. La memoria RAM alcanza casi el límite físico 8 GB aunque la capacidad de procesamiento se mantiene la mayor parte del tiempo por debajo del 10%.

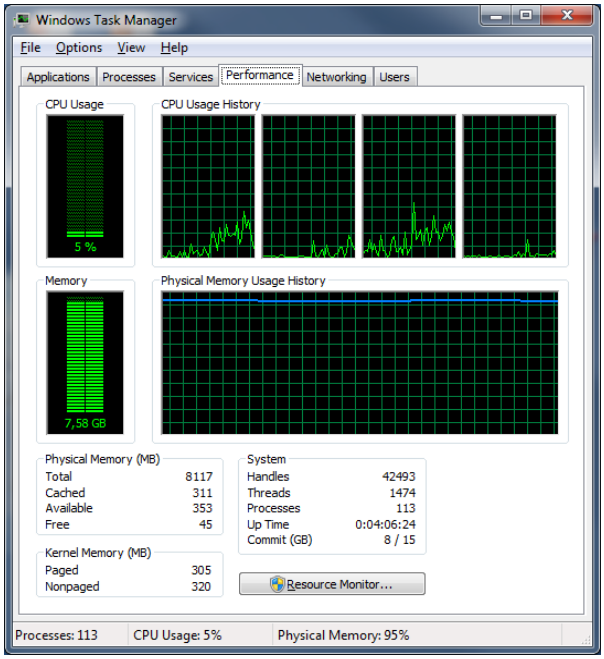


Figura 47. Rendimiento sistema en la fase 1. Fuente: elaboración propia.

6.2.4 Planificación

La planificación de esta primera fase ha sido la siguiente (figura 48)

Fase 1: Adquisición de datos en la nube e integración de componentes	30 days?	**6/01/14 8:00**	**14/02/14 17:00**
Determinar objetivos	2 days	**6/01/14 8:00**	**7/01/14 17:00**
Adquisición de datos	2 days	6/01/14 8:00	7/01/14 17:00
Integración de componentes	2 days	6/01/14 8:00	7/01/14 17:00
Análisis de riesgo	3 days?	**8/01/14 8:00**	**10/01/14 17:00**
Identificación de riesgos	1 day?	8/01/14 8:00	8/01/14 17:00
Análisis de riesgos	1 day?	8/01/14 8:00	8/01/14 17:00
Planificación de riesgos	1 day?	9/01/14 8:00	9/01/14 17:00
Supervisión de riesgos	1 day?	10/01/14 8:00	10/01/14 17:00
Desarrollo y verificación del producto	24 days?	**13/01/14 8:00**	**13/02/14 17:00**
Adquisición de datos	10 days?	**13/01/14 8:00**	**24/01/14 17:00**
Preparación del entorno	9 days?	**13/01/14 8:00**	**23/01/14 17:00**
Ejecución y verificación	1 day?	24/01/14 8:00	24/01/14 17:00
Integración componentes Oracle Endeca	10 days?	**27/01/14 8:00**	**7/02/14 17:00**
Integración Oracle Enterprise Linux	2 days?	27/01/14 8:00	28/01/14 17:00
Integración Oracle WebLogic	1 day	29/01/14 8:00	29/01/14 17:00
Oracle Database	2 days?	30/01/14 8:00	31/01/14 17:00
Oracle Endeca Server	2 days	3/02/14 8:00	4/02/14 17:00
Oracle Endeca Studio	2 days	5/02/14 8:00	6/02/14 17:00
Oracle Endeca Provisioning	1 day	7/02/14 8:00	7/02/14 17:00
Resultados y verificación	4 days?	**10/02/14 8:00**	**13/02/14 17:00**
Resultados adquisición de datos	2 days	10/02/14 8:00	11/02/14 17:00
Resultados integración de componentes	3 days	11/02/14 8:00	13/02/14 17:00
Resultados visualización	2 days	12/02/14 8:00	13/02/14 17:00
Resultado análisis de riesgos	1 day?	13/02/14 8:00	13/02/14 17:00
Planificación	1 day?	14/02/14 8:00	14/02/14 17:00

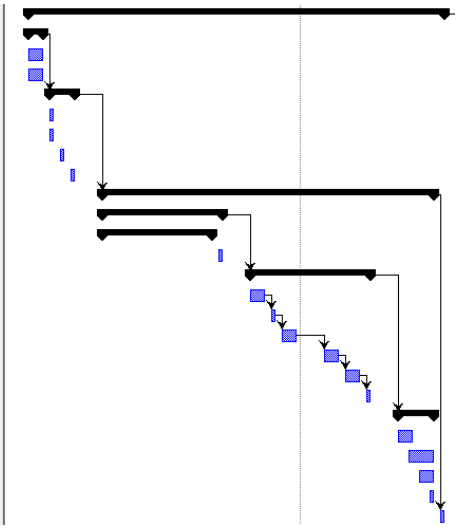
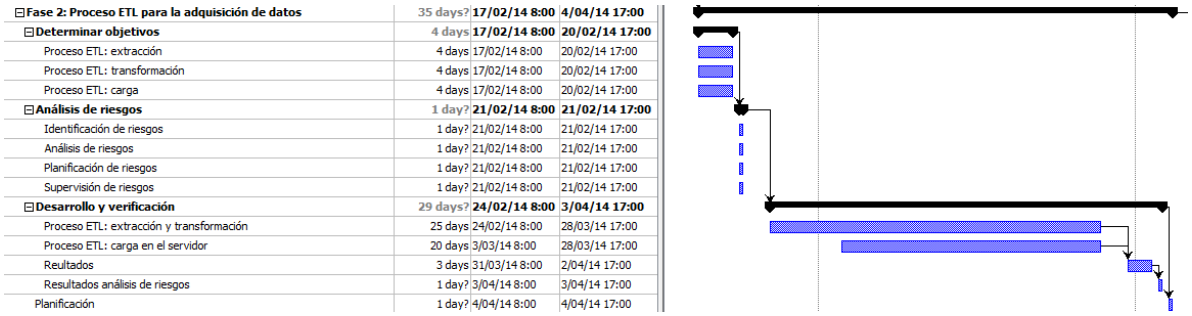


Figura 48. Planificación fase 1. Fuente: elaboración propia.

Para la siguiente fase, será necesario reestructurar el sistema de adquisición de datos, cambiando su arquitectura por completo. Además se tendrá cambiar el idioma de adquisición de tuits a español y definir nuevos términos de búsqueda más precisos.

La planificación que se llevará en la tercera fase será la siguiente (figura 49).



Fase 2: Proceso ETL para la adquisición de datos	35 days?	**17/02/14 8:00**	**4/04/14 17:00**
Determinar objetivos	4 days	**17/02/14 8:00**	**20/02/14 17:00**
Proceso ETL: extracción	4 days	17/02/14 8:00	20/02/14 17:00
Proceso ETL: transformación	4 days	17/02/14 8:00	20/02/14 17:00
Proceso ETL: carga	4 days	17/02/14 8:00	20/02/14 17:00
Análisis de riesgos	1 day?	**21/02/14 8:00**	**21/02/14 17:00**
Identificación de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Análisis de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Planificación de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Supervisión de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Desarrollo y verificación	29 days?	**24/02/14 8:00**	**3/04/14 17:00**
Proceso ETL: extracción y transformación	25 days	24/02/14 8:00	28/03/14 17:00
Proceso ETL: carga en el servidor	20 days	3/03/14 8:00	28/03/14 17:00
Reultados	3 days	31/03/14 8:00	2/04/14 17:00
Resultados análisis de riesgos	1 day?	3/04/14 8:00	3/04/14 17:00
Planificación	1 day?	4/04/14 8:00	4/04/14 17:00

Figura 49. Planificación fase 2. Fuente: elaboración propia.

6.3 Fase 2: Proceso ETL para la adquisición de datos

En esta fase se cambiará todo el sistema de adquisición de datos. No se utilizarán los servicios web proporcionados por Amazon ni la Streaming API de Twitter.

Se desarrollará un proceso ETL, el cual realice la conexión con la Rest API de Twitter, extraiga los campos solicitados, los transforme y los cargue directamente en el servidor de Oracle Endeca.

A diferencia de la Streaming API de Twitter, la Rest API no mantiene una conexión abierta y por tanto habrá que realizar manualmente la solicitud de tuits.

Esta fase comienza el 17 de febrero de 2014 y acabará el 7 de abril de 2014.

6.3.1 Determinar objetivos

Al igual que en la fase 1 del proyecto, la primera actividad según la metodología de desarrollo en espiral es determinar los objetivos de esta segunda fase.

6.3.1.1 Proceso ETL: Extracción

Este objetivo consistirá en la conexión a la Rest API de Twitter mediante una solicitud URL. Se tendrá en cuenta la siguiente configuración:

- Nueva definición de los términos de búsqueda (Anexo II).
- Extracción solo de tuits en español.

Para este proceso de extracción se utilizará una herramienta ETL llamada Oracle Integrator ETL.

6.3.1.2 Proceso ETL: Transformación

Una vez realizada la extracción, el servidor de Twitter enviará un fichero en formato JSON el cual habrá que transformarlo dividiéndolo en campos o metadatos.

Para este proceso de transformación se utilizará la herramienta ETL llamada Oracle Integrator ETL.

6.3.1.3 Proceso ETL: Carga

Con la transformación realizada y los metadatos estructurados, se procederá a una carga directa en el servidor de Oracle Endeca. En la fase anterior, se realizaban las conexiones a la base de datos proporcionada por Amazon desde la interfaz gráfica de Oracle Endeca Studio. En esta fase se realizará la carga al servidor mediante el uso de los servicios web proporcionados por Endeca Server.

Para este proceso de carga se utilizará la herramienta ETL llamada Oracle Integrator ETL.

6.3.1.4 Resultados

Con el proceso ETL completado se procederá a mostrar los resultados mediante la creación de un cuadro de mandos en Oracle Endeca Studio.

6.3.2 Análisis de riesgos

Después de los riesgos analizados en la fase 1 del proyecto, se mantendrán tres de los riesgos ya identificados.

6.3.2.1 Identificación de riesgos

En esta fase del proyecto se han identificado los siguientes riesgos.

Tabla 18. Identificación de riesgos de la fase 2. Fuente: elaboración propia.

Identificador	Riesgo	Tipo	Descripción
RIESGO-F02-01	Pérdida de datos	Proyecto	Pérdida de datos por corrupción de los mismos.
RIESGO-F02-02	Cantidad de datos adquiridos	Proyecto	No tener suficientes datos para realizar el estudio
RIESGO-F02-03	Rendimiento del sistema integrado	Producto	Capacidades de hardware

RIESGO-F02-01: Pérdida de datos

Durante el proceso ETL puede ocurrir una pérdida de datos o corrupción debido a una parada inesperada o incluso la modificación de algún parámetro podría afectar a la estructura de los metadatos y provocaría una pérdida de los datos.

RIESGO-F02-02: Cantidad de datos adquiridos

Al igual que en la primera fase, se estima que para realizar un estudio sobre las universidades de Madrid, se necesitan 250 mil tuits aproximadamente. Por ello se analizará y se supervisará la cantidad de datos adquiridos semanalmente.

RIESGO-F02-03: Rendimiento del sistema integrado

Toda la plataforma de análisis de datos se ha desarrollado sobre una máquina virtual con capacidades limitadas. Es necesario monitorizar el rendimiento de la misma a medida que se van agregando nuevos datos al servidor de Oracle Endeca.

6.3.2.2 Análisis de riesgos

Una vez identificados los riesgos se procede a su análisis y la probabilidad de que ocurran.

Tabla 19. Análisis de riesgos de la fase 2. Fuente: elaboración propia.

Identificador	Nombre	Probabilidad
RIESGO-F02-01	Pérdida de datos	Muy bajo
RIESGO-F02-02	Cantidad de datos adquiridos	Moderado
RIESGO-F02-03	Rendimiento del sistema integrado	Muy bajo

RIESGO-F02-01: Pérdida de datos

Se ha considerado que la probabilidad de pérdida de datos es muy baja, ya que al trabajar en local se pueden realizar copias de seguridad periódicamente. En el peor de los casos existiría una pérdida de los datos de una semana.

RIESGO-F02-02: Cantidad de datos adquiridos

A diferencia de la fase 1, se ha aumentado la probabilidad de que ocurra este riesgo ya que la Streaming API de la fase 1 no ha dado los resultados esperados. En esta fase se espera mejorar considerablemente la cantidad de datos adquiridos.

RIESGO-F02-03: Rendimiento del sistema integrado

Se ha calificado como riesgo muy bajo ya que el rendimiento con del sistema completo ha sido muy bueno en la fase 1. Además en caso de verse limitadas las capacidades hardware se puede proceder a una ampliación de las mismas.

6.3.2.3 Planificación de riesgos

Con los riesgos analizados se procede a su planificación.

RIESGO-F02-01: Pérdida de datos

- **Estrategia de prevención:** se realizarán backups semanales en local.
- **Plan de contingencia:** en caso que ocurra el riesgo identificado se procederá a la restauración del último backup guardado.

RIESGO-F02-02: Cantidad de datos adquiridos

- **Estrategia de prevención:** se monitorizará la cantidad de datos semanalmente.
- **Plan de contingencia:** al utilizar la Rest API de Twitter, se puede ejecutar con más frecuencia para adquirir mayor número de tuits, siempre y cuando se hayan generado nuevos tuits.

RIESGO-F02-03: Rendimiento del sistema integrado

- **Estrategia de prevención:** mediante el gestor de tareas de Windows se puede monitorizar las capacidades del sistema. Se pueden eliminar procesos que no sean esenciales para la ejecución de la máquina.
- **Plan de contingencia:** en caso de verse el hardware totalmente insuficiente se procederá a una ampliación del mismo.

6.3.2.4 Supervisión de riesgos

Con los riesgos planificados se procede a explicar el método de supervisión.

RIESGO-F02-01: Pérdida de datos

Mediante la tabla 20 se mantendrá un control de los backups realizados semanalmente (-1 backup no realizado, 1 backup realizado, 0 backup pendiente de realizar)

Tabla 20. RIESGO-F02-01. Fuente: elaboración propia.

RIESGO-F02-01	
Semana	Backup realizado
10	
11	
12	
13	
14	0
15	0
16	0

Las primeras cuatro semanas se dedicarán a adquirir realizar el proceso ETL y a realizar pruebas, por ello no se realizarán backups.

RIESGO-F02-02: Cantidad de datos adquiridos

Mediante una tabla (tabla 21) se mantendrá monitorizado la cantidad de datos adquiridos semanalmente, así como el objetivo semanal, la cantidad de datos adquiridos en total y el objetivo teórico total de cada semana. Además se mostrará el porcentaje total real, es decir, el porcentaje de datos adquiridos del objetivo de 250 mil tuits frente al porcentaje total teórico, es decir, el porcentaje esperado de tuits adquiridos.

Tabla 21. RIESGO-F02-02. Fuente: elaboración propia.

RIESGO-F02-02						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
10	-	-	-	-	0%	0%
11	-	-	-	-	0%	0%
12	-	-	-	-	0%	0%
13	-	-	-	-	0%	0%
14	-	16.667	-	16.667	0%	7%
15	-	16.667	-	33.334	0%	13%
16	-	16.667	-	50.001	0%	20%
TOTAL	-	50.001	-	250.000	0%	

La figura 50 muestra mediante un gráfico de barras la cantidad de tuits adquiridos frente al objetivo semanal. En el eje X se muestra el número de semana de la fase y en el eje Y la cantidad de tuits.

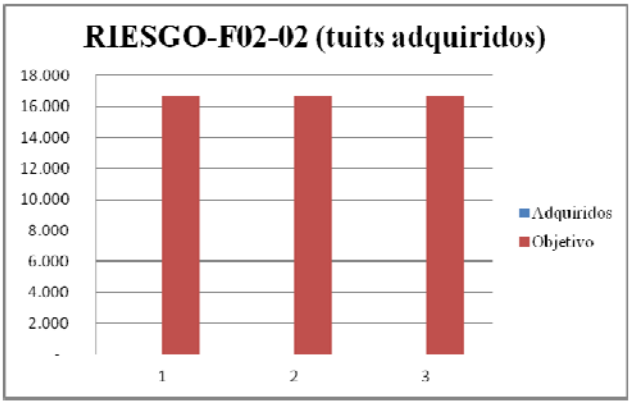


Figura 50. RIESGO-F02-02 (tuits adquiridos). Fuente: elaboración propia.

En la figura 51 se muestra la relación entre el porcentaje real de datos adquiridos frente al ideal de datos que se debería tener.

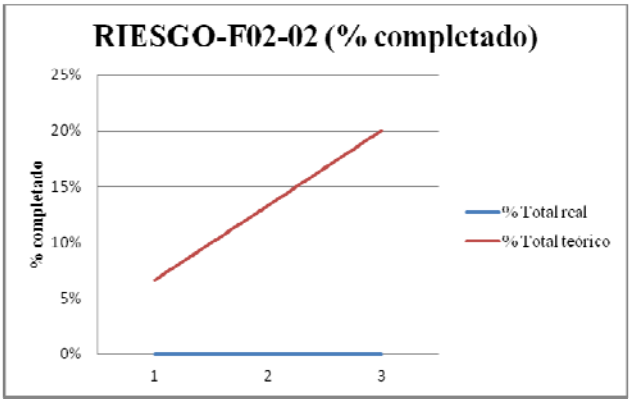


Figura 51. RIESGO-F02-02 (% completado). Fuente: elaboración propia.

RIESGO-F02-03: Rendimiento del sistema integrado

Este riesgo se supervisará de la misma manera que el RIESGO-F01-04. Mediante el administrador de tareas, a observar el uso de la CPU, el uso memoria física así como la paginación del sistema. Se sabrá que se está llegando al límite de las capacidades del ordenador cuando la interactuación con la máquina virtual sea demasiado lenta, se bloquee o no se pueda trabajar.

6.3.3 Desarrollo y verificación

Con el análisis de riesgos realizados se procede a la fase de desarrollo y verificación. En esta fase se desarrollarán los objetivos especificados. Para la realización del proceso ETL se utilizará la herramienta Oracle Integrator ETL 3.1.1.

6.3.3.1 Proceso ETL: Extracción y transformación

En la primera fase de extracción se tendrá que llamar a la Rest API de Twitter. La herramienta ETL permite realizar el proceso completo mediante componentes configurables.

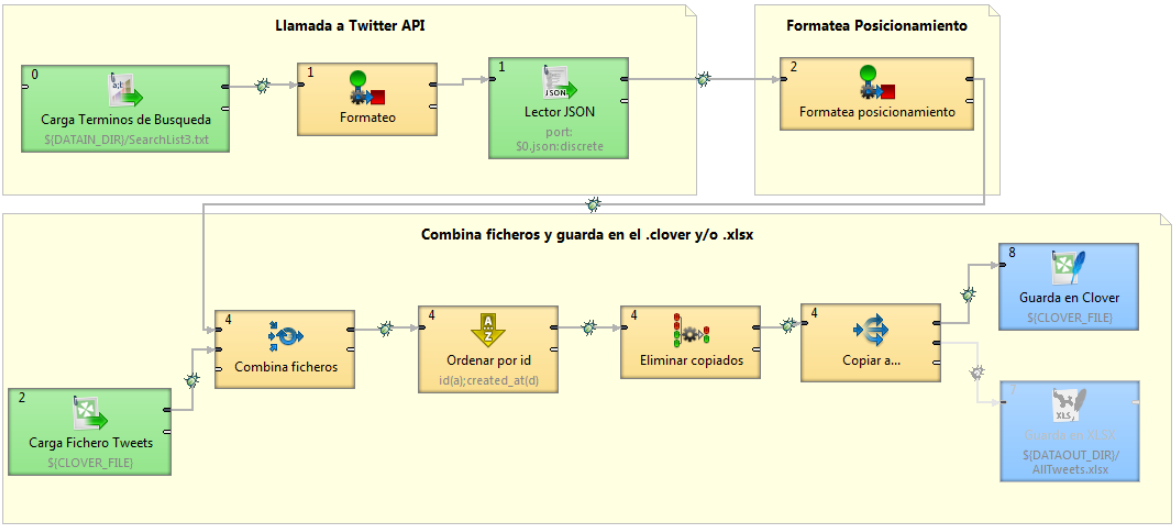


Figura 52. Proceso ETL para la Rest API de Twitter. Fuente: elaboración propia.

El flujo de datos realiza los siguientes procesos:

1. **Carga Términos de búsqueda:** este componente carga los términos de búsqueda que se pasarán como parámetro en la llamada a la Rest API. En el Anexo II se especifican los términos de búsqueda. Por motivos que se verán en el siguiente componente se han separado los términos de búsqueda en 3 ficheros.
2. **Formateo:** este componente, escrito en Java, realiza una cadena de conexión para cada término de búsqueda. Por cada término de búsqueda, se identificará mediante las claves de acceso a la API, se generará la cadena de conexión con el término de búsqueda, el idioma y los parámetros de configuración y se enviará al servidor de Twitter. Por cada llamada la API devuelve hasta 100 tuits, hasta un máximo de 180 llamadas cada 15 minutos (38). Cada respuesta se guardará en un array de JSONs que se enviarán al siguiente componente del proceso ETL. El ajuste de estos parámetros dependerá de cuantos resultados se obtengan de los términos de búsqueda. Los parámetros configurables son:
    - Número de términos a buscar por cada ejecución ETL: 33-34
    - Número de tuits devueltos por cada llamada a la API: 100 (el máximo)
    - Número de llamadas a la API por término a buscar: 150
    - Longitud del array JSONs: 100

    Una vez sobrepasado el límite de 180 llamadas en 15 minutos (38), la aplicación se bloqueará 15 minutos.

    Por ejemplo, imagínese buscar 5 términos con suficientes tuits. Con esta configuración y suponiendo que la cantidad de tuits enviados por el servidor sea siempre homogénea, se realizarían 150 llamadas para el primer término y cada respuesta contendrá 100

tuits, es decir, 150*100=15.000 tuits para el primer término y 30*100=3.000 tuits para el segundo. Después, la aplicación se bloquearía durante 15 minutos.

Después de un estudio con estos términos de búsqueda se ha elegido esta configuración óptima y se ha optado por separar los términos de búsqueda en 3 ficheros que se ejecutarán cada 15 minutos una vez al día.

3. **Lector JSON:** del componente anterior se recibe un array JSON con una longitud máxima de 100 JSONs por cada término de búsqueda. Este componente mapea, mediante el lenguaje XPath (figura 53), los campos del fichero JSON a metadatos previamente configurados.

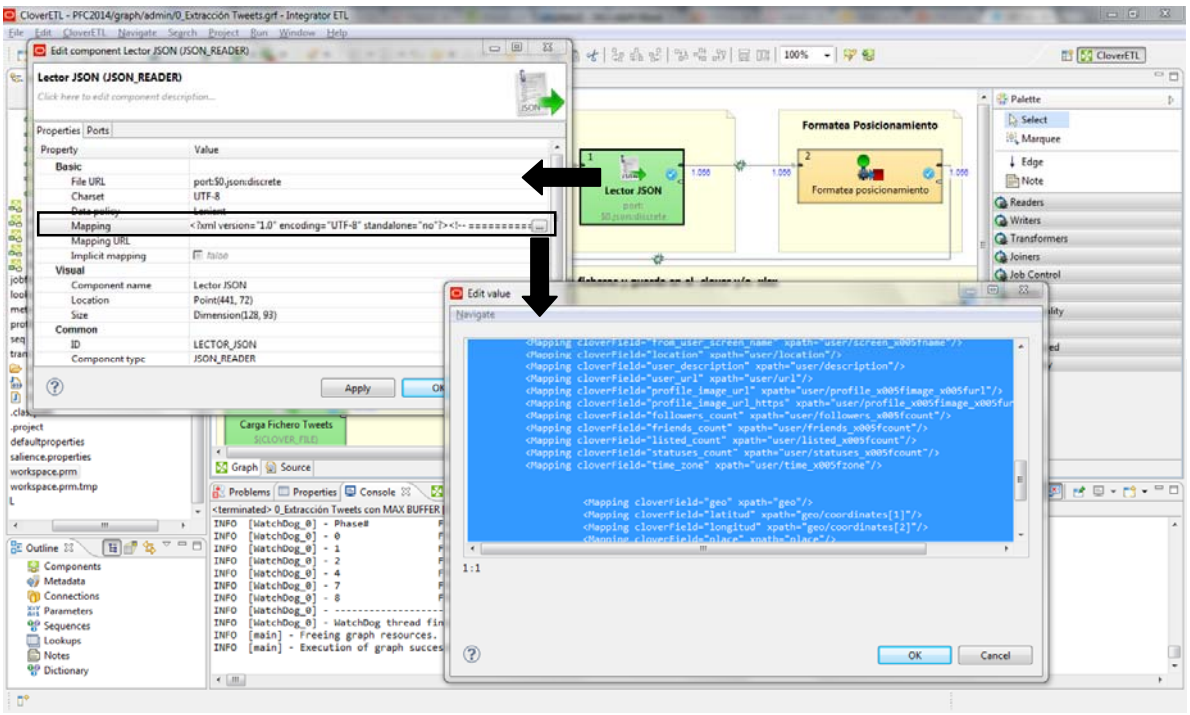


Figura 53. Mapeo del JSON en Oracle Integrator. Fuente: elaboración propia.

4. **Formatea posicionamiento:** este componente en Java, combina la latitud y longitud en un único metadato llamado “geo”, el formato final es “latitud, longitud”. Esto será necesario para mostrar coordenadas en Oracle Endeca.
5. **Combina ficheros:** Este componente combina los metadatos de los tuits recogidos en este proceso con los guardados en una ejecución anterior.
    a. **Carga ficheros tuits:** este componente cargar el fichero de una ejecución anterior. Este fichero contiene los tuits de la pasada ejecución.
6. **Ordenar por id:** al combinar los tuits es frecuente que existan repetidos, por ello se ordenan por la clave primaria, el propio id del tuit.
7. **Eliminar copiados:** este componente en Java, elimina los registros que tienen el mismo id.

8. **Copiar a:** este componente copia el flujo de entrada al flujo de salida. Es muy útil para realizar pruebas y guardar los datos en un fichero Excel por ejemplo.
9. **Guarda en Clover:** guarda los datos de la ejecución en un fichero con extensión clover. Este tipo de archivo es propietario de la herramienta. El fichero final contendrá los registros de la ejecución anterior (cargado en "carga ficheros") y los nuevos que se han adquirido (en el componente "lector JSON").

Para realizar el proceso completo hay que ejecutar el gráfico tres veces, cambiando el fichero de entrada del primer componente (searchlist1.txt, searchlist2.txt, searchlist3.txt). Por cada iteración normalmente se espera 15 minutos para que no bloquee la API a la mitad de la ejecución.

El fichero final con extensión ".clover" contiene los registros (tuits) preformateados con 33 metadatos cada uno. Estos metadatos son información de cada tuit y proporcionados por la API de Twitter como por ejemplo id, fecha de creación, texto, usuario, latitud, longitud, zona horaria, etc. En el Anexo IV se especifican los metadatos utilizados.

6.3.3.2 Proceso ETL: Carga

Para cargar datos en el servidor de Oracle Endeca mediante esta herramienta es necesario realizar una serie de pasos. Se ha automatizado el proceso completo de carga mediante un único gráfico (figura 54) que llama a los demás gráficos.

Figura 54. Proceso ETL de carga completo. Fuente: elaboración propia.

6.3.3.2.1 Inicializa DD

Este gráfico crea un dominio de datos (Data Domain) en el servidor de Oracle Endeca. El dominio de datos es el contenedor de todos los datos de la aplicación (incluyendo datos, configuración, cuadros de mando, variables, etc.). El flujo de datos viene especificado en la figura 55.

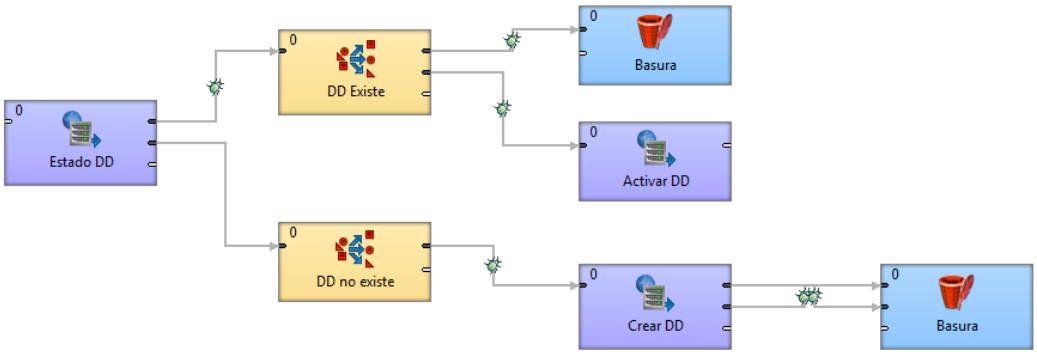


Figura 55. Proceso ETL para inicializar el dominio de datos. Fuente: elaboración propia.

Mediante WSDL se consulta al servidor una serie de peticiones.

1. **Estado DD:** se consulta al servidor si el domino de datos existe. La salida es enviada a todos los componentes conectados.
2. **DD Existe:** si el dominio de datos existe.
   1.1 **Activar DD:** mediante otra llamada WSDL se activa el dominio de datos.
   2.1 **Basura:** el resto de la consulta se descarta.
3. **DD no existe:** si el dominio de datos no existe.
   1.1 **Crear DD:** mediante una llamada WSLD se crea el dominio de datos y por defecto se activa.
   2.1 **Basura:** el resultado de crear el dominio de datos se descarta.

Es conveniente no tener activados todos los dominios de datos ya que estos consumen memoria en la máquina virtual y por tanto ralentizan las operaciones.

6.3.3.2.2 Resetea DD

El siguiente paso una vez creado o activado el dominio de datos es borrar el contenido y su configuración. Este paso es importante pasa asegurarse que no hay configuraciones que puedan generar conflictos.

Figura 56. Componente para resetear el dominio de datos. Fuente: elaboración propia.

6.3.3.2.3 Carga la preconfiguración

Una vez borrado el dominio de datos, es necesario configurar los datos que se van a cargar. Mediante esta configuración se establecerán los atributos que tendrán los registros.

Aunque Oracle Endeca puede trabajar sin un modelo de datos, para realizar una carga mediante llamadas WSDL es necesario preconfigurar los datos que se van a cargar.

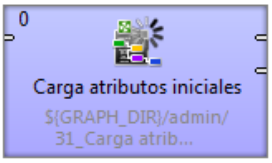


Figura 57. Componente para cargar los atributos iniciales. Fuente: elaboración propia.

6.3.3.2.3.1 Carga de atributos – Metadatos

Todos los parámetros de configuración se encuentran en un fichero Excel (configuration.xls). Este contiene diferentes pestañas.

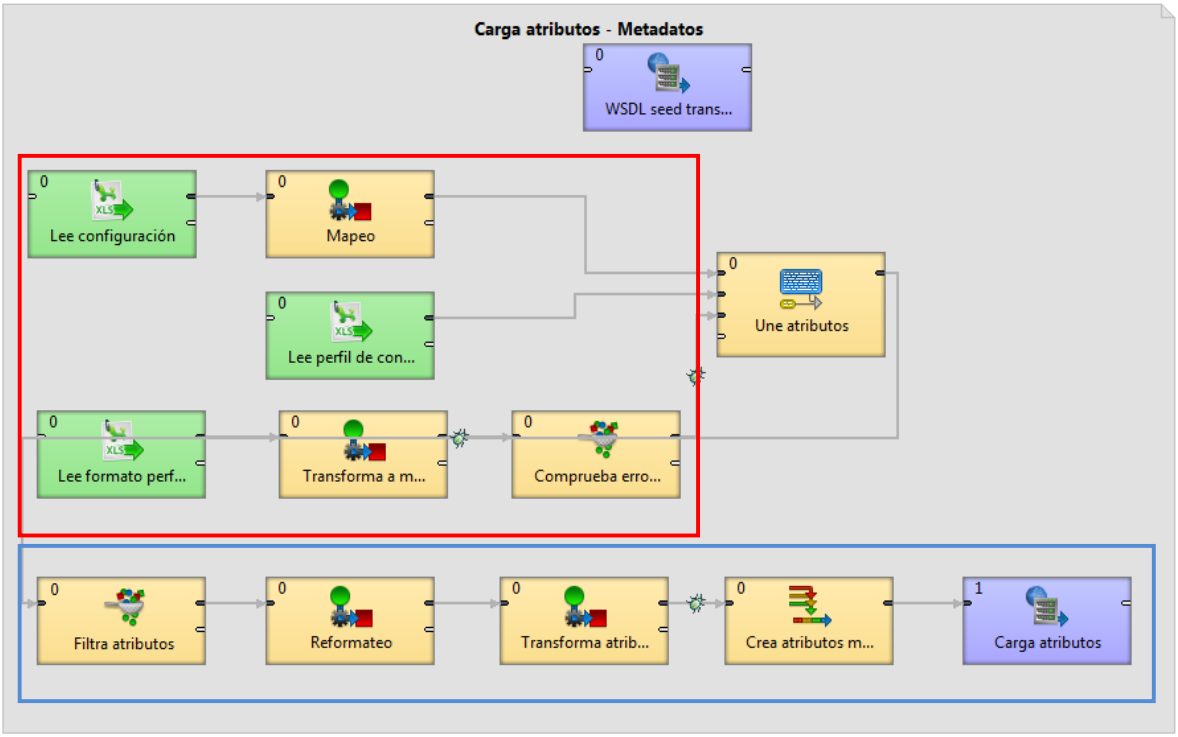


Figura 58. Proceso EL de carga de metadatos. Fuente: elaboración propia.

En primer lugar se seleccionarán una serie de atributos (recuadro en rojo de la figura 58)

- **Lee configuración**: este componente lee la configuración del fichero de configuración de la hoja "configuración", en donde se encuentran los metadatos especificados en el Anexo IV (id, texto, coordenadas, usuario, descripción del usuario, url de la imagen del perfil del usuario, etc.) y se define su tipo (string, int, double), perfil (si es una métrica o una dimensión), si es único, si es indexable y si hay que ordenarlo.
    - **Mapeo**: este componente mapea el flujo de entrada con el de salida, seleccionando algunos de los metadatos que se necesitarán.

- **Lee perfil de configuración**: este componente lee la configuración del fichero de configuración de la hoja "perfil de configuración", en donde se especifica los posibles perfiles indicados en el paso anterior. Los perfiles utilizados son dimensión, métrica, texto, fecha y "por defecto" (el perfil "por defecto" se utiliza para aquellos que no tengan perfil asignado).
- **Lee formato perfil**: este componente lee la configuración del fichero de configuración de la hoja "formato de perfil", en donde se especifican los tipos de datos. Los tipos de datos utilizados son entero, decimal, moneda, porcentaje, geocode (para coordenadas de posicionamiento), booleano y "por defecto" (utilizado para aquellos que no tengan tipo de dato asignado). Para cada tipo de dato, se especifica las posibles operaciones que se pueden realizar, por ejemplo, suma, media, máximo, mínimo, contar, contar distintos, etc. Además se especifica para cada tipo de dato si tiene que tener algún formato específico como por ejemplo, $, €, %, etc.
    - **Transforma a mayúsculas**: mediante Java, se convierte a mayúsculas el flujo de entrada.
    - **Comprueba errores**: mediante Java, este componente solo selecciona aquellos tipos de datos que tengan sentido. Por ejemplo, la fecha no puede tener % o €. Este paso es necesario para evitar errores ya que, como se está accediendo mediante servicios web (WSDL), es necesario configurar los metadatos del servidor de Oracle Endeca mediante una hoja de Excel.
- **Une atributos**: este componente une los atributos anteriormente cargados (figura 58).

Una vez se tenga todos los metadatos, tipos de datos y perfiles configurados se procede a cargarlos al servidor mediante los siguientes pasos (recuadro azul de la Figura 58):

1. **Filtra atributos**: filtra aquellos atributos que cumplen algún criterio como por ejemplo que no sean nulos, o que sean indexables, o que sean únicos, etc.
2. **Reformateo**: este componente es utilizado para pruebas, por defecto la salida es igual a la entrada.
3. **Transforma atributos**: el servidor de Endeca necesita recibir un fichero de configuración XML. Es por ello que hay que poner etiquetas a cada uno de los atributos que se han cargado y transformado.
4. **Crea atributos metadatos**: este componente recoge todo el flujo de entrada con sus etiquetas y genera un fichero XML.
5. **Carga atributos**: como entrada recibe el fichero XML, generado a partir de la configuración del fichero Excel. Realiza una llamada WSDL y carga el fichero de configuración al servido de Oracle Endeca.

Una vez están configurados los metadatos, tipos de datos y perfiles de configuración, el servidor de Oracle Endeca está preparado para recibir los datos.

6.3.3.2.4 Carga de datos

Este gráfico carga los datos estructurados en el servidor de Oracle Endeca. Para ello lee el fichero extensión ".clover" que contiene los datos y mediante el componente "Carga o remplaza registros" carga directamente los datos en el dominio de datos creado previamente.

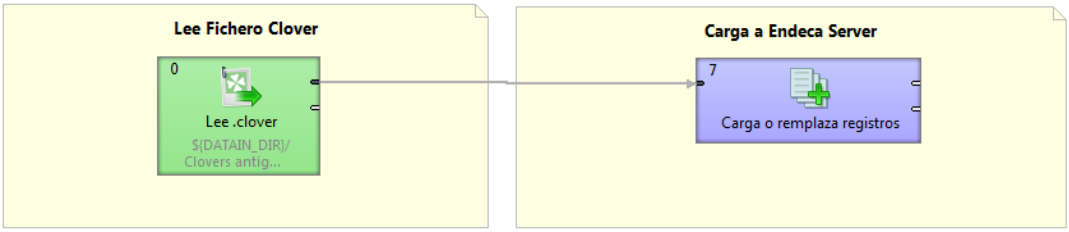


Figura 59. Proceso ETL de carga de datos. Fuente: elaboración propia.

6.3.3.2.5 Carga posconfiguración

Una vez están cargados los datos, se puede cargar una posconfiguración (figura 60) para refinar los resultados. En este caso, debido a que existen numerosos metadatos, se ha optado por crear grupos de metadatos y por otra parte, especificar al sistema las palabras vacías (StopWords) como artículos, pronombres, etc. Las StopWords son palabras sin significado como artículos, determinantes, pronombres, preposiciones etc. Es conveniente indicar al sistema que no tenga en cuenta estas palabras para que no aparezcan en las búsquedas. Esto mejorará el posterior enriquecimiento de texto.

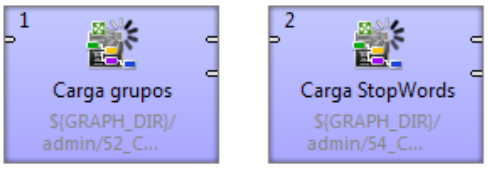


Figura 60. Proceso ETL de carga de configuración. Fuente: elaboración propia.

6.3.3.2.6 Carga de atributos – Grupos

Para crear los grupos hay que modificar el fichero Excel de configuración y añadir una nueva columna indicando para cada metadato en que grupo se encuentra. Se han creado tres grupos de metadatos:

- TwitterAPI: todos los metadatos de tipo texto que devuelve la API de Twitter.
- Fecha: para aquellos metadatos que sean de tipo fecha
- Medida: para los metadatos enteros, como el número de retuits, seguidores, amigos, etc.

En otras palabras, los metadatos "TwitterAPI" serán las dimensiones y los metadatos "Medida" serán las métricas.

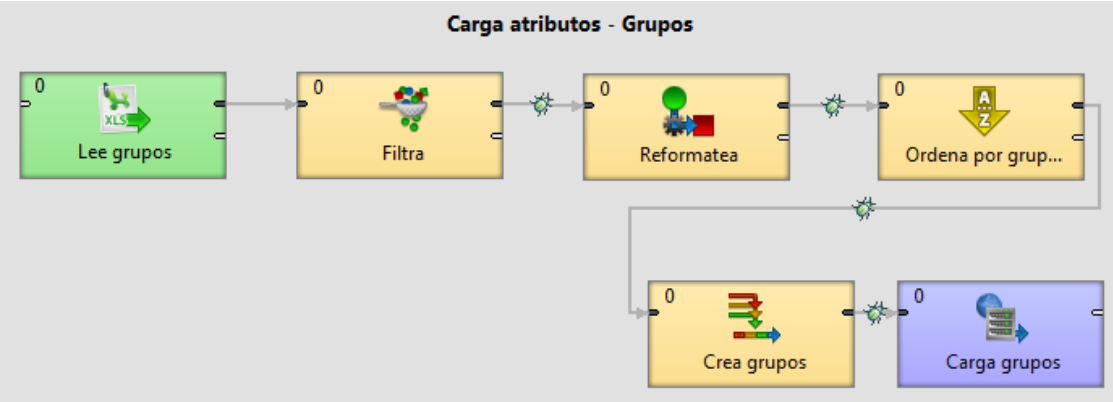


Figura 61. Proceso ETL de carga de atributos. Fuente: elaboración propia.

El proceso es el siguiente:

1. **Lee grupos**: se accede al fichero de configuración y la pestaña de configuración. Esta contiene los metadatos que se han cargado.
2. **Filtra**: se filtran aquellos metadatos que tienen grupo.
3. **Reformatea**: mediante Java, se crean dos nuevos metadatos a la salida, nombre del grupo que será el propio nombre del grupo e id del grupo.
4. **Ordena por grupos:** se ordena por id del grupo.
5. **Crea grupos**: mediante Java, se recoge el nombre del grupo y se crea el fichero XML que posteriormente será cargado.
6. **Carga grupos**: mediante una llamada WSDL y la función "putGroups" se carga el fichero XML de configuración de grupos en el servidor de Oracle Endeca.

6.3.3.2.7 Carga StopWords

Para el reconocimiento de entidades se puede indicar al servidor de Oracle Endeca una seria de palabras vacías, es decir, aquellas que no aportan información y que se repiten mucho, es el caso de los artículos, preposiciones, pronombres, etc.

El procedimiento para cargar la lista de StopWords es similar al anterior.

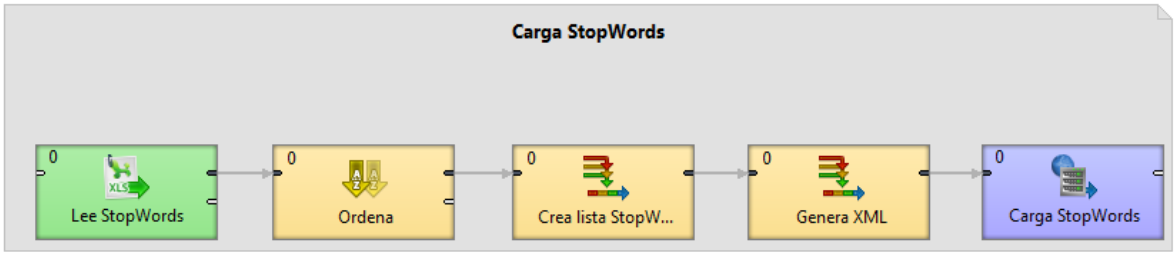


Figura 62. Proceso ETL de carga de StopWords. Fuente: elaboración propia.

1. **Lee StopWords**: abre el fichero de configuración en la pestaña StopWords y lee la lista de palabras. Se han definido 310 palabras vacías (55).
2. **Ordena**: ordena ascendentemente las palabras leídas.

3. **Crea lista StopWords**: mediante Java, para cada palabra se añaden las etiquetas correspondientes para posteriormente generar el fichero XML.
4. **Genera XML:** este componente genera el fichero XML a partir de las etiquetas creadas en el componente anterior.
5. **Carga StopWords**: mediante el servicio web de configuración y la función "putConfigDocuments" se añade el fichero XML "stop_words".

6.3.3.3 Resultados

Una vez cargados los datos al servidor de Endeca, se configura el dominio de datos desde el Oracle Endeca Studio, se crea la aplicación y los resultados son los siguientes.

Nada más acceder se crea una barra resumen y se puede ver en la figura 63 que se han cargado 29.784 tuits, todos en español.

A la izquierda se encuentran los metadatos que se han configurado.

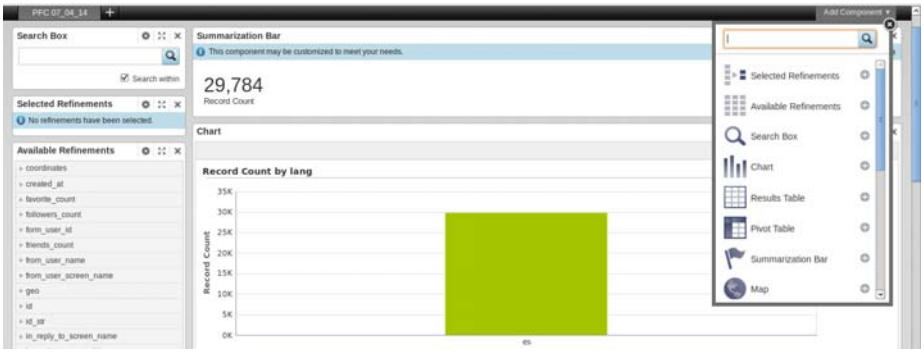


Figura 63. Resultados idioma fase 2 en Oracle Endeca Studio. Fuente: elaboración propia.

Se puede crear una gráfica y ordenar aquellos usuarios que más tuits han generado, en este caso se puede observar en la figura 64, que el usuario ESNE ha generado más de 300 tuits y la Universidad Francisco de Vitoria ocupa la posición 15ª con 100 tuits aproximadamente. Para la siguiente fase será necesario asociar los nombres de usuario de Twitter a un mismo término, es decir, "Universidad Francisco de Vitoria" será la agrupación de tuits que correspondan a "universidad francisco de vitoria", "ufv", "ufvmadrid", etc.

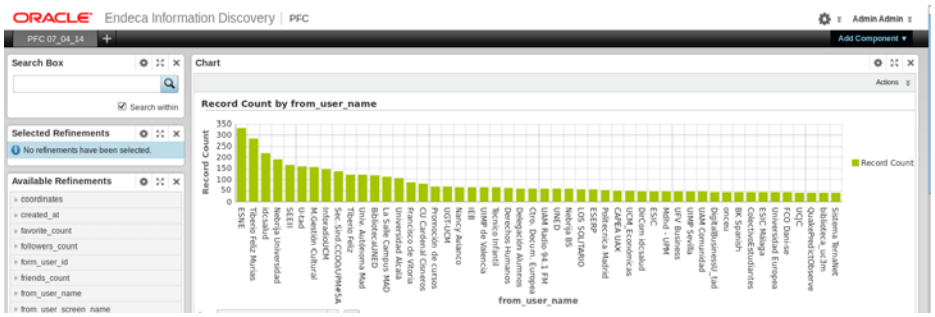


Figura 64. Resultados usuarios fase 2 en Oracle Endeca Studio. Fuente: elaboración propia.

En la figura 65 se puede leer en la columna “text”, los tuits en español almacenados en el servidor.

Results Table

Details | 0 records selected | View Options

Record ID	place	profile_image_url	profile_image_url_https	text	user_description
0		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	Incidentes en la primera ...	Titulares aleatorias de m...
1		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	La Universidad Autónom...	De vacaciones: ¡Oh, qui...
2		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	Competitividad, innovaci...	Somos proyecto, somos ...
3		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	Si todavía no te has ente...	Somos proyecto, somos ...
4		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	BOE : http://t.co/6m2Ot1J...	Canal de oposiciones p...
5		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	La importancia de la for...	CASTELLÓ-SIRVENT C...
6		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	La importancia de la for...	Por Un Chile Más Justo -...
7	c6a1f1f8e7db637aPolyg...	http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	No se como chucha me l...	Ex alumna Lincoln Intern...
8	dbf989427f114112Polyg...	http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	I'm at Universidad San P...	Apasionada del balonce...
9		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	I'm at Universidad San P...	soy profesor de natación...
10		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	Buenos días de parte de...	por los hijos de puta que...
11	1a27537478dd8e38Pol...	http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	@ColorPlus_BCN esper...	Hasta un 70% de #ahorr...
12		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	@Chivatazoescuni a la ...	RRPP MADRID DE ELE...
13		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	RT @dolorymov: El prog...	Grados: Magisterio Infan...
14		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	¡Buenos días! Queda me...	An international challen...
15		http://pbs.twimg.com/prof...	https://pbs.twimg.com/pr...	ESCUNI	Estilos de Apr...	Cuenta oficial

Figura 65. Tabla detalles de los resultados de la fase 2. Fuente: elaboración propia.

Desde el propio Oracle Endeca Studio, en la configuración de la aplicación, se puede realizar un enriquecimiento de texto muy básico: puede realizarse una extracción de términos o un etiquetado de palabras clave a partir de la lista whitelist comentada anteriormente (figura 66).

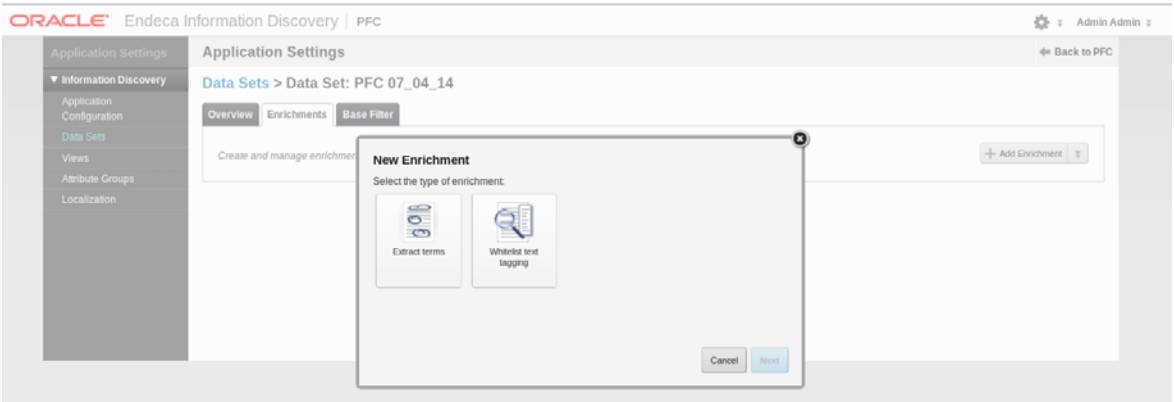


Figura 66. Enriquecimiento de texto desde Oracle Endeca Studio. Fuente: elaboración propia.

Se selecciona extracción de entidades y en el campo text, se indica que como máximo reconozca 5 entidades por registro y que guarde los resultados en el campo “extract terms” (figura 67).

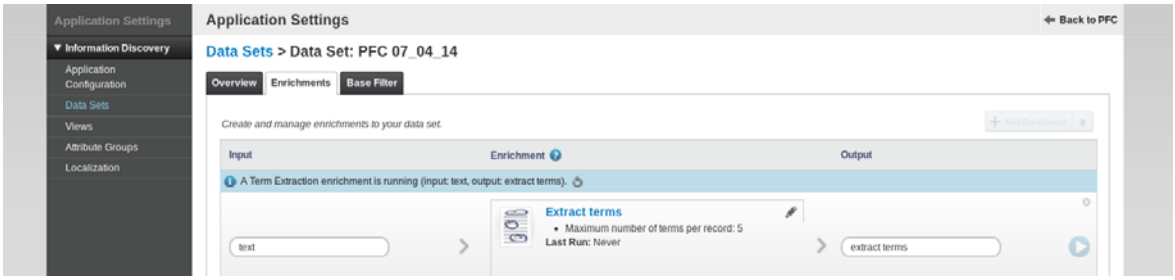


Figura 67. Enriquecimiento de texto 2 desde Oracle Endeca Studio. Fuente: elaboración propia.

Después de 30 minutos (para casi 30 mil registros) de procesamiento los resultados dejan bastante que desear (figura 68). El extractor de términos de Oracle Endeca no ha tenido en cuenta la lista de StopWords ya que el proceso hay que realizarlo desde el la herramienta ETL.

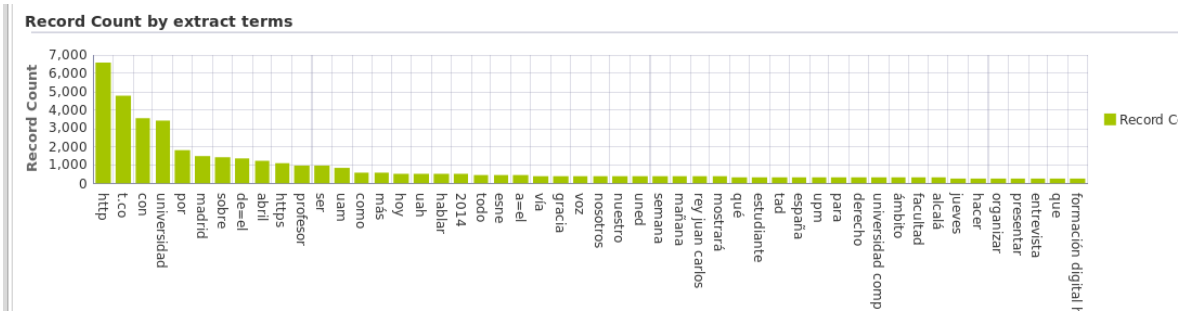


Figura 68. Palabras reconocidas por el enriquecimiento de texto de Oracle Endeca Studio. Fuente: elaboración propia.

Para mejorar los resultados, en la siguiente fase del desarrollo se realizará un análisis de sentimiento con extracción de entidades y palabras clave.

6.3.3.4 Resultados análisis de riesgos

Los riesgos identificados en esta fase se han ido supervisando mediante las técnicas anteriormente descritas.

RIESGO-F02-01: Pérdida de datos

Se han realizado backups semanalmente, desde la semana 14 del proyecto. Las cuatro primeras semanas se han dedicado a la creación de los procesos ETL.

Tabla 22. RIESGO-F02-01. Fuente: elaboración propia.

RIESGO-F02-01	
Semana	Backup realizado
10	-1
11	-1
12	-1
13	-1
14	1
15	1
16	1

RIESGO-F02-02: Cantidad de datos adquiridos

Aunque los resultados no son los esperados ha habido un progreso y al finalizar la fase 2 se ha adquirido el 12% del objetivo propuesto.

Tabla 23. RIESGO-F02-02. Fuente: elaboración propia.

RIESGO-F02-02						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
10	-	-	-	-	0%	0%
11	-	-	-	-	0%	0%
12	-	-	-	-	0%	0%
13	-	-	-	-	0%	0%
14	8.216	16.667	8.216	16.667	3%	7%
15	10.457	16.667	18.673	33.334	7%	13%
16	11.075	16.667	29.748	50.001	12%	20%
TOTAL	29.748	50.001	29.748	250.000	12%	

Mostrando los resultados en forma de gráfico de barras (figura 69) se puede apreciar que efectivamente ha habido una evolución en la adquisición de tuits pero en ningún caso se ha llegado al objetivo semanal.

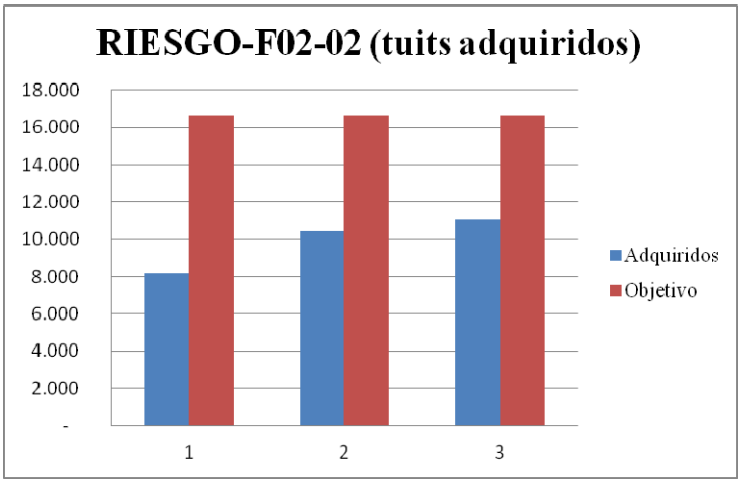


Figura 69. RIESGO-F02-02 (tuits adquiridos). Fuente: elaboración propia.

En la figura 70 se puede observar la tendencia, siendo casi paralela a la teórica, indica que tiene casi el mismo crecimiento. Hay una mejora frente a la fase 1.

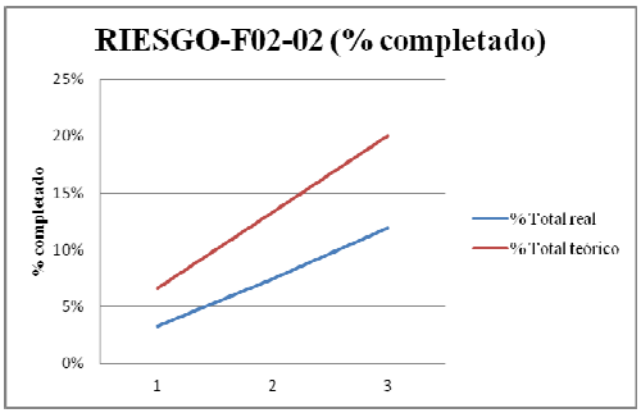


Figura 70. RIESGO-F02-02 (% completado). Fuente: elaboración propia.

Se espera que en la siguiente fase el crecimiento sea mayor.

RIESGO-F02-03: Rendimiento del sistema integrado

Una vez puesto en marcha la máquina virtual, arrancada la base de datos, levantados los dominios de Oracle Endeca Server, Oracle Endeca Studio y Oracle Endeca Provisioning Service y con la aplicación ejecutándose (con 29 mil tuits cargados), el sistema consume toda la memoria física disponible y pagina 9 GB de los 15 GB asignados. Por otra parte, la carga de procesamiento CPU no es elevada.

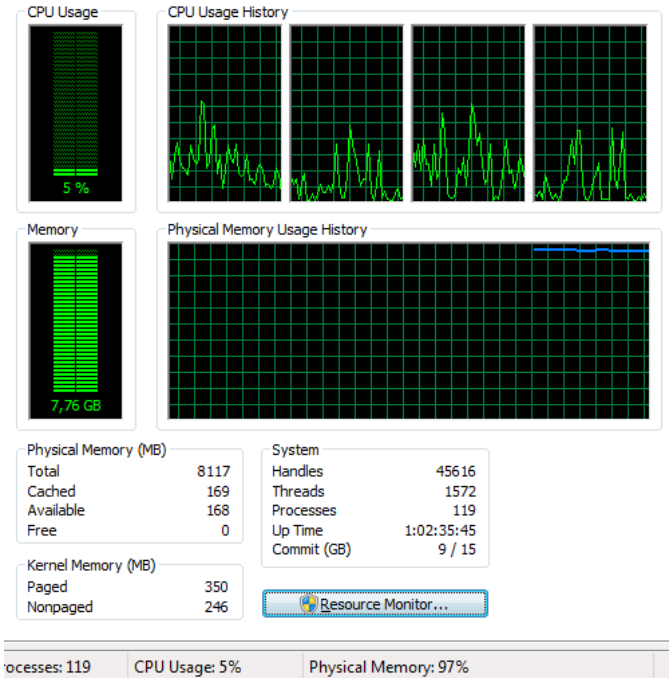


Figura 71. Administrador de tareas con el sistema completo ejecutándose. Fuente: elaboración propia.

6.3.4 Planificación

La planificación llevada a cabo en la segunda fase ha sido la siguiente (figura 72). Como se puede apreciar la mayor parte del tiempo se ha dedicado al desarrollo de los procesos ETL.

Fase 2: Proceso ETL para la adquisición de datos	35 days?	**17/02/14 8:00**	**4/04/14 17:00**
Determinar objetivos	4 days	**17/02/14 8:00**	**20/02/14 17:00**
Proceso ETL: extracción	4 days	17/02/14 8:00	20/02/14 17:00
Proceso ETL: transformación	4 days	17/02/14 8:00	20/02/14 17:00
Proceso ETL: carga	4 days	17/02/14 8:00	20/02/14 17:00
Análisis de riesgos	1 day?	**21/02/14 8:00**	**21/02/14 17:00**
Identificación de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Análisis de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Planificación de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Supervisión de riesgos	1 day?	21/02/14 8:00	21/02/14 17:00
Desarrollo y verificación	29 days?	**24/02/14 8:00**	**3/04/14 17:00**
Proceso ETL: extracción y transformación	25 days	24/02/14 8:00	28/03/14 17:00
Proceso ETL: carga en el servidor	20 days	3/03/14 8:00	28/03/14 17:00
Reultados	3 days	31/03/14 8:00	2/04/14 17:00
Resultados análisis de riesgos	1 day?	3/04/14 8:00	3/04/14 17:00
Planificación	1 day?	4/04/14 8:00	4/04/14 17:00

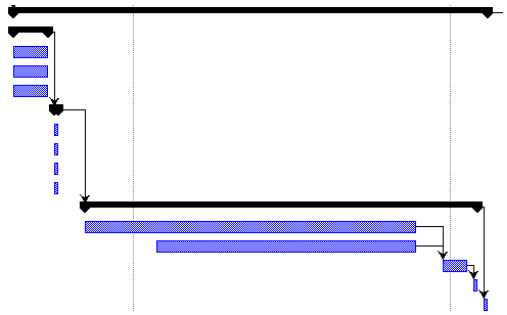


Figura 72. Planificación fase 2. Fuente: elaboración propia.

En la tercera fase se seguirá la siguiente planificación (figura 73).

Fase 3: Análisis de sentimiento	19 days?	**8/04/14 8:00**	**2/05/14 17:00**
Determinar objetivos	3 days	**8/04/14 8:00**	**10/04/14 17:00**
Análisis de sentimiento	2 days	8/04/14 8:00	9/04/14 17:00
Modificación proceso ETL	2 days	9/04/14 8:00	10/04/14 17:00
Análisis de riesgos	1 day?	**11/04/14 8:00**	**11/04/14 17:00**
Identificación de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Análisis de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Planificación de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Supervisión de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Desarrollo y verificación	14 days?	**14/04/14 8:00**	**1/05/14 17:00**
Modificar ETL	10 days	14/04/14 8:00	25/04/14 17:00
Análisis de sentimiento	12 days	16/04/14 8:00	1/05/14 17:00
Resultados análisis de riesgo	1 day?	1/05/14 8:00	1/05/14 17:00
Planificación	1 day?	2/05/14 8:00	2/05/14 17:00

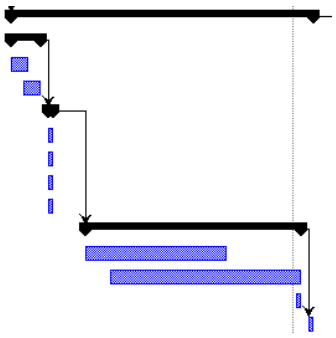


Figura 73. Planificación fase 3. Fuente: elaboración propia.

6.4 Fase 3: Análisis de sentimiento

Para realizar el estudio sobre las universidades de Madrid en Twitter, es necesario implementar un análisis de sentimiento, extraer términos o entidades, puntuar los comentarios de alguna manera y así poder tener una visión global de todos los datos sin tener que leerlos uno a uno.

En la tercera fase del proyecto se realizarán algunas mejoras en el proceso ETL y se utilizará Lexalytics (32) como componente para el análisis de sentimiento. Este componente se integrará dentro del proceso ETL para realizar una carga completa en el servidor de Oracle Endeca.

Esta fase dura 4 semanas, comenzando el 8 de abril de 2014 hasta 4 de mayo de 2014.

6.4.1 Determinar objetivos

La primera actividad de cada fase según la metodología de desarrollo en espiral es determinar los objetivos que se van a implementar en esta fase. Los objetivos serán los siguientes:

6.4.1.1 Análisis de sentimiento

Mediante el componente de Lexalytics (32) se realizará el análisis de sentimiento. Se ha escogido este componente debido a su capacidad de integración con el resto de componentes ya en uso, instalados y configurados.

6.4.1.2 Modificar el proceso ETL

Será necesario añadir nuevos componentes para poder realizar el análisis de sentimiento.

6.4.1.3 Añadir nuevo grupo de metadatos

Durante el proceso ETL de la fase 2 se han definido grupos de metadatos para que posteriormente, desde Oracle Endeca Studio, se tenga una visión organizada de los metadatos. Se añadirá un nuevo grupo para el análisis de sentimiento.

6.4.1.4 Whitelist

Oracle Endeca permite definir listas de relaciones entre palabras claves seleccionadas. Se definirá una lista de universidades de Madrid con sus centros asociados (Anexo II).

Una vez definidos los objetivos se procede al análisis de riesgos de esta tercera fase de desarrollo y verificación.

6.4.2 Análisis de riesgos

Durante el análisis de riesgos se identificarán los riesgos, se analizarán, se planificarán y por último se indicará la manera de supervisarlos.

6.4.2.1 Identificación de riesgos

En esta fase del proyecto se han identificado los siguientes riesgos.

Identificador	Riesgo	Tipo	Descripción
RIESGO-F03-01	Pérdida de datos	Proyecto	Pérdida de datos por corrupción de los mismos.
RIESGO-F03-02	Cantidad de datos adquiridos	Proyecto	No tener suficientes datos para realizar el estudio
RIESGO-F03-03	Rendimiento del sistema integrado	Producto	Capacidades de hardware

Tabla 24. Identificación de riesgos de la fase 3. Fuente: elaboración propia.

RIESGO-F03-01: Pérdida de datos

Durante el proceso ETL puede ocurrir una pérdida de datos o corrupción debido a una parada inesperada o incluso la modificación de algún parámetro podría afectar a la estructura de los metadatos y provocaría una pérdida de los datos.

RIESGO-F03-02: Cantidad de datos adquiridos

Al igual que en la segunda fase, se estima que para realizar un estudio sobre las universidades de Madrid, se necesitan 250 mil tuits aproximadamente.

RIESGO-F03-03: Rendimiento del sistema integrado

Toda la plataforma de análisis de datos se ha desarrollado sobre una máquina virtual con capacidades limitadas. Es necesario monitorizar el rendimiento de la misma a medida que se van agregando nuevos datos al servidor de Oracle Endeca.

6.4.2.2 Análisis de riesgos

Una vez identificados los riesgos se procede al análisis, estudiando la probabilidad de que ocurran. La probabilidad se muestra en la tabla 25.

Identificador	Nombre	Probabilidad
RIESGO-F03-01	Pérdida de datos	Muy bajo
RIESGO-F03-02	Cantidad de datos adquiridos	Moderado
RIESGO-F03-03	Rendimiento del sistema integrado	Muy bajo

Tabla 25. Análisis de riesgo de la fase 3. Fuente: elaboración propia.

RIESGO-F03-01: Pérdida de datos

Al igual que en la fase 2, se ha considerado que la probabilidad es muy baja ya que se ha demostrado que se pueden realizar copias de seguridad fácilmente en local.

RIESGO-F03-02: Cantidad de datos adquiridos

Se ha aumentado la cantidad de datos necesarios por semana a 18.354, esto es debido a que aunque los resultados han sido buenos, no se ha cumplido el objetivo semanal durante la fase 2. En esta fase se espera mejorar considerablemente la cantidad de datos adquiridos.

RIESGO-F03-03: Rendimiento del sistema integrado

Se ha calificado como riesgo muy bajo ya que el rendimiento con del sistema completo ha sido muy bueno en la fase 2. Además en caso de verse limitadas las capacidades hardware se puede proceder a una ampliación de las mismas.

6.4.2.3 Planificación de riesgos

Con los riesgos analizados se procede a su planificación.

RIESGO-F03-01: Pérdida de datos

- **Estrategia de prevención:** se realizarán backups semanales en local.
- **Plan de contingencia:** en caso que ocurra el riesgo identificado se procederá a la restauración del último backup guardado.

RIESGO-F03-02: Cantidad de datos adquiridos

- **Estrategia de prevención:** se monitorizará la cantidad de datos semanalmente.
- **Plan de contingencia:** al utilizar la Rest API de Twitter, se puede ejecutar con más frecuencia para adquirir mayor número de tuits, siempre y cuando se hayan generado nuevos tuits.

RIESGO-F03-03: Rendimiento del sistema integrado

- **Estrategia de prevención:** mediante el gestor de tareas de Windows se puede monitorizar las capacidades del sistema. Se pueden eliminar procesos que no sean esenciales para la ejecución de la máquina.
- **Plan de contingencia:** en caso de verse el hardware totalmente insuficiente se procederá a una ampliación del mismo.

6.4.2.4 Supervisión de riesgos

Con los riesgos planificados se procede a explicar el método de supervisión.

RIESGO-F03-01: Pérdida de datos

Mediante la tabla 26 se mantendrá un control de los backups realizados semanalmente (-1 backup no realizado, 1 backup realizado, 0 backup pendiente de realizar).

Tabla 26. RIESGO-F03-01. Fuente: elaboración propia

RIESGO-F03-01	
Semana	Backup realizado
17	0
18	0
19	0
20	0

RIESGO-F03-02: Cantidad de datos adquiridos

Mediante una tabla se mantendrá monitorizado la cantidad de datos adquiridos semanalmente, así como el objetivo semanal, la cantidad de datos adquiridos en total y el objetivo teórico total de cada semana. Además se mostrará el porcentaje total real, es decir, el porcentaje de datos adquiridos del objetivo de 250 mil tuits frente al porcentaje total teórico, es decir, el porcentaje esperado de tuits adquiridos.

Tabla 27. RIESGO-F03-02. Fuente: elaboración propia.

RIESGO-F03-02						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
Fase 2	29.748	-	29.748			
17	-	18.354	29.748	36.708	12%	15%
18	-	18.354	29.748	55.062	12%	22%
19	-	18.354	29.748	73.416	12%	29%
20	-	18.354	29.748	91.770	12%	37%
TOTAL	-	73.416	29.748	250.000	12%	

La figura 74 muestra mediante un gráfico de barras la cantidad de tuits adquiridos frente al objetivo semanal. En el eje X se muestra el número de semana de la fase y en el eje Y la cantidad de tuits.

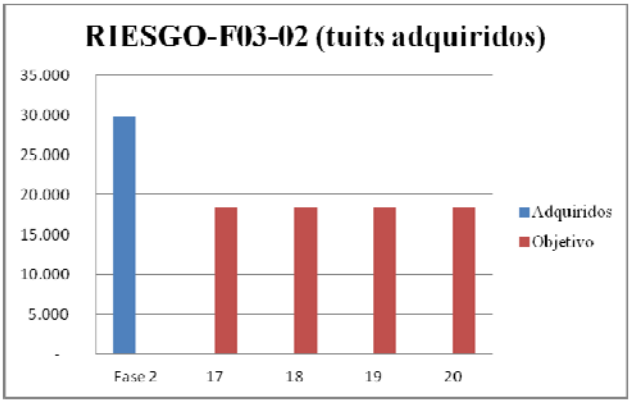


Figura 74. RIESGO-F03-02 (tuits adquiridos). Fuente: elaboración propia.

En la figura 75 se muestra la relación entre el porcentaje real de datos adquiridos frente al ideal de datos que se debería tener.

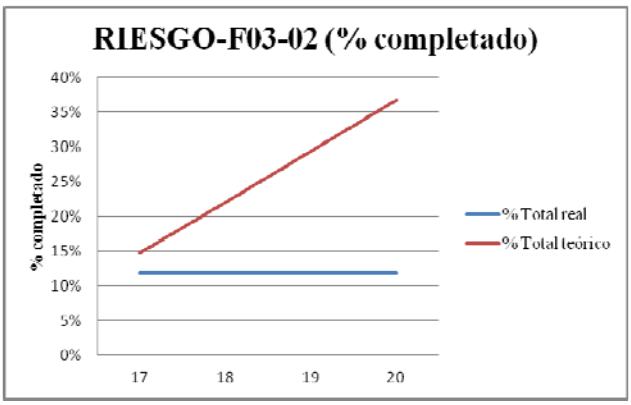


Figura 75. RIESGO-F03-02 (% completado). Fuente: elaboración propia.

RIESGO-F03-03: Rendimiento del sistema integrado

Este riesgo se supervisará de la misma manera que el RIESGO-F02-03. Mediante el administrador de tareas, a observar el uso de la CPU, el uso memoria física así como la paginación del sistema. Se sabrá que se está llegando al límite de las capacidades del ordenador cuando la interactuación con la máquina virtual sea demasiado lenta, se bloquee o no se pueda trabajar.

6.4.3 Desarrollo y verificación

En esta tercera fase se desarrollo y verificación se cumplirán los objetivos descritos. Para ello se modificará el proceso ETL añadiendo el componente de enriquecimiento de texto, se añadirán la lista “Whitelist” y se modificarán los grupos de los atributos de los metadatos.

6.4.3.1 Modificar ETL

En el proceso de carga de la fase 2 del proyecto se realizó un ETL de carga. Este cargaba directamente desde un fichero extensión clover al servidor de Oracle Endeca.

Para añadir el nuevo componente de enriquecimiento de texto es necesario modificar el proceso ETL para tratar los datos antes de cargarlos. El flujo de datos viene descrito en la figura 76.

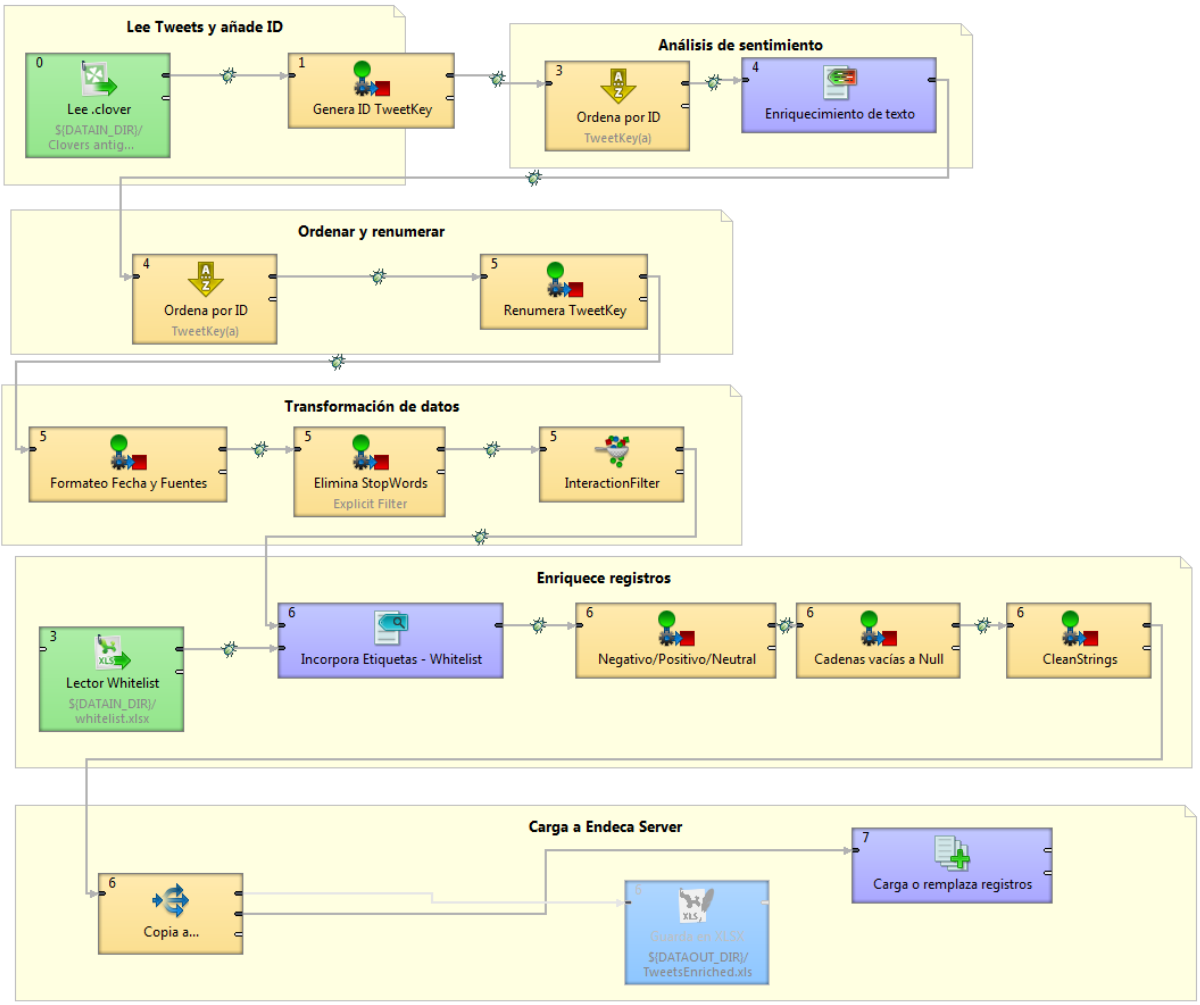


Figura 76. Proceso ETL con análisis de sentimiento. Fuente: elaboración propia

El flujo de datos que se llevará a cabo en el proceso ETL es el siguiente:

LEE Tuits y añade ID

1 **Lee .clover**: este componte lee el fichero extensión clover de entrada. Este fichero contendrá todos los datos de los tuits que se han adquirido mediante el ETL de extracción de tuits.

2 **Genera ID:** mediante java, añade un campo con un ID único.

Análisis de sentimiento

3 **Ordena por ID**: se ordenan los registros por el ID único
4 **Enriquecimiento de texto**: este componente hace una llamada al motor de Lexalytics que realiza el enriquecimiento de texto.

Para configurar el componente (figura 77) se especifica el campo que se quiere enriquecer (Input field), en este caso, el campo "text". Se indica la ruta del fichero de propiedades del motor de Lexalytics (Salience) (31), la ruta y la licencia de prueba.

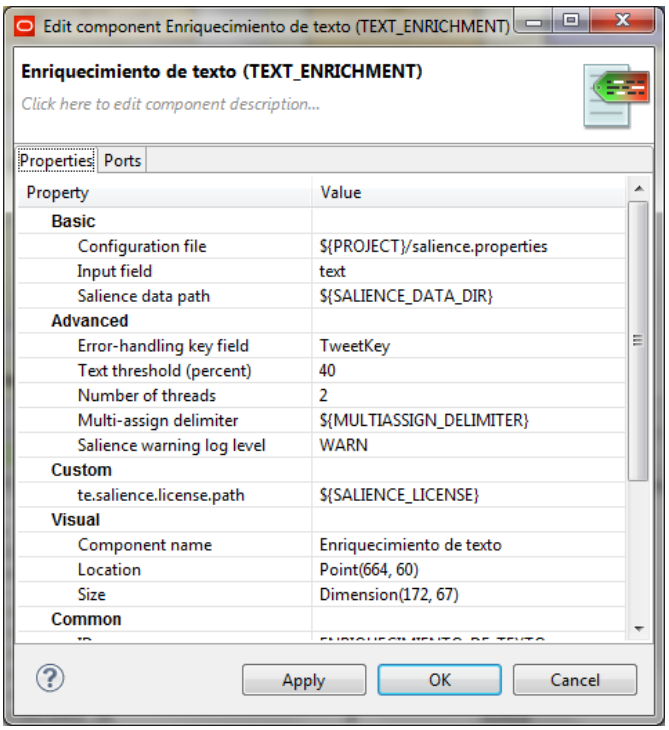


Figura 77. Propiedades del componente de enriquecimiento de texto. Fuente: elaboración propia.

Previamente es necesario crear unos nuevos metadatos de entrada. Cada registro que pase a través del componente se le agregará nuevos datos en los metadatos creados. Para ello se modifican los metadatos y se crean los siguientes, especificados en la figura 78 (Anexo IV).

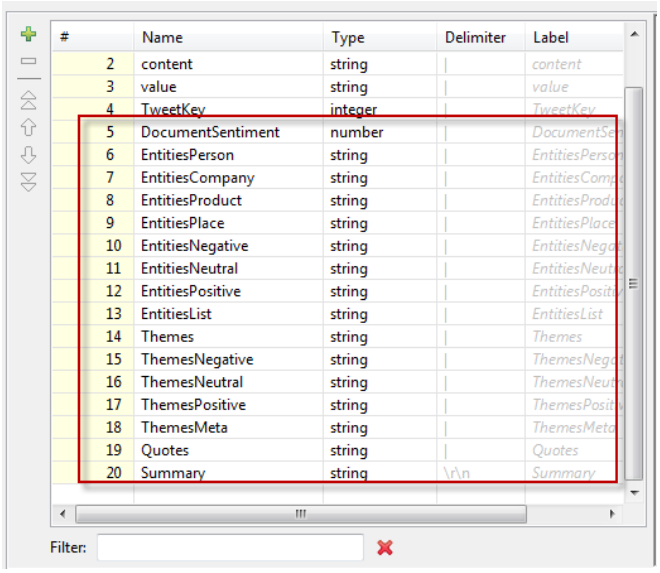


Figura 78. Nuevos metadatos creados para el componente de enriquecimiento de texto. Fuente: elaboración propia.

Ordenar y reenumerar

5 **Ordenar por ID:** una vez se ha enriquecido el texto se habrán rellenado los metadatos correspondientes en función de la salida del componente. Se ordenan los campos por el ID único creado anteriormente.

6 **Reenumerar TweetKey**: es posible que haya habido algún error en el enriquecimiento, por ejemplo, con los caracteres especiales. El motor de Lexalytics devuelve un error en tiempo de ejecución y continúa el proceso sin interrumpirse. Como consecuencia se pierde el registro, por ello es necesario volver a numerar los registros. Mediante Java se reenumeran el campo TweetKey.

Transformación de datos

7 **Formateo Fecha**: mediante Java, se formatea la fecha ya que la API devuelve la fecha en 3 campos separados (año, mes, día).

8 **Elimina StopWords**: como se obtuvieron problemas a la hora de insertar directamente la lista de StopWords (especificado en la segunda fase de desarrollo), se realiza una eliminación de las mismas de manera manual.

9 **Filtro:** se eliminan las palabras etiquetadas en el paso anterior.

Enriquece registros

10 **Lector Whitelist**: se lee la Whitelist (Anexo II)

11 **Incorpora Etiquetas Whitelist:** mediante este componente se buscan en los registros y si encuentra alguna se incorpora en el campo “texttagged”.

12 **Negativo/Positivo**: mediante Java, se realiza una conversión de la calificación obtenida en el componente de enriquecimiento de texto. Si es > 0 será positiva, < 0 negativa y 0 neutral.
13 **Cadenas vacías a Null:** se llama a una clase java que pone todos los campos vacíos a null. Así se evitan problemas en la carga.
14 **CleanString:** este componente llama a una clase java que elimina los caracteres inválidos.

Carga a Endeca Server

15 **Copia a**…: copia el flujo de entrada a la salida. Normalmente es muy útil para realizar pruebas.
16 **Carga o remplaza registros**: carga los registros con los metadatos modificados en todo el proceso ETL en el servidor de Oracle Endeca.

Este proceso ETL se puede ejecutarse una vez se ha ejecutado el proceso ETL general desarrollado en la fase 2 de desarrollo.

6.4.3.2 Análisis de sentimiento

Como se quiere analizar los comentarios de las personas que hablan de las universidades es necesario realizar un análisis de sentimiento, el cual debe poder analizar la mayor parte de los datos.

En este apartado se quiere estudiar y mejorar la eficacia del componente de enriquecimiento de texto de Lexalytics. Para saber si el componente puntúa correctamente los tuits, se necesita un conjunto de datos puntuados manualmente. En este proyecto se ha utilizado un corpus de más de 7.000 tuits etiquetado a mano, TASS 2014 (56).

Este conjunto de datos son tuits desde noviembre de 2011 a marzo de 2012 de ámbito general. La polaridad de los mismos viene calificada de la siguiente manera:

Tabla 28. Relación de calificaciones de Lexalytics, TASS y el proyecto. Fuente: elaboración propia.

TASS		LEXALYTICS	ESTUDIO
Muy positivo	P+	> 0	POSITIVO
Positivo	P		
Neutral	NEU	0	NEUTRAL
Sin sentimiento	NONE		
Negativo	N	< 0	NEGATIVO
Muy negativo	N-		

El corpus proporcionado, TASS 2014 (56), utiliza una calificación discreta (evaluada por alguien de manera manual) mientras que Lexalytics (32) utiliza decimales. Para unificarlo, en el estudio se realizará una conversión según la tabla 28.

Además del conjunto de datos se utilizará una matriz de confusión (tabla 29) en la que se calificarán los resultados obtenidos después del procesamiento. Esta matriz será el cruce ente los datos de referencia, con el resultado de la clasificación en eso datos.

Tabla 29. Matriz de confusión de ejemplo utilizada para el estudio. Fuente: elaboración propia.

MATRIZ DE CONFUSIÓN			
CLASIFICADOS→	POSITIVO	NEUTRAL	NEGATIVO
POSITIVO	10	7	2
NEUTRAL	8	9	1
NEGATIVO	3	6	5

6.4.3.2.1 Diccionario de Lexalytics

Ya que se desconoce el componente y las capacidades con texto escrito informal, como son los tuits, se espera que un alto porcentaje de los tuits analizados sean neutrales ya que el 40% (19) de los tuits están calificados como “charlas sin sentido” y el 4% “spam”.

Se desarrolla un proceso ETL (figura 79) para evaluar el componente de Lexalytics. Se ha elegido utilizar Excel como fichero de salida, y no el servidor de Endeca, porque es más ágil con pocos datos. En este caso se analizarán 7.194 tuits.

Figura 79. Proceso ETL para el estudio del análisis de sentimiento.

En la tabla 30 se ha estudiado el reparto de tuits del corpus TASS 2014 (56) en las categorías positivo, negativo y neutra. Se puede observar que 2.868 son positivos, 2.144 neutrales y 2.182 negativos.

Tabla 30. Matriz de confusión de TASS 2014. Fuente: elaboración propia.

MATRIZ DE CONFUSIÓN - TASS 2014			
CLASIFICADOS→	POSITIVO	NEUTRAL	NEGATIVO
POSITIVO	2868	0	0
NEUTRAL	0	2144	0
NEGATIVO	0	0	2182

Utilizando los mismos datos de entrada y con el componente de Lexalytics (32), después del realizar el proceso ETL de la figura 76 son los siguientes.

La precisión ("accuracy", clasificados correctamente entre el total) fue del 44%, es decir, aquellos que fueron calificados correctamente (3.136 tuits).

Hubo 1.995 tuits calificados como neutrales siendo positivos y 1.447 tuits calificados como neutrales siendo negativos. Este estudio no se centrará en los neutrales ya que se considera que es menos importante que un positivo se clasifique como neutral a que un positivo se clasifique como negativo. De la misma manera con los negativos.

Tabla 31. Matriz de confusión de Lexalytics. Fuente: elaboración propia.

MATRIZ DE CONFUSIÓN - LEXALYTICS			
CLASIFICADOS→	POSITIVO	NEUTRAL	NEGATIVO
POSITIVO	734	1995	139
NEUTRAL	164	1854	125
NEGATIVO	187	1447	548

Para el estudio que se quiere realizar interesará principalmente aquellos tuits que se clasifique como positivos siendo positivos y aquellos negativos que se clasifiquen como negativos.

- **Fracción de verdaderos positivos (TPF):** relación de positivos clasificados correctamente respecto al total de positivos.

$$TPF = \frac{Positivos\ clasificados\ correctamente}{Total\ los\ positivos} = \frac{734}{734 + 1995 + 139} \approx 0{,}26$$

- **Fracción de verdaderos negativos (TNF):** relación de negativos clasificados correctamente respecto al total de negativos.

$$TNF = \frac{Negativos\ clasificados\ correctamente}{Total\ los\ negativos} = \frac{548}{187 + 1447 + 548} \approx 0{,}25$$

- **Fracción de falsos positivos (FPF):** relación entre negativos clasificados como positivos.

$$FPF = \frac{Negativos\ clasificados\ como\ positivos}{Todos\ los\ negativos} = \frac{187}{187 + 1447 + 548} \approx 0{,}09$$

- **Fracción de falsos negativos (FFN):** relación entre positivos clasificados como negativos.

$$FFN = \frac{Positivos\ clasificados\ como\ negativos}{Todos\ los\ positivos} = \frac{139}{734 + 1995 + 139} \approx 0{,}05$$

6.4.3.2.2 Mejora del diccionario

Para mejorar los resultados en la clasificación el motor Salience de Lexalytics (31) cuenta con una herramienta llamada Salience Workbench (57) versión 1.

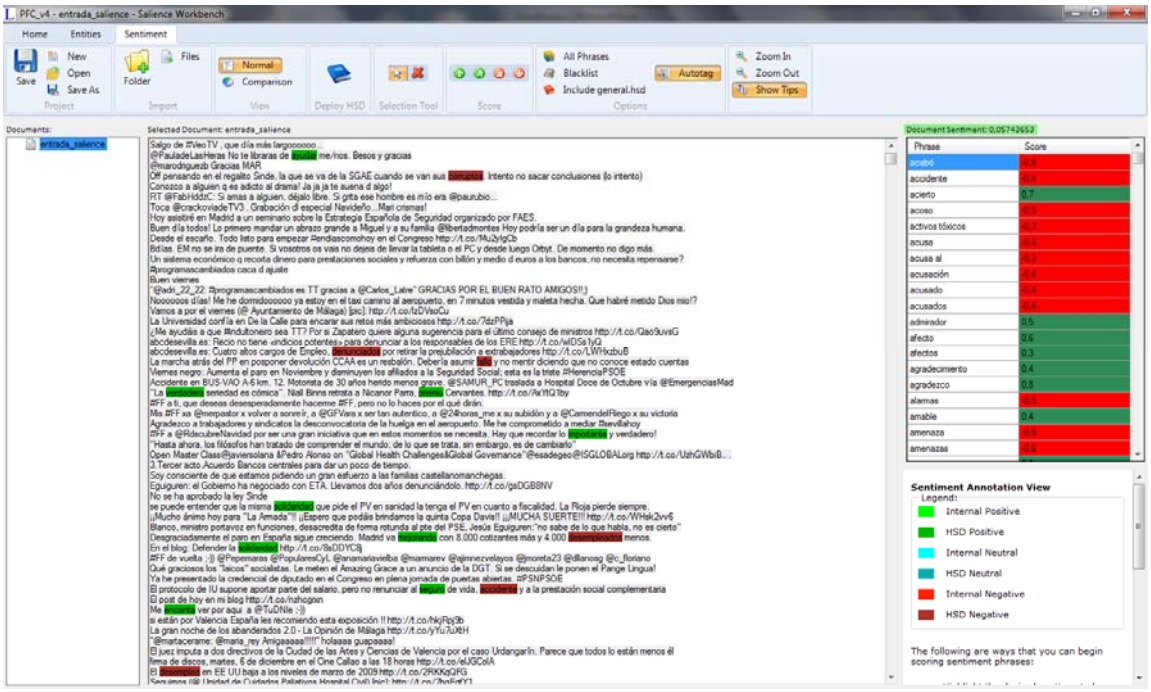


Figura 80. Salience Workbench de Lexalytics (57). Fuente: elaboración propia.

El diccionario en español es un conjunto de librerías de aproximadamente 290 MB y esta herramienta permite ajustar algunos parámetros en la calificación del análisis de sentimiento. Dentro de la librería se encuentra un fichero llamado “general.hsd” que contiene una relación de palabras con su valoración (figura 81). Este será fichero que se sustituirá o se añadirán las nuevas palabras que se hayan etiquetado en la herramienta Salience Workbench.

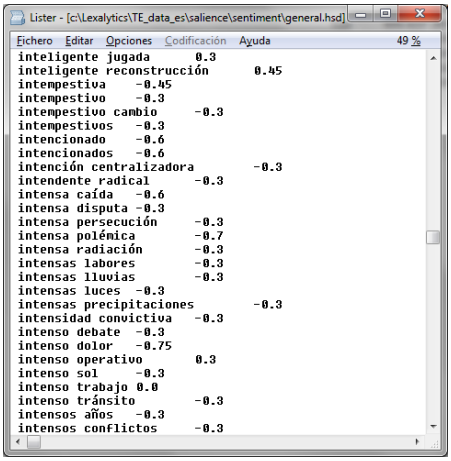


Figura 81. Fichero "general.hsd". Fuente: elaboración propia.

El proceso para valorar las palabras mediante la herramienta Salience Workbench es seleccionando la palabra e indicando la valoración que se estima (figura 82).

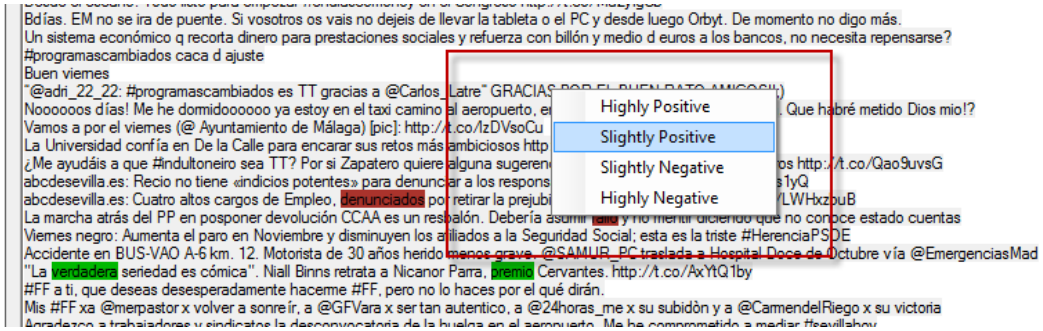


Figura 82. Clasificación de palabras en Salience Workbench. Fuente: elaboración propia.

Una vez valorada, aparece en la sección izquierda con la puntuación (figura 83).

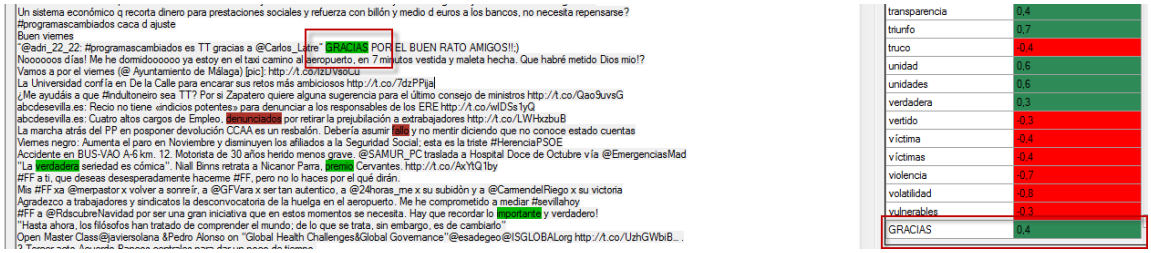


Figura 83. Clasificación manual realizada en Salience Workbench. Fuente: elaboración propia.

En lugar de una clasificación con la herramienta Salience Workbench, para este estudio se ha analizado la frecuencia de las palabras que aparecen en el corpus general de TASS 2014 (56) y se han identificado, para su posterior eliminación, aquellas las que son StopWords (figura 84). Posteriormente se han ordenado por frecuencia y se han etiquetado manualmente indicando las muy positivas 0,6, positivas 0,3, neutrales 0, negativas -0,3 y muy negativas -0,6.

	A	B	C	D
1	PALABRA	FRECUENCIA	STOPWORD	VALORACIÓN
2	=	34489021	FALSE	
3	??	25090	FALSE	
4	???	13887	FALSE	
5	?????	10651	FALSE	
6	?	7121	FALSE	
7	De	5196	TRUE	
8	La	3471	TRUE	
9	El	3179	TRUE	

Figura 84. Análisis de palabras del corpus TASS 2014 con su frecuencia y categoría. Fuente: elaboración propia.

Eliminando las stopwords, se han valorado 376 nuevas palabras (figura 85). Estas se han añadido al fichero "general.hsd" de la librería de Lexalytics.

	A	B	C	D
1	PALABRA	FRECUENCIA	STOPWORD	VALORACIÓN
39	Gracias	280	FALSE	0,3
69	Buenos	145	FALSE	0,6
133	Feliz	90	FALSE	0,6
134	Bien	89	FALSE	0,6
139	Buenas	87	FALSE	0,6
201	déficit	71	FALSE	-0,6
209	Bueno	65	FALSE	0,6
237	Buena	58	FALSE	0,6

Figura 85. Lista de palabras después de la clasificación manual. Fuente: elaboración propia.

Con la librería modificada se vuelve a ejecutar el proceso ETL (figura 76) con el componente de enriquecimiento de texto y los mismos parámetros de configuración. Los resultados son los siguientes (tabla 32).

La precisión ("accuracy", clasificados correctamente entre el total) es de un 52%, se clasificaron correctamente 3.764 tuits añadiendo 376 nuevas palabras al diccionario de datos.

Tabla 32. Matriz de confusión con la librería de Salience modificada. Fuente: elaboración propia.

MATRIZ DE CONFUSIÓN - LEXALYTICS MODIFICADO			
CLASIFICADOS→	POSITIVO	NEUTRAL	NEGATIVO
POSITIVO	1215	1456	197
NEUTRAL	288	1629	226
NEGATIVO	323	939	920

- **Fracción de verdaderos positivos (TPF):** relación de positivos clasificados correctamente respecto al total de positivos

$$TPF = \frac{Positivos\ clasificados\ correctamente}{Total\ positivos} = \frac{1215}{1215 + 1456 + 197} \approx 0{,}42$$

- **Fracción de verdaderos negativos (TNF):** relación de negativos clasificados correctamente respecto al total de negativos

$$TNF = \frac{Negativos\ clasificados\ correctamente}{Total\ negativos} = \frac{920}{323 + 939 + 920} \approx 0{,}42$$

- **Fracción de falsos positivos (FPF):** relación entre negativos clasificados como positivos.

$$FPF = \frac{Negativos\ clasificados\ como\ positivos}{Todos\ los\ negativos} = \frac{323}{323 + 939 + 920} \approx 0{,}15$$

- **Fracción de falsos negativos (FFN):** relación entre positivos clasificados como negativos.

$$FFN = \frac{Positivos\ clasificados\ como\ negativos}{Todos\ los\ positivos} = \frac{197}{1215 + 1456 + 197} \approx 0{,}07$$

En resumen, ya que Lexalytics no permite acceder ni modificar la estructura interna de sus librerías, y solo permite afinar el diccionario de datos, la mejora de la clasificación es muy complicada y lenta. Aún así, se ha conseguido mejorar la precisión un 9% llegando a clasificar correctamente el 42% de los tuits positivos y negativos. El coste de mejorar los resultados en cuanto a la clasificación tiene como consecuencia un aumento en la fracción de falsos negativos un 2% y en la fracción de falsos positivos un 6%. Esto quiere decir que mejorando la clasificación en general habrá más tuits positivos clasificados como negativos y más tuits negativos clasificados como positivos. Para este estudio compensa la ganancia en la mejora de clasificación frente al error.

En la siguiente fase, cuando se obtengan más tuits, se repetirá el proceso para obtener resultados similares o mejores.

6.4.3.3 Resultados análisis de riesgos

Los riesgos identificados en esta fase se han ido supervisando mediante las técnicas anteriormente descritas.

RIESGO-F03-01: Pérdida de datos

Se han realizado backups semanales como se había planificado.

RIESGO-F03-01	
Semana	Backup realizado
17	1
18	1
19	1
20	1

Figura 86. RIESGO-F03-01. Fuente: elaboración propia.

RIESGO-F03-02: Cantidad de datos adquiridos

En esta tercera fase se ha avanzado mucho en cuanto a la adquisición de tuits, al final de la fase se tiene el 41% de tuits del objetivo propuesto (tabla 33).

Tabla 33. RIESGO-F03-02. Fuente: elaboración propia.

RIESGO-F03-02						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
Fase 2	29.748	-	29.748			
17	15.137	18.354	44.885	36.708	18%	15%
18	13.047	18.354	69.293	55.062	28%	22%
19	24.408	18.354	90.397	73.416	36%	29%
20	21.104	18.354	103.444	91.770	41%	37%
TOTAL	73.696	73.416	103.444	250.000	41%	

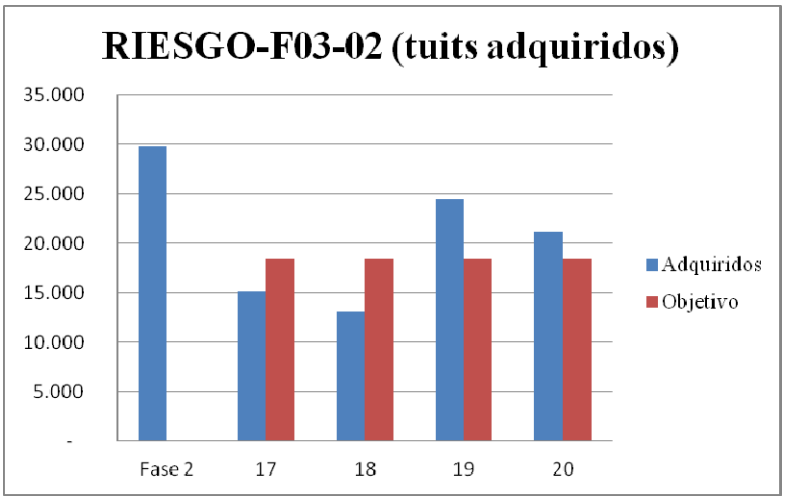


Figura 87. RIESGO-F03-02 (tuits adquiridos). Fuente: elaboración propia.

En la figura 87 se puede observar que la semana 19 y 20 se adquirieron más tuits del objetivo teórico. El crecimiento es superior al esperado (figura 88).

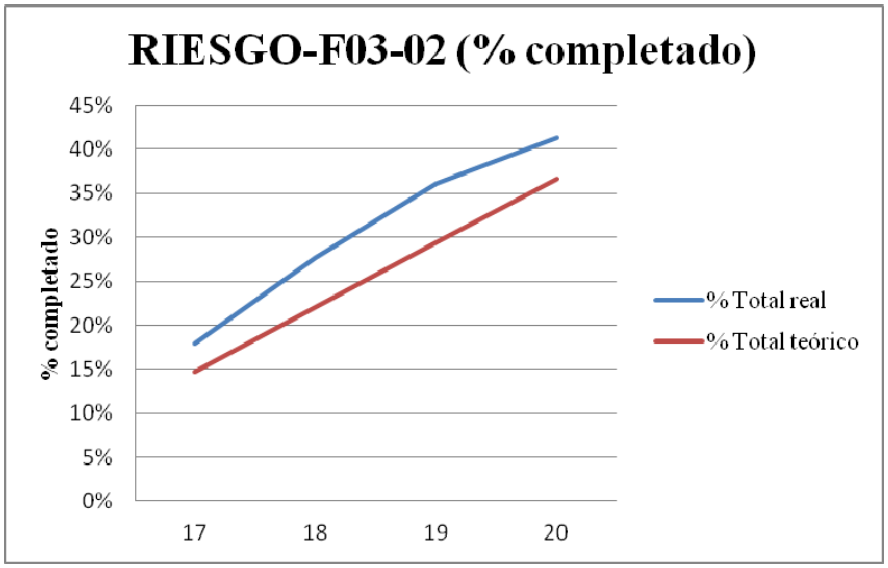


Figura 88. RIESGO-F03-02(% completado). Fuente: elaboración propia.

RIESGO-F03-03: Rendimiento del sistema integrado

En cuanto al rendimiento del sistema integrado sin arrancar el componente de Oracle Endeca Studio se pueden realizar cargas al servidor de Endeca sin consumir tantos recursos (figura 89).

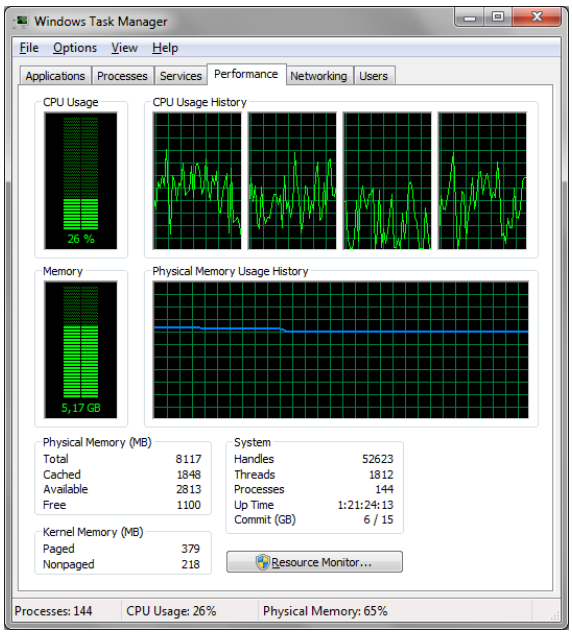


Figura 89. Gráfica de consumo de recursos durante la carga en la tercera fase. Fuente: elaboración propia.

6.4.4 Planificación

En la tercera fase se ha llevado a cabo la siguiente planificación (figura 90).

Fase 3: Análisis de sentimiento	19 days?	**8/04/14 8:00**	**2/05/14 17:00**
Determinar objetivos	3 days	**8/04/14 8:00**	**10/04/14 17:00**
Análisis de sentimiento	2 days	8/04/14 8:00	9/04/14 17:00
Modificación proceso ETL	2 days	9/04/14 8:00	10/04/14 17:00
Análisis de riesgos	1 day?	**11/04/14 8:00**	**11/04/14 17:00**
Identificación de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Análisis de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Planificación de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Supervisión de riesgos	1 day?	11/04/14 8:00	11/04/14 17:00
Desarrollo y verificación	14 days?	**14/04/14 8:00**	**1/05/14 17:00**
Modificar ETL	10 days	14/04/14 8:00	25/04/14 17:00
Análisis de sentimiento	12 days	16/04/14 8:00	1/05/14 17:00
Resultados análisis de riesgo	1 day?	1/05/14 8:00	1/05/14 17:00
Planificación	1 day?	2/05/14 8:00	2/05/14 17:00

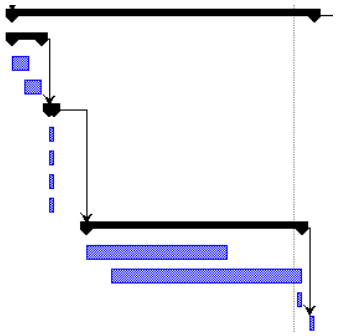


Figura 90. Planificación fase 3. Fuente: elaboración propia.

Una vez concluida esta fase, se procede a la visualización de los datos. Para ello se diseñará en la siguiente fase cuadros de mandos, se realizarán las mejoras en el análisis de sentimiento que se ha visto en esta fase y se integrarán datos estructurados sobre las universidades y alumnos.

Durante la cuarta fase se seguirá la planificación de la figura 91.

Fase 4: Visualización	20 days?	**5/05/14 8:00**	**30/05/14 17:00**
Determinar objetivos	4 days?	**5/05/14 8:00**	**8/05/14 17:00**
Integración datos universidades y alumnos	2 days	5/05/14 8:00	6/05/14 17:00
Diseño de los cuadros de mando	2 days	6/05/14 8:00	7/05/14 17:00
Análisis de sentimiento	1 day?	8/05/14 8:00	8/05/14 17:00
Análisis de riesgos	1 day?	**9/05/14 8:00**	**9/05/14 17:00**
Identificación de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Análisis de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Planificación de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Supervisión de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Desarrollo y verificación	15 days?	**12/05/14 8:00**	**30/05/14 17:00**
Análisis de sentimiento con datos de las universidades y pruebas	10 days	12/05/14 8:00	23/05/14 17:00
Desarrollo cuadro de mandos: unviersides	5 days	16/05/14 8:00	22/05/14 17:00
Desarrollo cuadro de mandos: Análisis de sentimiento	5 days	19/05/14 8:00	23/05/14 17:00
Desarrollo cuadro de mandos: usuarios	5 days	22/05/14 8:00	28/05/14 17:00
Desarrollo cuadro de mandos: mapas	4 days	26/05/14 8:00	29/05/14 17:00
Verificación y pruebas	9 days	19/05/14 8:00	29/05/14 17:00
Resultados análisis de riesgos	1 day?	30/05/14 8:00	30/05/14 17:00

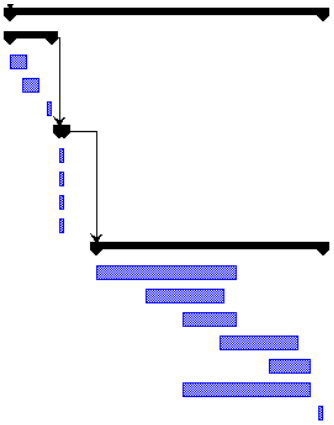


Figura 91. Planificación fase 4. Fuente: elaboración propia.

6.5 Fase 4: Visualización

Con los datos cargados en el servidor se diseñarán varios cuadros de mando desde el cual se podrá navegar por la información, consiguiendo así dar una visión global del conjunto de datos, buscar relaciones entre los mismos y poder llegar a conclusiones.

Mediante los cuadros de mando se accederá a información detallada de los usuarios, los temas sobre los que hablan, las tendencias, usuarios más influentes en la red social, etc.

Por último se procederá a una mejora en el enriquecimiento de texto mediante el procedimiento visto en la tercera fase.

6.5.1 Determinar objetivos

En la cuarta fase se han definido los siguientes objetivos.

6.5.1.1 Diseño de los cuadros de mando

Se diseñarán los siguientes cuadros de mando para poder visualizar la información:

- Cuadro de mandos sobre las universidades, temas y entidades
- Cuadro de mandos sobre usuarios y análisis de sentimiento
- Mapas

6.5.1.2 Análisis de sentimiento

Siguiendo el procedimiento de mejora en el enriquecimiento de texto, explicado en la fase 3, se utilizarán los datos ya descargados para la agregación de nuevas palabras al diccionario.

6.5.2 Análisis de riesgos

Durante el análisis de riesgos se identificarán los riesgos, se analizarán, se planificarán y por último se indicará la manera de supervisarlos.

6.5.2.1 Identificación de riesgos

En esta fase del proyecto se han identificado los siguientes riesgos.

Identificador	Riesgo	Tipo	Descripción
RIESGO-F04-01	Pérdida de datos	Proyecto	Pérdida de datos por corrupción de los mismos.
RIESGO-F04-02	Cantidad de datos adquiridos	Proyecto	No tener suficientes datos para realizar el estudio
RIESGO-F04-03	Rendimiento del sistema integrado	Producto	Capacidades de hardware

Tabla 34. Identificación de riesgos de la fase 4. Fuente: elaboración propia.

RIESGO-F04-01: Pérdida de datos

Durante el proceso ETL puede ocurrir una pérdida de datos o corrupción debido a una parada inesperada o incluso la modificación de algún parámetro podría afectar a la estructura de los metadatos y provocaría una pérdida de los datos.

RIESGO-F04-02: Cantidad de datos adquiridos

Al igual que en la tercera fase, se estima que para realizar un estudio sobre las universidades de Madrid, se necesitan 250 mil tuits aproximadamente. Por ello se analizará y se supervisará la cantidad de datos adquiridos semanalmente.

RIESGO-F04-03: Rendimiento del sistema integrado

Toda la plataforma de análisis de datos se ha desarrollado sobre una máquina virtual con capacidades limitadas. Es necesario monitorizar el rendimiento de la misma a medida que se van agregando nuevos datos al servidor de Oracle Endeca.

6.5.2.2 Análisis de riesgos

Una vez identificados los riesgos se procede al análisis, estudiando la probabilidad de que ocurran. La probabilidad se muestra en la tabla 35.

Identificador	Nombre	Probabilidad
RIESGO-F04-01	Pérdida de datos	Muy bajo
RIESGO-F04-02	Cantidad de datos adquiridos	Bajo
RIESGO-F04-03	Rendimiento del sistema integrado	Muy bajo

Tabla 35. Análisis de riesgo de la fase 3. Fuente: elaboración propia.

RIESGO-F04-01: Pérdida de datos

Al igual que en la fase 3, se ha considerado que la probabilidad es muy baja ya que se ha demostrado que se pueden realizar copias de seguridad fácilmente en local.

RIESGO-F04-02: Cantidad de datos adquiridos

Se ha aumentado la cantidad de datos necesarios por semana a 36.639, esto es debido a que aunque los resultados han sido buenos, todavía hay que alcanzar el objetivo de 250.000. En esta fase se espera mejorar considerablemente la cantidad de datos adquiridos.

RIESGO-F04-03: Rendimiento del sistema integrado

Se ha calificado como riesgo muy bajo ya que el rendimiento con del sistema completo ha sido muy bueno en la fase 3. Además en caso de verse limitadas las capacidades hardware se puede proceder a una ampliación de las mismas.

6.5.2.3 Planificación de riesgos

Con los riesgos analizados se procede a su planificación.

RIESGO-F04-01: Pérdida de datos

- **Estrategia de prevención:** se realizarán backups semanales en local.
- **Plan de contingencia:** en caso que ocurra el riesgo identificado se procederá a la restauración del último backup guardado.

RIESGO-F04-02: Cantidad de datos adquiridos

- **Estrategia de prevención:** se monitorizará la cantidad de datos semanalmente.
- **Plan de contingencia:** al utilizar la Rest API de Twitter, se puede ejecutar con más frecuencia para adquirir mayor número de tuits, siempre y cuando se hayan generado nuevos tuits.

RIESGO-F04-03: Rendimiento del sistema integrado

- **Estrategia de prevención:** mediante el gestor de tareas de Windows se puede monitorizar las capacidades del sistema. Se pueden eliminar procesos que no sean esenciales para la ejecución de la máquina.
- **Plan de contingencia:** en caso de verse el hardware totalmente insuficiente se procederá a una ampliación del mismo.

6.5.2.4 Supervisión de riesgos

Con los riesgos planificados se procede a explicar el método de supervisión.

RIESGO-F04-01: Pérdida de datos

Mediante la tabla 36 se mantendrá un control de los backups realizados semanalmente (-1 backup no realizado, 1 backup realizado, 0 backup pendiente de realizar)

Tabla 36. RIESGO-F04-01. Fuente: elaboración propia

RIESGO-F04-01	
Semana	Backup realizado
21	0
22	0
23	0
24	0

RIESGO-F04-02: Cantidad de datos adquiridos

Mediante una tabla se mantendrá monitorizado la cantidad de datos adquiridos semanalmente, así como el objetivo semanal, la cantidad de datos adquiridos en total y el objetivo teórico total de cada semana. Además se mostrará el porcentaje total real, es decir, el porcentaje de datos adquiridos del objetivo de 250 mil tuits frente al porcentaje total teórico, es decir, el porcentaje esperado de tuits adquiridos.

Tabla 37. RIESGO-F04-02. Fuente: elaboración propia.

RIESGO-F03-02						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
Fase 3	103.444	-	103.444			
21		36.639	103.444	140.083	41%	56%
22		36.639	103.444	176.722	41%	71%
23		36.639	103.444	213.361	41%	85%
24		36.639	103.444	250.000	41%	100%
TOTAL	-	146.556	103.444	250.000	41%	

La figura 92 muestra mediante un gráfico de barras la cantidad de tuits adquiridos frente al objetivo semanal. En el eje X se muestra el número de semana de la fase y en el eje Y la cantidad de tuits.

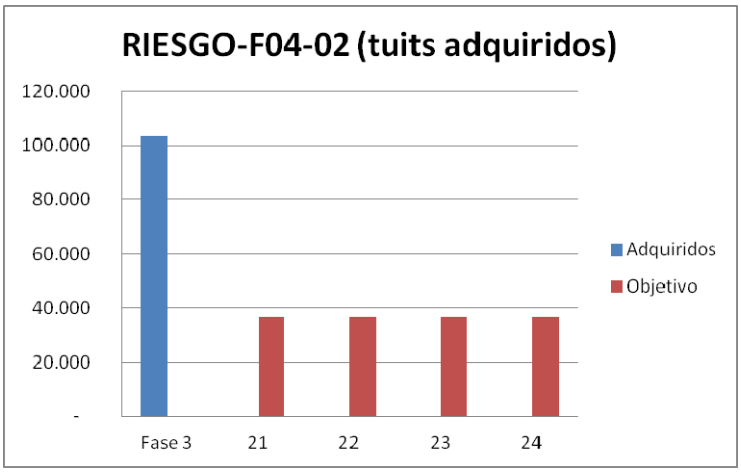


Figura 92. RIESGO-F04-02 (tuits adquiridos). Fuente: elaboración propia.

En la figura 9 se muestra la relación entre el porcentaje real de datos adquiridos frente al ideal de datos que se de3bería tener.

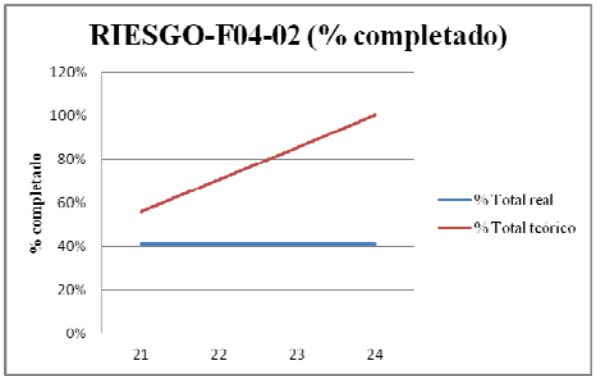


Figura 93. RIESGO-F04-02 (% completado). Fuente: elaboración propia.

RIESGO-F04-03: Rendimiento del sistema integrado

Este riesgo se supervisará de la misma manera que el RIESGO-F03-03. Mediante el administrador de tareas, a observar el uso de la CPU, el uso memoria física así como la paginación del sistema. Se sabrá que se está llegando al límite de las capacidades del ordenador cuando la interactuación con la máquina virtual sea demasiado lenta, se bloquee o no se pueda trabajar.

6.5.3 Desarrollo y verificación

Una vez analizados los riesgos de la fase y siguiendo la metodología de desarrollo en espiral, en esta fase se desarrollarán los objetivos planificados.

6.5.3.1 Análisis de sentimiento con datos de las universidades y pruebas

Utilizando la metodología vista en la tercera fase para mejorar el análisis de sentimiento, se ha estudiado la frecuencia de las palabras y se han etiquetado manualmente.

Se han etiquetado 3.254 nuevas palabras con una valoración de -0,6 para muy negativas -0,3 para negativas, 0,3 para las positivas y 0,6 las muy positivas.

Los resultados del análisis se pueden ver más adelante en la sección de resultados.

6.5.3.2 Vistas

Oracle Studio permite definir conjuntos más pequeños de datos llamados vistas. Para definir las vistas requiere realizar una consulta al servidor de Oracle en un lenguaje propio llamado EQL (Endeca Query Language), parecido a SQL. Ver anexo IV.

Durante la segunda fase, se ha definido un metadato llamado “texttagged” que contiene, en función de las palabras clave del tuit, la universidad sobre la que está hablando.

En este proyecto se han definido 16 vistas, una para cada universidad. De esta manera, utilizando una vista, por ejemplo, la vista “Universidad Francisco de Vitoria” automáticamente se filtrarán todos aquellos datos que la columna “texttagged” sea igual a “Universidad Francisco de Vitoria”. Esto será muy útil para los mapas.

En el Anexo IV se definen las vistas utilizadas.

6.5.3.3 Desarrollo cuadro de mandos: datos universidades

Para no sobrecargar la máquina virtual, se ha utilizado un modelo de diseño basado en pestañas. De esta manera cuando se carga el cuadro de mandos, no se cargan todas las gráficas al mismo tiempo (figura 94).

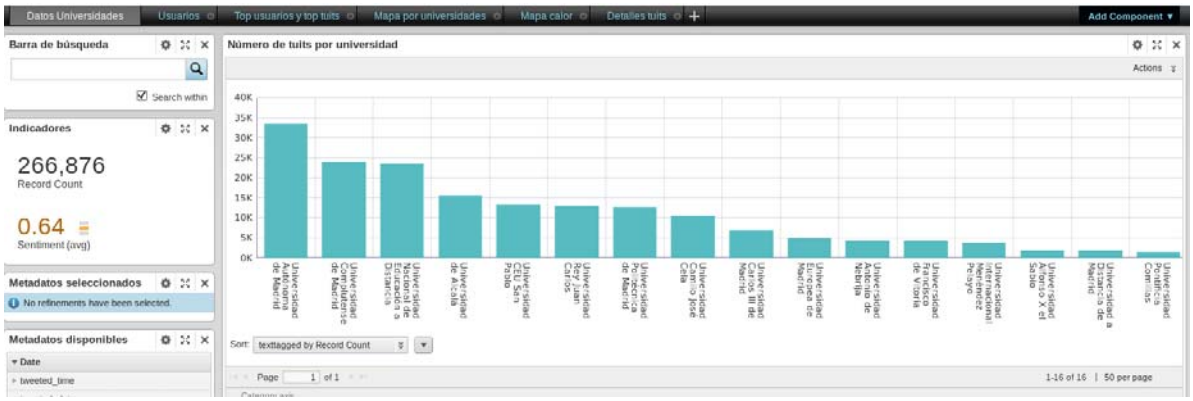


Figura 94. Cuadro de mandos pestaña datos usuario. Fuente: elaboración propia.

El primer gráfico que se crea son unos indicadores del número de tuits almacenados y la puntuación media de todos los tuits según el analizador de sentimiento. Como se puede ver se han cargado más de 260 mil tuits.

También en la figura 94 se ha creado un gráfico de barras indicando el número de tuits por universidad, en este caso, 16 universidades de la Comunidad de Madrid (43).

Para realizar estos diagramas desde Oracle Endeca Studio, se añaden componentes según las necesidades del usuario (figura 95). Posteriormente requieren configuración indicando métricas y dimensiones.

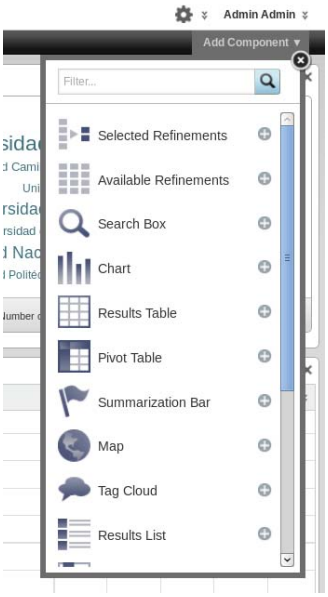


Figura 95. Componentes Oracle Endeca Studio. Fuente: elaboración propia.

Una vez seleccionado el componente se procede a su configuración. La configuración de los gráficos se basa en los siguientes apartados:

- **Selección de datos** (figura 96): se elige el conjunto de datos que se quiere mostrar. Puede ser "Base", es decir, el conjunto entero de datos o alguna de las vistas definidas previamente. Las vistas creadas permiten delimitar el conjunto de los datos a la universidad que se elija.

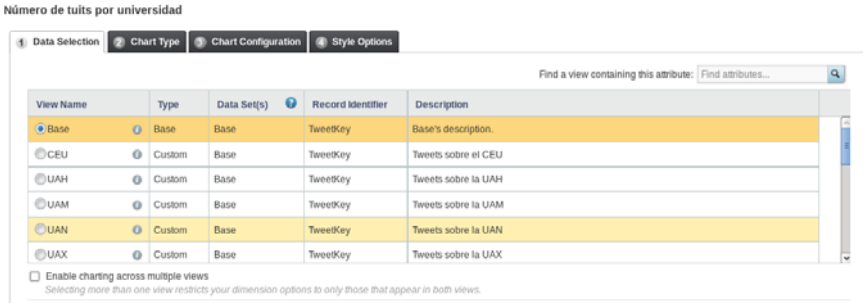


Figura 96. Selección de datos para diseñar una gráfica. Fuente: elaboración propia.

- **Selección del tipo de gráfico** (figura 97): se especifica el tipo de gráfico: barras (verticales, horizontales), lineal, barras-lineal, circular o burbujas.

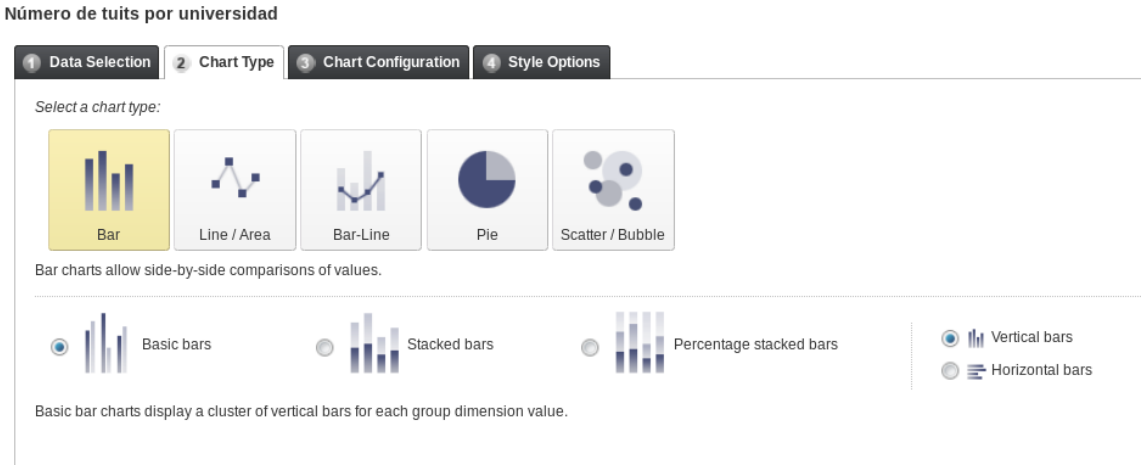


Figura 97. Selección del tipo de gráfico. Fuente: elaboración propia.

- **Configuración del gráfico** (figura 98): se especifican las métricas y dimensiones que se van a mostrar.

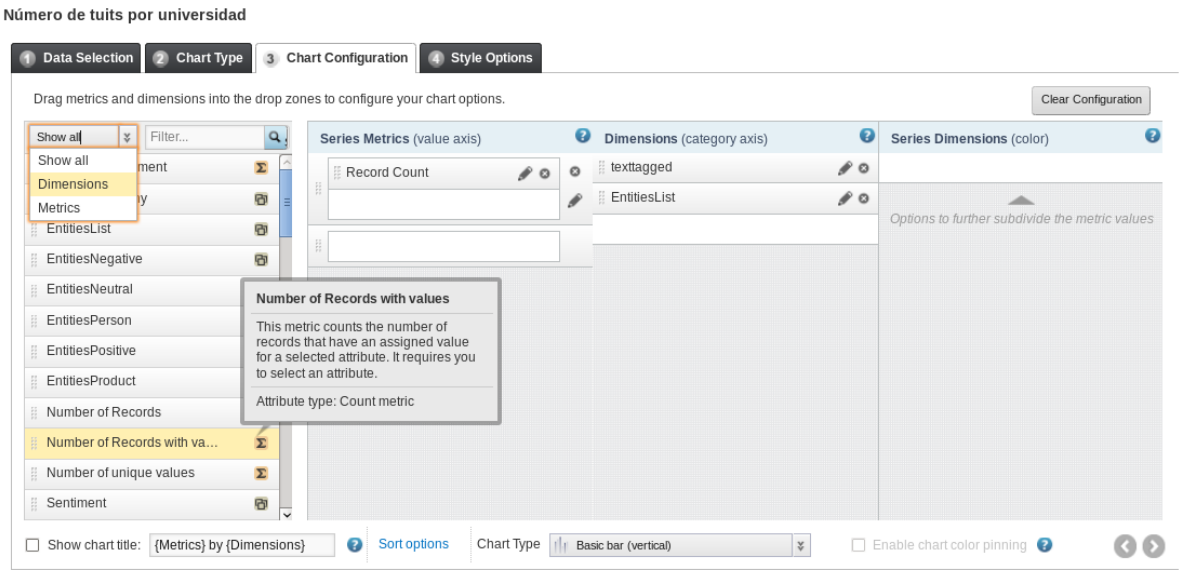


Figura 98. Configuración de un gráfico. Fuente: elaboración propia.

- **Opciones de estilo** (figura 99): se selecciona la configuración de la leyenda así como la información que aparecerá en los ejes.

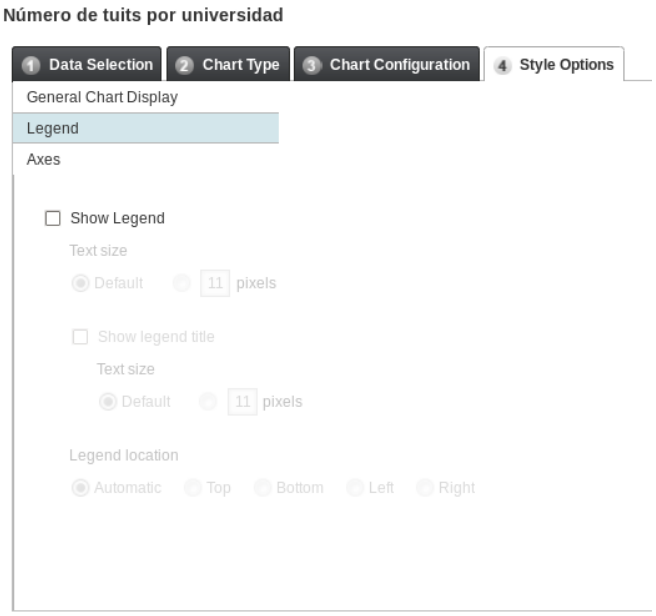


Figura 99. Opciones de estilo de un gráfico. Fuente: elaboración propia.

Después de su configuración el gráfico se muestra como el de la figura 94.

En la figura 100, mediante un diagrama de burbujas que pueden representar varias dimensiones, se ha relacionado el número de tuits (eje y), frente al sentimiento medio (eje x) por universidad. El tamaño de la burbuja indica también el número de tuits por universidad.

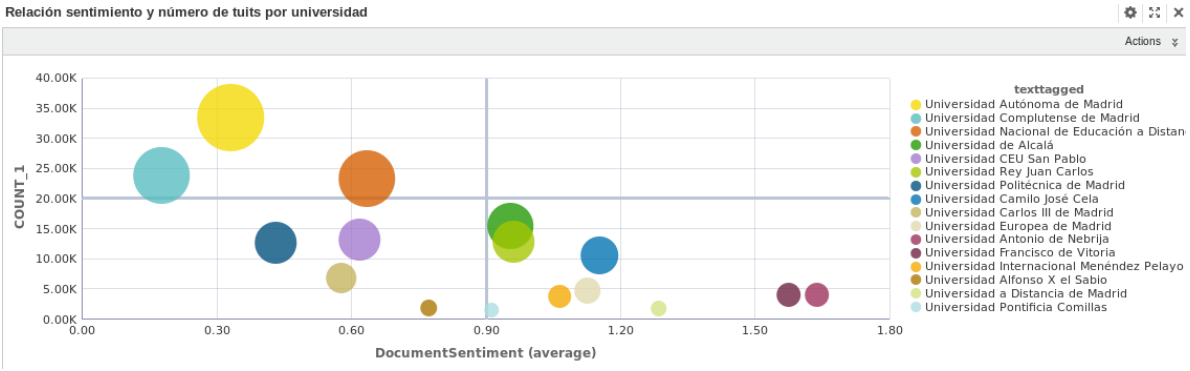


Figura 100. Relación sentimiento y número de tuits por universidad. Fuente: elaboración propia.

En la figura 101 se muestran diferentes gráficas y nubes de palabras repartidos por pestañas. La nube de palabras indica las entidades que ha reconocido el analizador de sentimiento, indicando por ejemplo el que el incidente del "sótano de los horrores" (58) de la Universidad Complutense fue muy comentado. Además en el gráfico de la derecha se observa como la mayoría de los tuits son neutrales y que no aportan valor de manera desagregada.

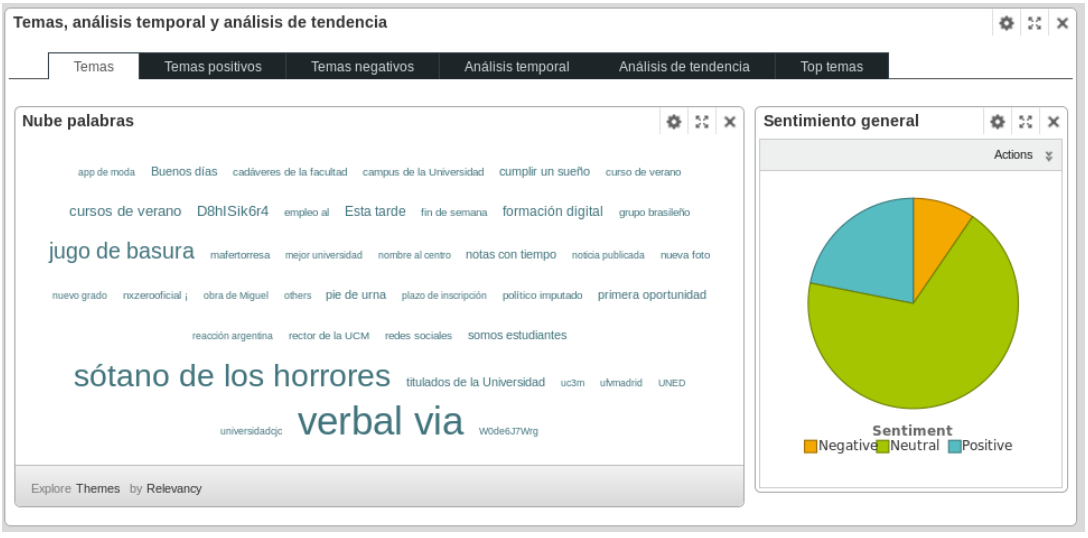


Figura 101. Temas extraídos. Fuente: elaboración propia.

En la figura 102, se ha realizado una nube de palabras o temas positivos identificados por el analizador de sentimiento.

Figura 102. Temas positivos. Fuente: elaboración propia.

En el caso de la figura 103 se ha optado por una nube de palabras con los temas negativos.

Figura 103. Temas negativos. Fuente: elaboración propia.

También se ha realizado un análisis temporal (figura 104), indicando el número de tuits (eje y de la derecha) por día y el sentimiento medio de los tuits descargados (eje y de la izquierda).

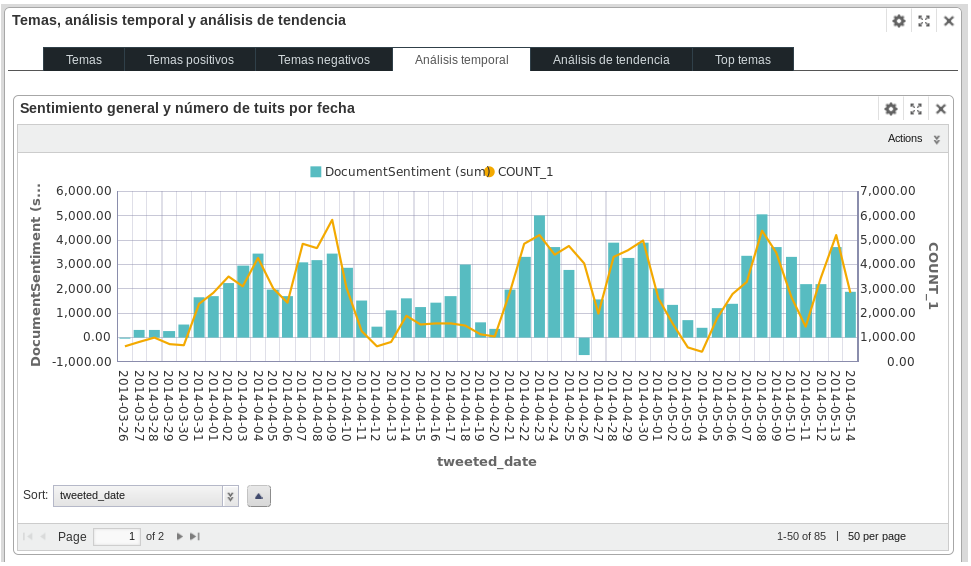


Figura 104. Sentimiento general y número de tuits por fecha. Fuente: elaboración propia.

En el caso de la figura 105, se ha relacionado el número de tuits de un usuario con su sentimiento medio. Esto puede dar un claro ejemplo de qué usuario comenta más positivamente sobre un tema en específico.

Figura 105. Relación sentimiento y número de tuits por tema. Fuente: elaboración propia.

En la figura 106, se muestra una lista de los temas positivos y negativos más frecuentes.

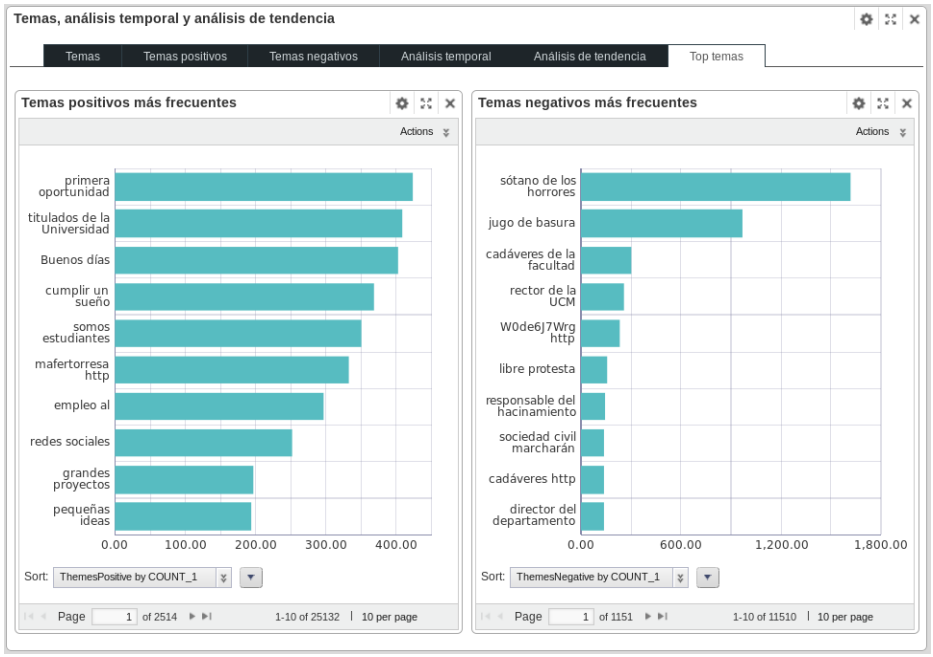


Figura 106. Temas frecuentes positivos y negativos. Fuente: elaboración propia.

6.5.3.4 Desarrollo cuadro de mandos: usuarios

Además de la pestaña creada sobre datos de las universidades se ha creado una específica para analizar los datos sobre los usuarios.

Nada más acceder se ha creado unos indicadores resumen del número de registros y el sentimiento medio. Además de la gráfica de tuits positivos, negativos y neutrales (figura 107).

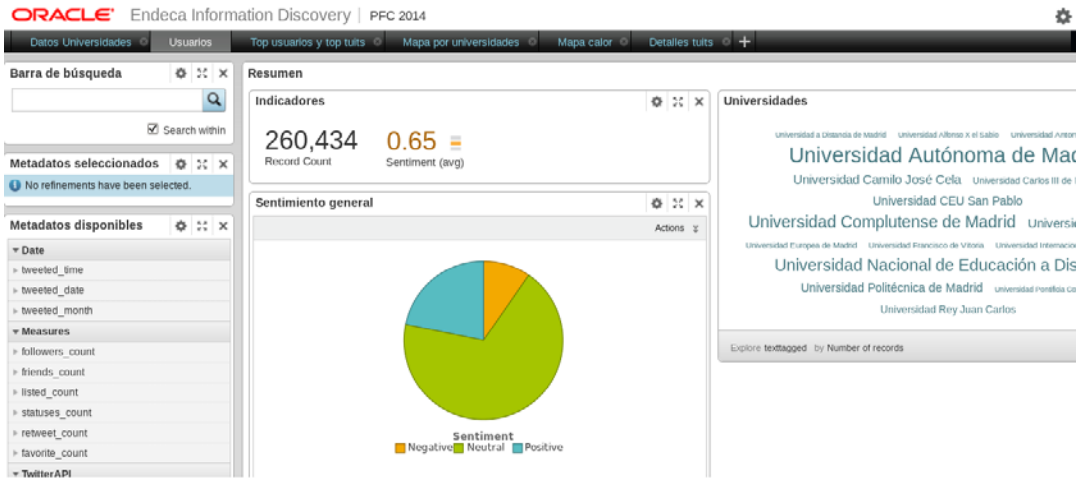


Figura 107. Pestaña usuarios. Fuente: elaboración propia.

En la figura 108 se muestra la relación entre tuits enviados (eje x) y el número de amigos (en el eje y, el número de personas que sigue) que tiene cada usuario. La mayoría de los usuarios

no siguen a más de 17 mil personas, salvo excepciones como el usuario “hay que saberlo” (a la derecha en rosa).

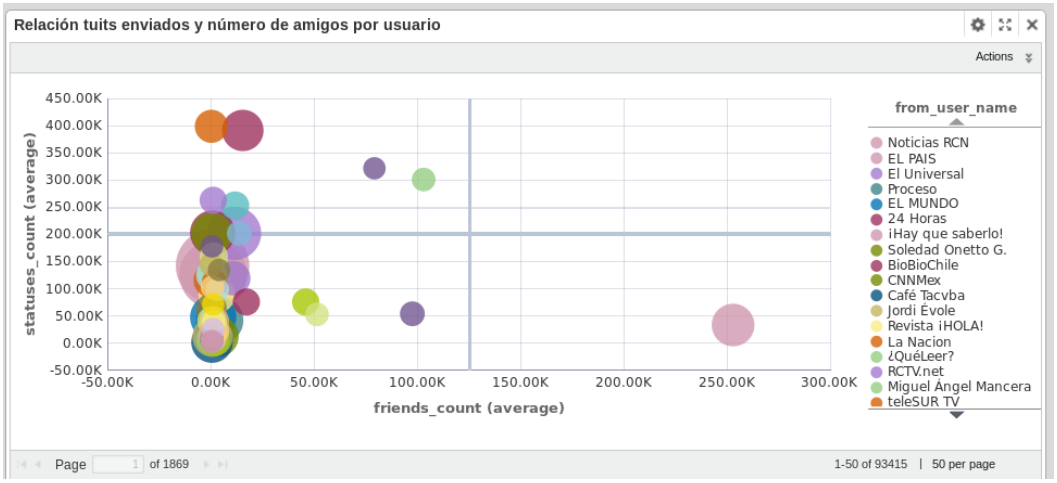


Figura 108. Relación tuits enviados y número de amigos por usuario. Fuente: elaboración propia.

Se ha optado por utilizar pestañas, para no sobrecargar el servidor con muchas consultas al mismo tiempo. En la figura 109, se muestra una nube de palabras de temas positivos y temas negativos.

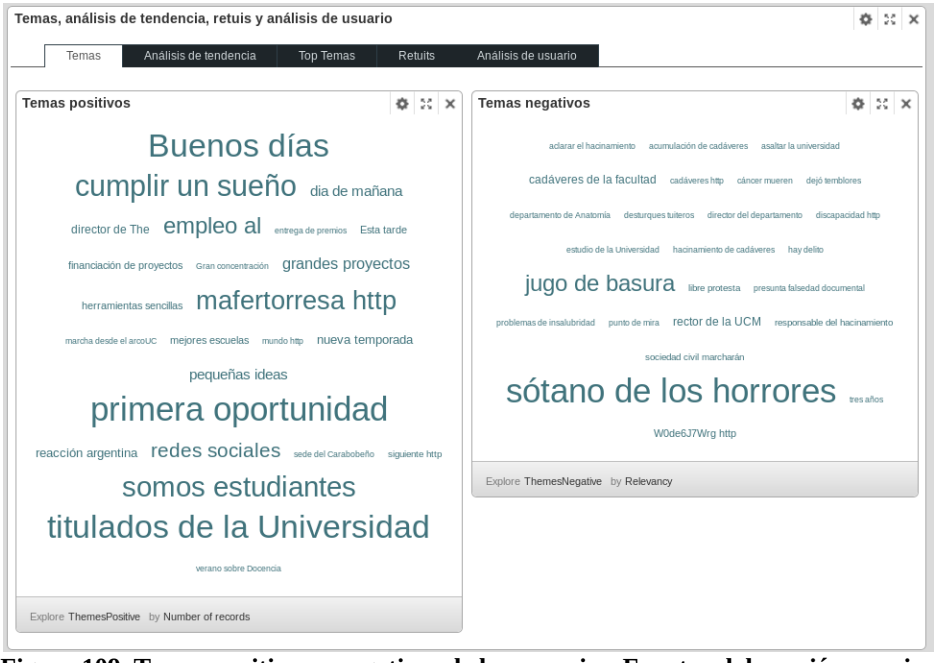


Figura 109. Temas positivos y negativos de los usuarios. Fuente: elaboración propia.

Mientras que en la figura 110, se puede ver cuántas veces se ha mencionado un tema específico y qué valoración tiene, por ejemplo “buenos días” (en azul a la derecha) tiene una

puntuación muy elevada con pocos comentarios (unos 400) frente a “sótano de los horrores” (en verde a la izquierda).

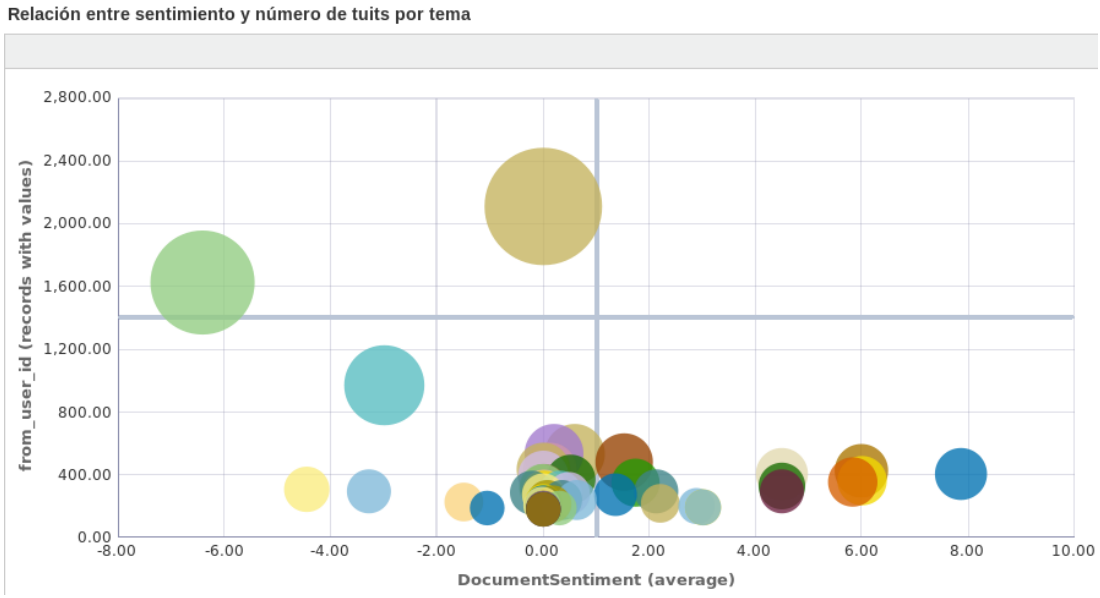


Figura 110. Análisis de tendencia de los usuarios. Fuente: elaboración propia.

En la figura 111 se analiza los diez temas más positivos y negativos.

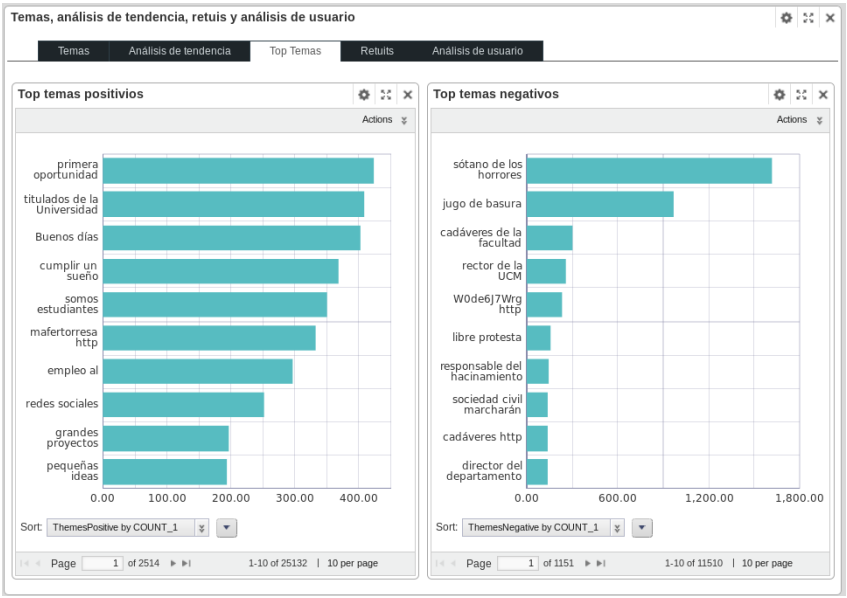


Figura 111. Top temas usuarios. Fuente: elaboración propia.

Las siguientes gráficas de la figura 112 muestran los tuits más retuiteados, en este caso es el de la “Universidad Europea” (figura 113).

Figura 112. Retuits. Fuente: elaboración propia.

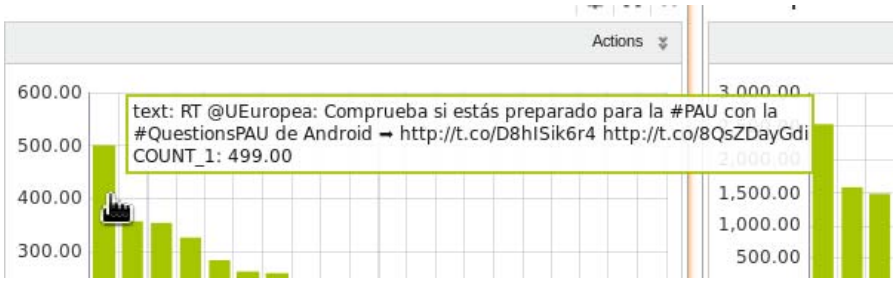


Figura 113. Tuit más retuiteado. Fuente: elaboración propia.

La figura 114 muestra información de los usuarios, por ejemplo, desde qué plataforma han emitido el tuit (arriba a la izquierda en la figura), cuántos seguidores tienen (arriba a la derecha en la figura), cuántos tuits envían (abajo a la derecha en la figura) o incluso la localización que marcan (abajo a la derecha de la figura). Hay que tener en cuenta que lo más preciso en cuanto a la localización es la posición GPS que se verá en la pestaña de mapas. La localización de la gráfica la introduce el usuario de Twitter en la configuración de la cuenta.

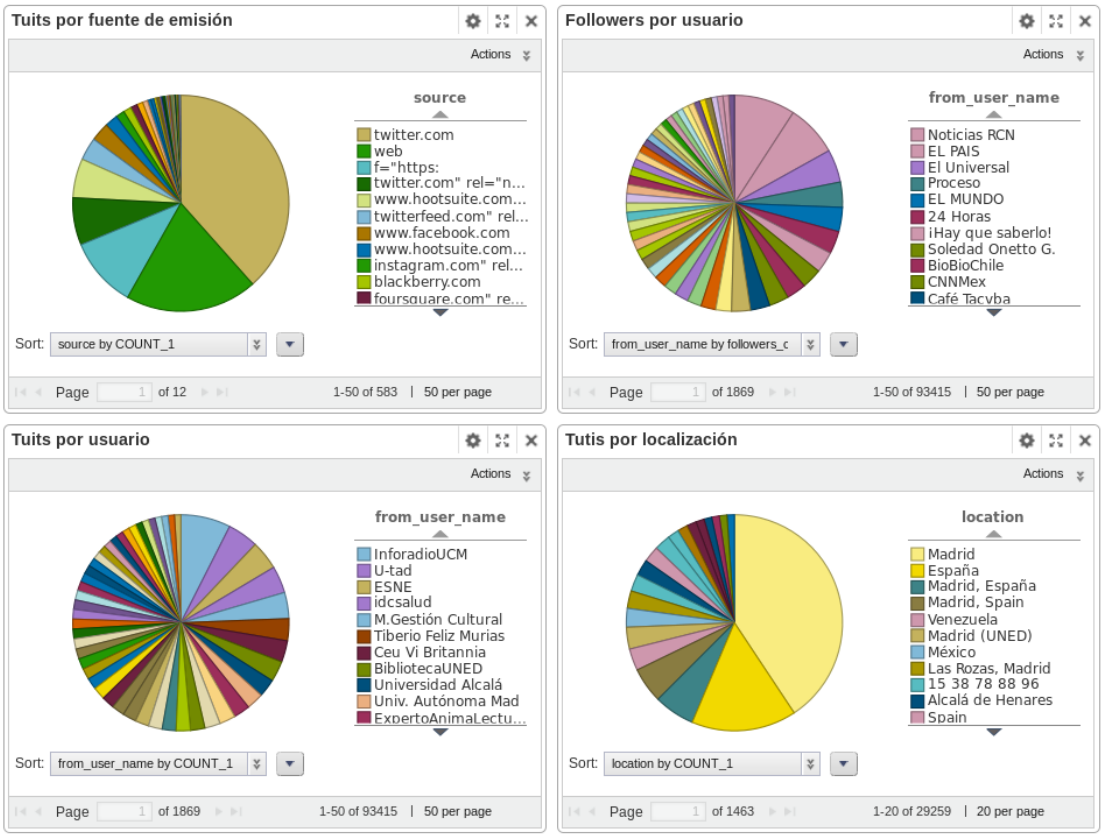


Figura 114. Análisis de usuarios. Fuente: elaboración propia.

6.5.3.5 Desarrollo cuadro de mandos: mapas

Para el desarrollo del mapa, como se ha mencionado anteriormente se hace uso de las vistas. Se va a representar los tuits correspondientes a cada universidad por medio de capas de puntos, por tanto se han implementado 16 capas de datos (figura 115).

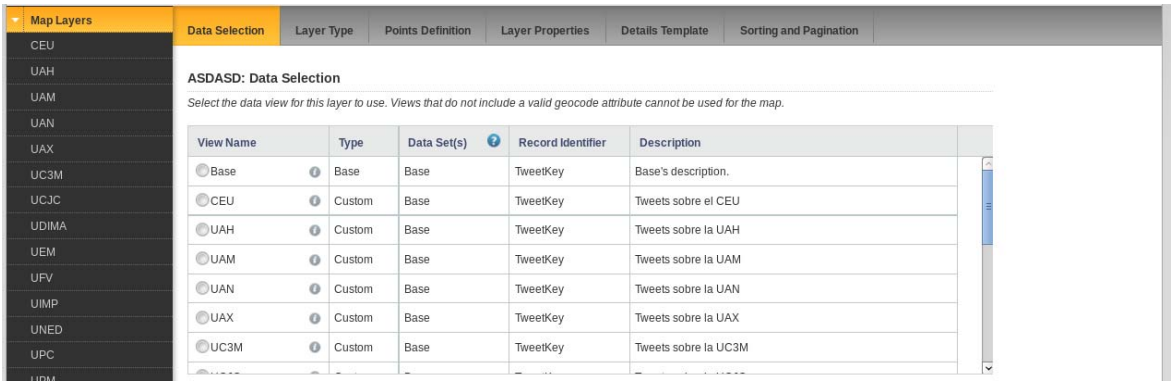


Figura 115. Capas creadas para los mapas. Fuente: elaboración propia.

Una vez cargue el mapa (figura 116) aparecen las capas y se puede seleccionar qué mapa se quiere mostrar, además a la derecha aparecerá el texto del tuit.

Figura 116. Capas de los mapas. Fuente: elaboración propia.

6.5.3.6 Resultados análisis de riesgos

Los riesgos identificados en esta fase se han ido supervisando y los resultados son los siguientes.

RIESGO-F04-01: Pérdida de datos

Se han realizado backups semanales como se había planificado.

RIESGO-F04-01	
Semana	Backup realizado
21	1
22	1
23	1
24	1

Figura 117. RIESGO-F04-01. Fuente: elaboración propia.

RIESGO-F04-02: Cantidad de datos adquiridos

En esta cuarta fase se ha avanzado mucho en cuanto a la adquisición de tuits, al final de la fase se tiene el 88% de tuits del objetivo propuesto (tabla 38). Ya que el proceso está automatizado, durante los días siguientes se incorporarán nuevos datos.

Tabla 38. RIESGO-F04-02. Fuente: elaboración propia.

RIESGO-F04-02						
Semana	Adquiridos (tuits)	Objetivo (tuits)	Adquiridos Total (tuits)	Objetivo Total (tuits)	% Total real	% Total teórico
Fase 3	103.444	-	103.444			
21	25.172	36.639	128.616	140.083	51%	56%
22	22.740	36.639	170.574	176.722	68%	71%
23	41.958	36.639	198.185	213.361	79%	85%
24	27.611	36.639	220.925	250.000	88%	100%
TOTAL	117.481	146.556	220.925	250.000	88%	

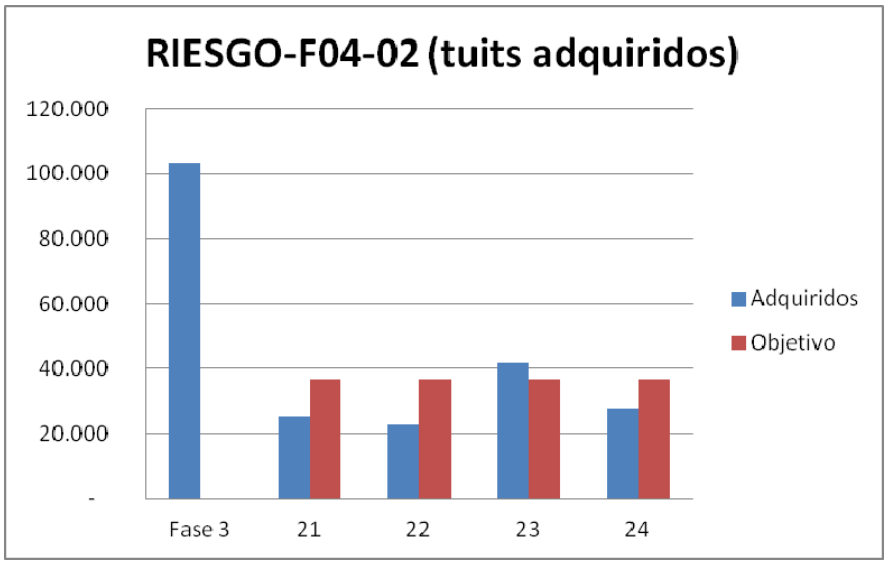


Figura 118. RIESGO-F04-02 (tuits adquiridos). Fuente: elaboración propia.

En la figura 118 se puede observar que la semana 23 se obtuvo más tuits del objetivo teórico. El crecimiento es superior al esperado (figura 119).

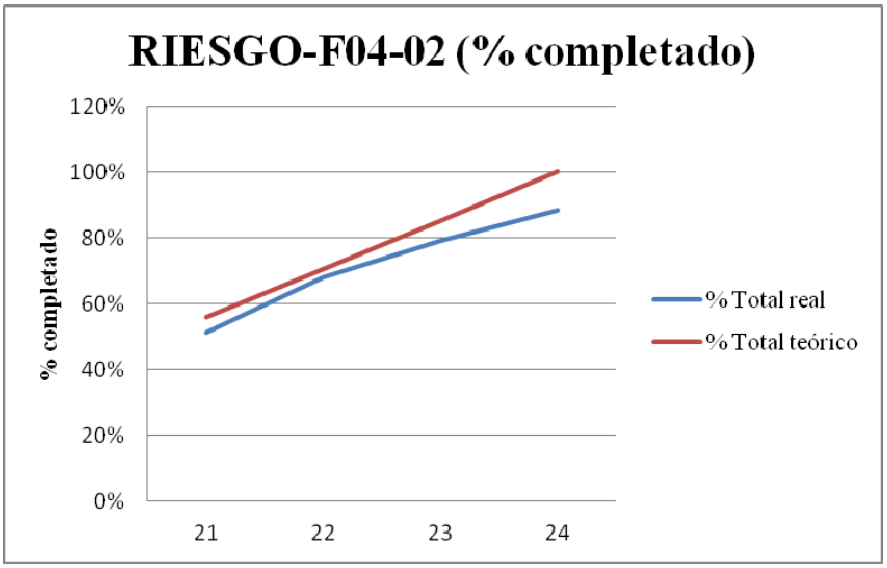


Figura 119. RIESGO-F03-02 (% completado). Fuente: elaboración propia.

RIESGO-F04-03: Rendimiento del sistema integrado

Durante el trabajo con Oracle Endeca Studio el sistema fue lento y complicado de trabajar, por ello para el diseño de los gráficos se utilizaron pestañas ya que cuando se carga una página se envían las consultas al servidor de Endeca a la vez.

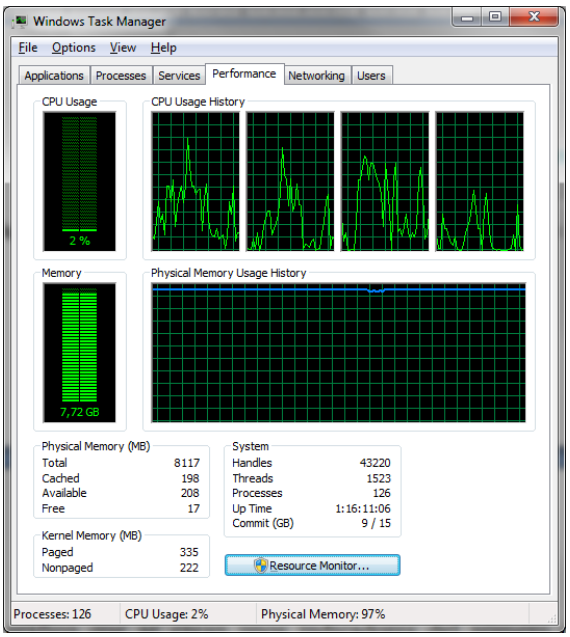


Figura 120. Gráfica de consumo de recursos durante la carga en la cuarta fase. Fuente: elaboración propia.

6.5.4 Planificación

En la cuarta fase se ha seguido planificación de la figura 121.

Fase 4: Visualización	**20 days?**	**5/05/14 8:00**	**30/05/14 17:00**
Determinar objetivos	**4 days?**	**5/05/14 8:00**	**8/05/14 17:00**
Integración datos universidades y alumnos	2 days	5/05/14 8:00	6/05/14 17:00
Diseño de los cuadros de mando	2 days	6/05/14 8:00	7/05/14 17:00
Análisis de sentimiento	1 day?	8/05/14 8:00	8/05/14 17:00
Análisis de riesgos	**1 day?**	**9/05/14 8:00**	**9/05/14 17:00**
Identificación de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Análisis de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Planificación de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Supervisión de riesgos	1 day?	9/05/14 8:00	9/05/14 17:00
Desarrollo y verificación	**15 days?**	**12/05/14 8:00**	**30/05/14 17:00**
Análisis de sentimiento con datos de las universidades y pruebas	10 days	12/05/14 8:00	23/05/14 17:00
Desarrollo cuadro de mandos: unviersides	5 days	16/05/14 8:00	22/05/14 17:00
Desarrollo cuadro de mandos: Análisis de sentimiento	5 days	19/05/14 8:00	23/05/14 17:00
Desarrollo cuadro de mandos: usuarios	5 days	22/05/14 8:00	28/05/14 17:00
Desarrollo cuadro de mandos: mapas	4 days	26/05/14 8:00	29/05/14 17:00
Verificación y pruebas	9 days	19/05/14 8:00	29/05/14 17:00
Resultados análisis de riesgos	1 day?	30/05/14 8:00	30/05/14 17:00

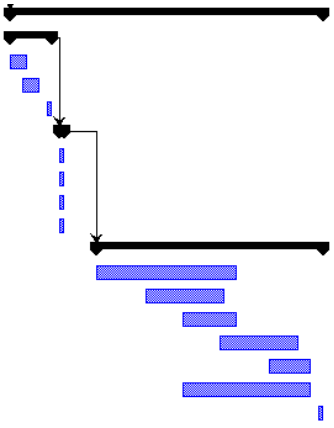


Figura 121. Planificación cuarta fase. Fuente: elaboración propia.

6.6 Resultados obtenidos

Después de más de 6 meses de desarrollo, se ha conseguido una plataforma que combina más de 260 mil tuits de las universidades de la Comunidad de Madrid. El proceso de adquisición comenzó el 26 de Marzo de 2014 y finalizó el 21 de Junio de 2014. Para ello se han integrado datos estructurados sobre universidades y usuarios, con datos no estructurados como los comentarios publicados en Twitter. Además se ha realizado un enriquecimiento de texto mediante extracción de entidades y análisis de sentimiento. Todo ello siguiendo una metodología de desarrollo en espiral a lo largo de las cuatro fases explicadas anteriormente.

Este apartado de resultados se divide en 18 puntos. El primero pretende mostrar cómo el usuario puede ir navegando por la aplicación haciéndose preguntas al momento y buscando las respuestas, incluso encontrará respuestas a preguntas que desconocía. El segundo punto es una comparativa de las universidades analizadas. En los siguientes 16 puntos se realiza un análisis por cada universidad.

6.6.1 Navegación natural

En este apartado se muestra la típica navegación natural que un usuario realiza cuando entra en la aplicación buscando nuevas preguntas.

Nada más acceder a la aplicación final se muestra el número de tuits que se han cargado y una gráfica sobre cómo están repartidos por universidad (figura 122).

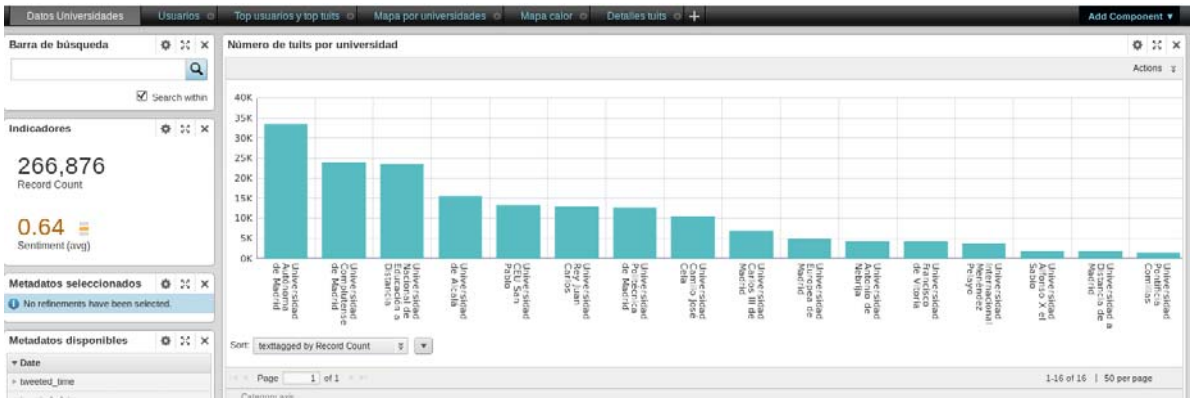


Figura 122. Pantalla inicial "Datos Universidades". Fuente: elaboración propia.

Si seleccionamos la Universidad Francisco de Vitoria automáticamente se filtra la aplicación con los datos referentes a la universidad.

Al filtrar se ve (figura 123) que se han recogido 4.095 tuits sobre la universidad y la calificación media es muy positiva 1,57 (a partir de 0 es positiva).

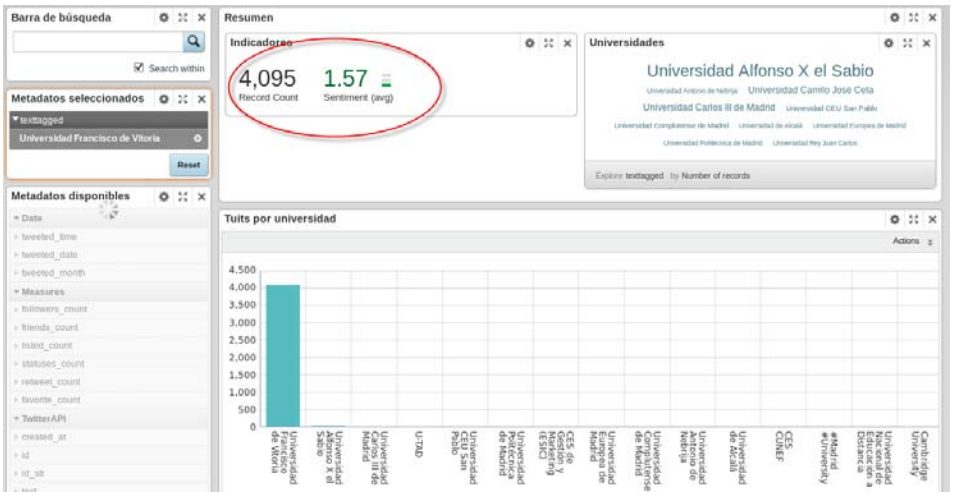


Figura 123. Cuadro de mando con el filtro "Universidad Francisco de Vitoria" aplicado. Fuente: elaboración propia.

En la figura 124 se muestra la nube de palabras "Temas positivos" con aquellos temas que en analizador de sentimiento ha calificado como positivos.

Figura 124. Nube de palabras de temas positivos sobre la Universidad Francisco de Vitoria. Fuente: elaboración propia.

Seleccionando en “alumna periodismo” (figura 124) y desplazándose al final del cuadro de mando, se puede acceder automáticamente a la pieza de información (figura 125).

source: twitter.com
from_user_name: coldwind DocumentSentiment: 1.50
text: RT @C_dPaz: Una alumna de periodismo de la @ufvmadrid acompaña a la @policia en la reciente operación antiterrorista http://t.co /13bZLBJuH6...
created_at: 2014-06-17T05:59:33+02:14
followers_count: 25

Figura 125. Tuit seleccionando “alumna periodismo”. Fuente: elaboración propia.

También vemos que con ese nuevo filtro (Universidad Francisco de Vitoria y alumna de periodismo) existen 29 tuits con una clasificación positiva (figura 126).

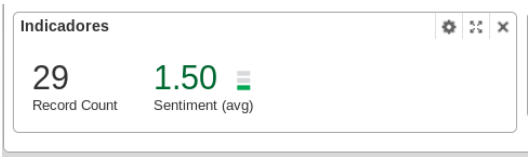


Figura 126. Resultado después de aplicar dos filtros. Fuente: elaboración propia.

Según muestra la figura 125 una alumna de periodismo acompañó a la policía en una operación antiterrorista, ¿cómo hubiésemos sabido esto?

Cambiando de tema decidimos fijarnos en el término “accidente grave” de la nube de palabras negativas. Automáticamente se filtra el cuadro de mando.

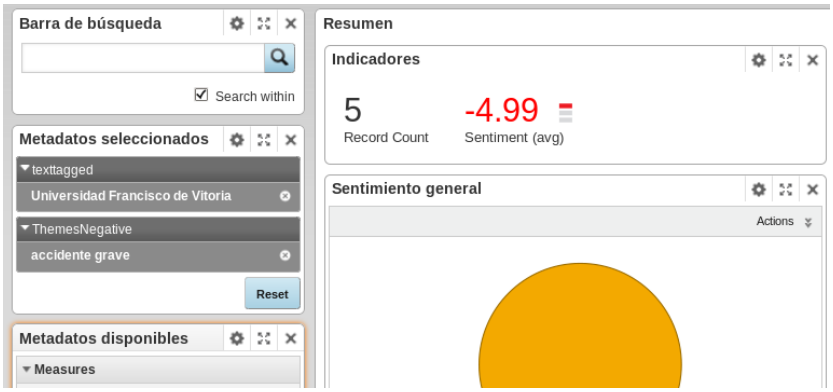


Figura 127. Filtro “accidente grave” y “Universidad Francisco de Vitoria” aplicado. Fuente: elaboración propia.

Parece que se refiere al atasco que se formó el día 7 de abril en la A6 (figura 128).

source: www.hootsuite.com" rel= nofollow >HootSuite<
from_user_name: Onda Universitaria DocumentSentiment: -4.99
text: ¡Alumnos @ufvmadrid! Evitar A6 majadahonda-madrid. accidente grave y retenciones de más de una hora vía @anavallepro
created_at: 2014-04-07T05:59:50+02:14
followers_count: 558

Figura 128. Twitter de "Onda Universitaria". Fuente: elaboración propia.

¿Cuáles son los usuarios que más comentan sobre la Universidad Francisco de Vitoria?
Accedemos al cuadro de mando "tuits por usuario" (figura 129) y podemos ver que son la Universidad Francisco de Vitoria, seguido por UFV Business, Mirada21.es, etc. También aparece un usuario con el nombre de Marta_UFV. Vamos a analizar qué comenta el usuario Marta_UFV.

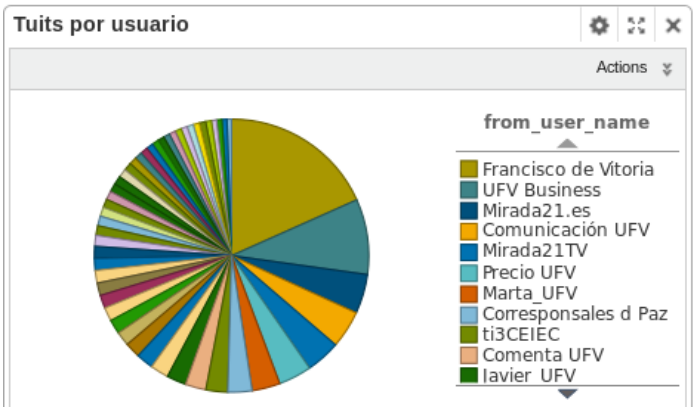


Figura 129. Tuits por usuario. Fuente: elaboración propia.

Vamos a filtrar todos los datos por Marta_UFV. Encontramos 328 registros (figura 130), muchos positivos.

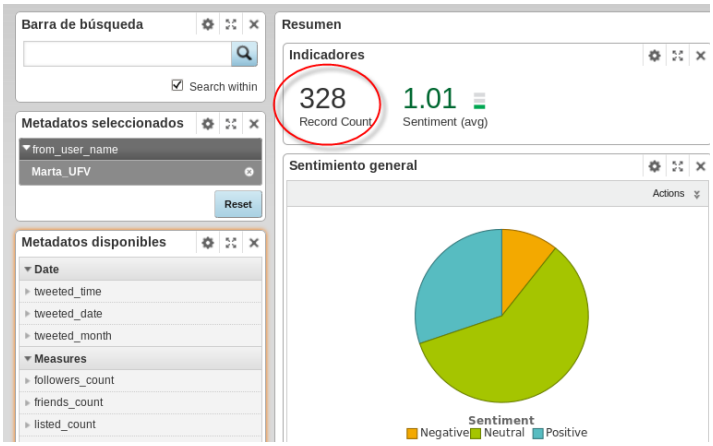


Figura 130. Filtro Marta_UFV aplicado. Fuente: elaboración propia.

Si profundizamos en los temas que Marta_UFV comenta por Twitter se obtienen los siguientes (figura 131).

Figura 131. Temas más comentados de Marta_UFV. Fuente: elaboración propia.

¿Desde dónde se habla de la Universidad Francisco de Vitoria?

Para ver desde dónde se publican los tuits que hablan de la Universidad Francisco de Vitoria se accede al mapa (figura 132).

Figura 132. Mapa con los tuits de la Universidad Francisco de Vitoria. Fuente: elaboración propia.

Se puede observar que los usuarios de Twitter que utilizan el término “ufv” se encuentran en España, Méjico y Brasil. Podemos aplicar zoom hasta mostrar solo mostrar los de España (figura 133).

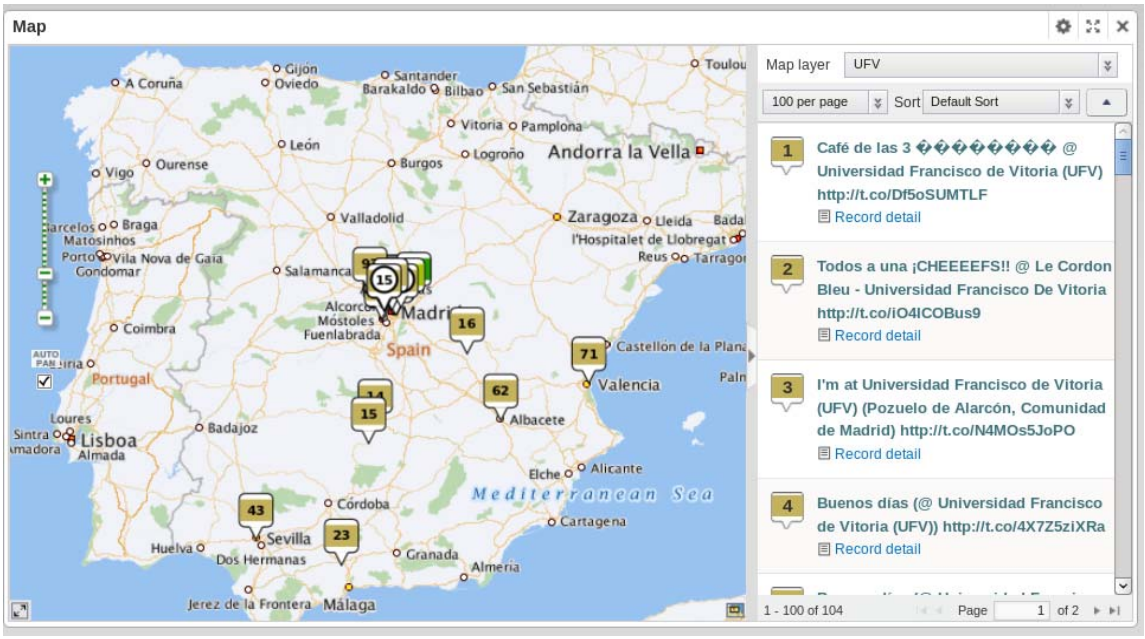


Figura 133. Ubicación de los tuits que hablan sobre la Universidad Francisco de Vitoria. Fuente: elaboración propia.

También se puede hacer zoom en Madrid y se obtiene el siguiente mapa (figura 134).

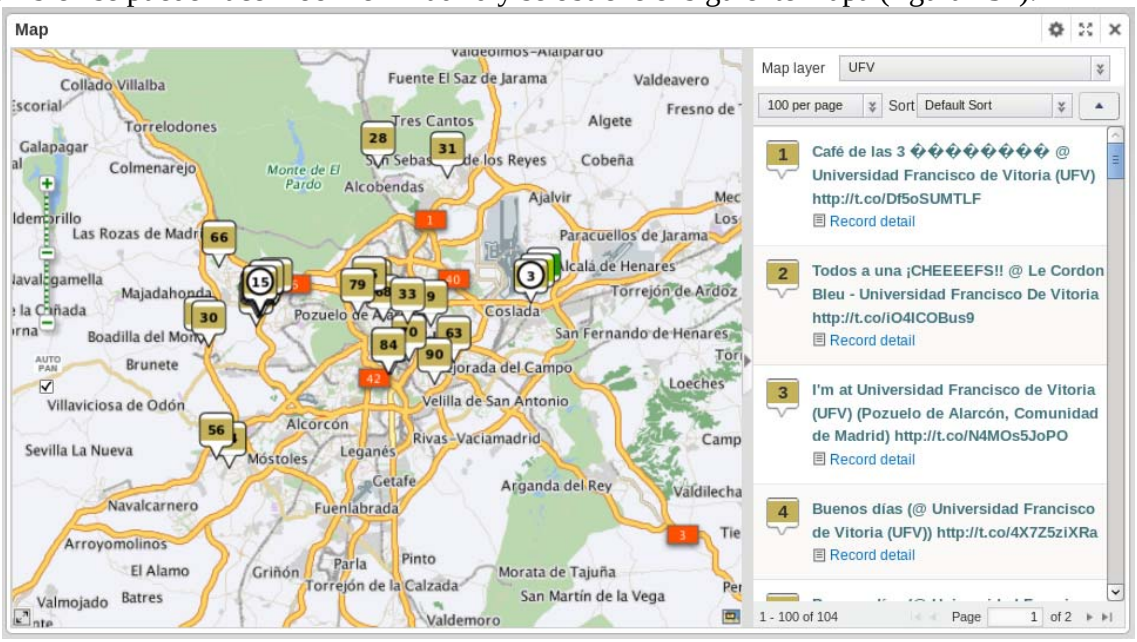


Figura 134. Ubicación tuits Universidad Francisco de Vitoria en Madrid. Fuente: elaboración propia.

Se podría acceder al contenido del tuit, por ejemplo, el número 2 de la figura 134 habla sobre "Le Cordon Bleu". Se podría hacer un estudio de la influencia del "Cordon Bleu".

¿Cuáles son las universidades mejor valoradas en relación con su número de tuits?
Al ver la siguiente gráfica (figura 135), vemos que la Universidad Antonio de Nebrija con un número de tuits muy parecido a la Universidad Francisco de Vitoria (tamaño del círculo) está un poco mejor puntuada. Vamos a intentar analizar qué comentarios han influido más en esta puntuación.

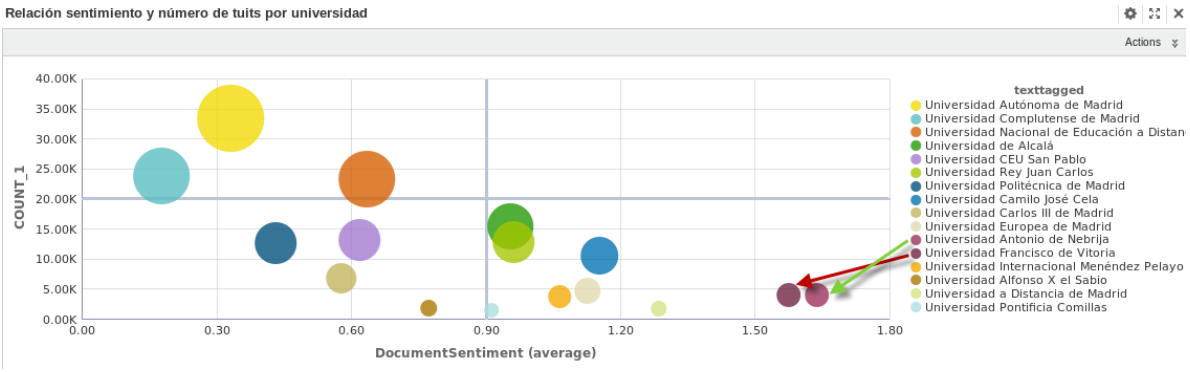


Figura 135. Relación sentimiento y número de tuits por universidad. Fuente: elaboración propia.

Accedemos a la universidad seleccionando sobre la universidad y vemos que en la nube de temas positivos (figura 136) tiene una etiqueta sobre un premio nacional.

Figura 136. Temas positivos Universidad Antonio de Nebrija. Fuente: elaboración propia.

Aplicando el filtro accedemos a la pieza de información. Parece que se hizo un concurso "Tourism Experience". Se tendría que buscar por internet en qué consiste ese concurso.

Figura 137. Tuit sobre el premio Nebrija Tourism Experience. Fuente: elaboración propia.

¿Cuáles son los temas más negativos referentes a las universidades?

Se detectan 24.897 tuits negativos, el incidente más renombrado es el comentado anteriormente sobre la Universidad Complutense y el "sótano de los horrores" (58)

Figura 138. Temas negativos. Fuente: elaboración propia.

Filtrando por la etiqueta "sótano de los horrores" hay 1.619 tuits (figura 139) hablando sobre ese tema (58), todos referentes a la Universidad Complutense de Madrid.

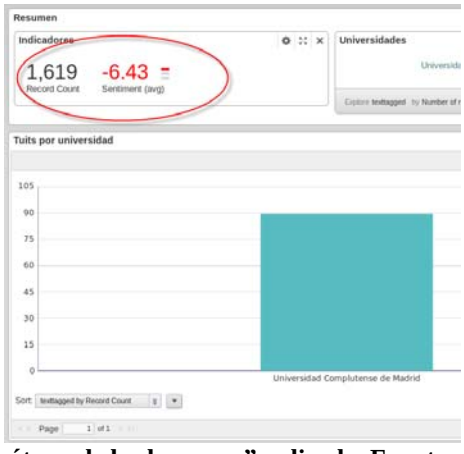


Figura 139. Filtro "sótano de los horrores" aplicado. Fuente: elaboración propia.

Se puede llegar a la propia pieza de información, el tuit, para leer exactamente lo que dicen los usuarios (figura 140).

Figura 140

¿Qué centro asociado tiene más comentarios y por qué?

Si añadimos los centros asociados, se agregan también los centros asociados a las universidades. Aparece el centro U-TAD entre los primeros (figura 141). Vamos a intentar profundizar sobre qué opinan los usuarios.

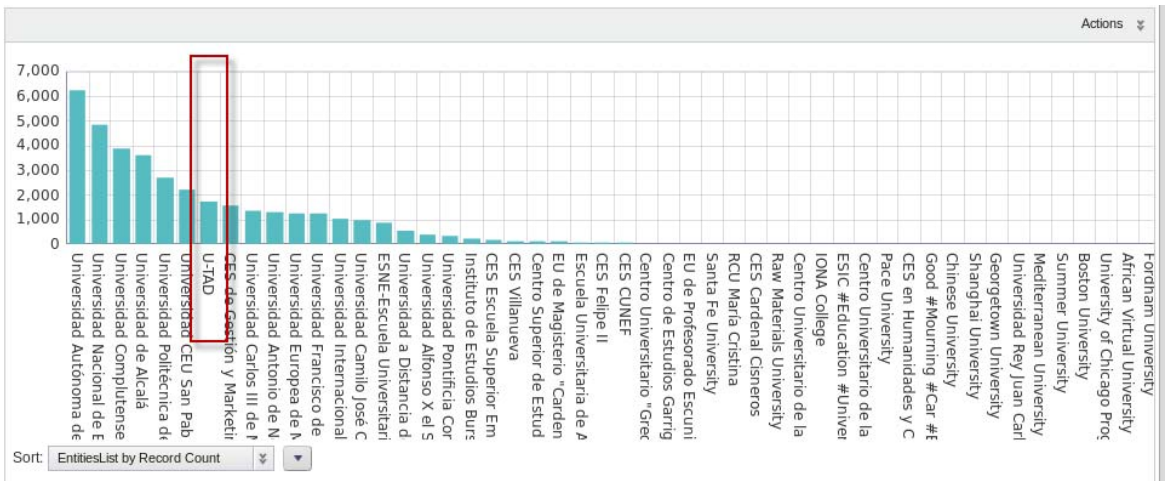


Figura 141. Centros asociados a las universidades. Fuente: elaboración propia.

Si observamos en la nube de palabras (figura 142) nos encontramos algunos datos curiosos. Parecen iniciativas que está haciendo el centro U-TAD y que son comentados muy positivamente. La mayoría de los tuits que hablan sobre el centro U-TAD son positivos (gráfico circular a la derecha en la figura 142).

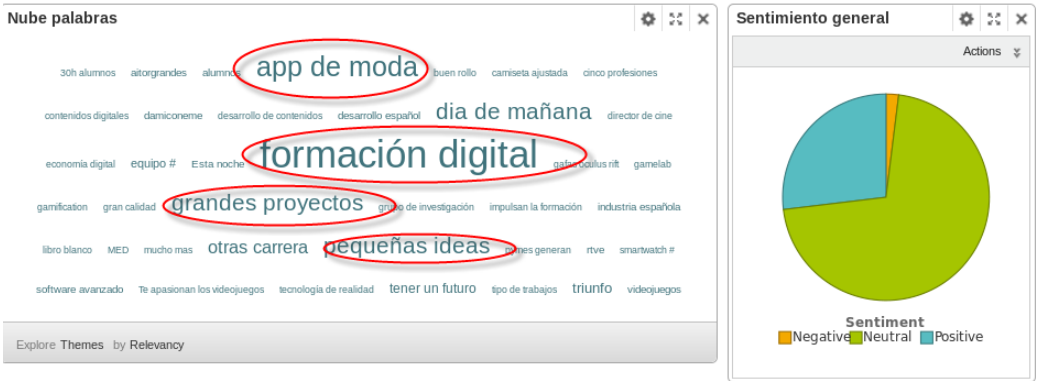


Figura 142. Filtro U-TAD aplicado. Fuente: elaboración propia.

Se resaltan algunos temas (marcados en rojo en la figura 142). Al entrar en los temas vemos:

App de moda (figura 143): alumno de U-TAD dice ser cofundador de una App.

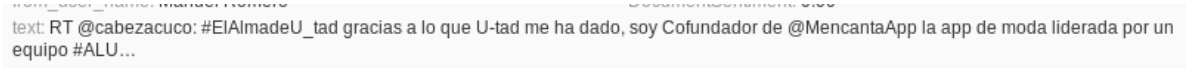


Figura 143. Tuit U-TAD. Fuente: elaboración propia.

Formación digital (figura 144): U-TAD participa en la exposición OMExpo, sobre formación digital.

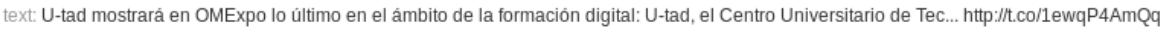


Figura 144. Tuit U-TAD. Fuente: elaboración propia.

¿Cuáles son los usuarios más influyentes que hablan de las universidades?

Simplemente nos desplazamos a la pestaña usuarios y en la gráfica "Relación tuits enviados y número de seguidores" (figura 145) se pueden ver los resultados. En el eje x muestra el número de seguidores, en el eje y el número mensajes enviados y el tamaño de la burbuja indica el número de amigos, es decir, aquellas personas que el usuario está siguiendo.

Los que aparezcan más a la derecha de la gráfica serán aquellos más influyentes.

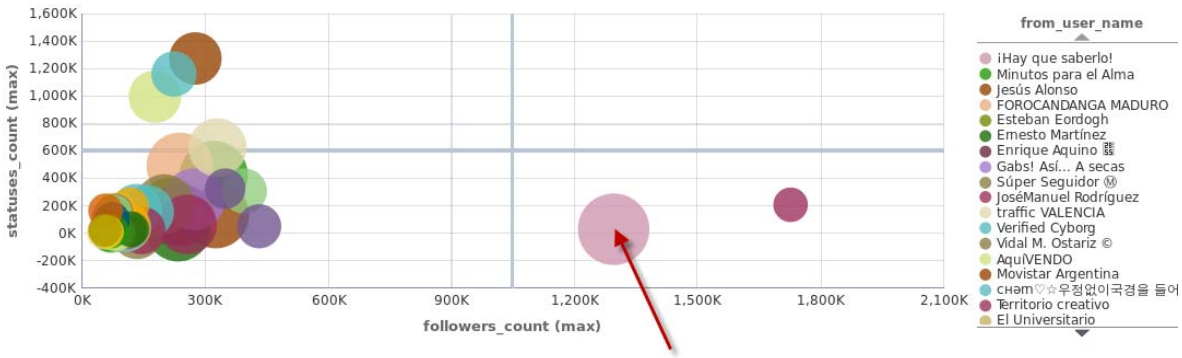


Figura 145. Relación seguidores y número de tuits enviados. Fuente: elaboración propia.

Por ejemplo, el usuario "¡Hay que saberlo!" tiene más de 1,3 millones de seguidores, está siguiendo a más de 250 mil usuarios y ha publicado algo más de 32 mil tuits. Habría que estudiar a qué personas está siguiendo para saber la importancia de sus seguidores.

6.6.2 Resultados sobre las universidades

En este apartado se muestran los resultados sobre las universidades de la Comunidad de Madrid. Estos resultados han sido tomados a partir de más de 266 mil tuis que se han ido adquiriendo a lo largo de 3 meses. Primero se analizarán los resultados globales sobre las universidades para más tarde particularizar el estudio sobre cada universidad.

En la figura 146 aparecen las 16 universidades ordenadas de mayor a menor por la cantidad de tuits que se han adquirido en este proceso.

Se puede observar que la universidad con mayor número de tuits es la Universidad Autónoma de Madrid con casi 35 mil tuits.

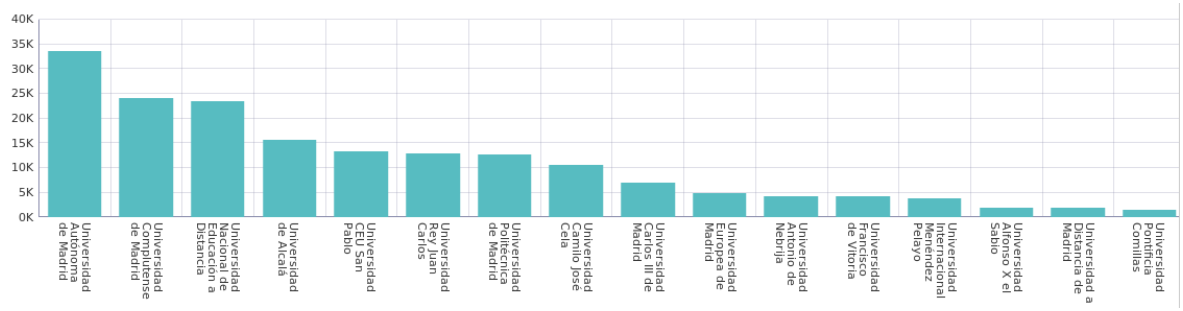


Figura 146. Número de tuits por universidad. Fuente: elaboración propia.

Con los tuits analizados y evaluados se ha hecho una media de los mismos por universidad. La siguiente gráfica, figura 147, muestra la puntuación media obtenida a partir del análisis de sentimiento por universidad. Hay que tener en cuenta que como la mayoría de los tuits son neutrales (calificados con un 0) los números son muy próximos a 0.

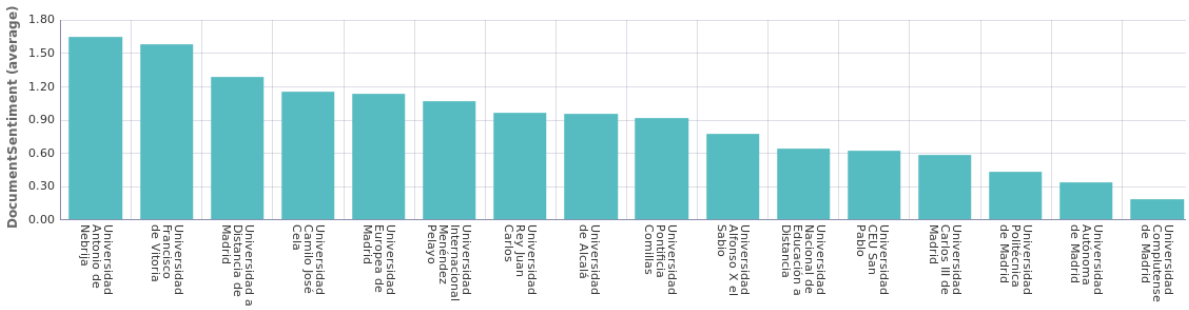


Figura 147. Media del análisis de sentimiento por universidad. Fuente: elaboración propia.

Se puede observar que con una puntuación mayor a 1,5 se encuentra la Universidad Antonio de Nebrija seguida de la Universidad Francisco de Vitoria.

Si colocamos las dos métricas anteriores en una gráfica de burbujas (figura 148) se puede ver que las universidades con más de 20 mil tuits han obtenido calificaciones por debajo de 0,7. Además se puede ver que no hay una relación clara entre las universidades con menos de 10 mil tuits, ya que la Universidad Carlos III de Madrid y la Universidad Rey Juan Carlos tienen un número parecido de tuits y existe una diferencia de más de 0,5 puntos en la media del análisis de sentimiento.

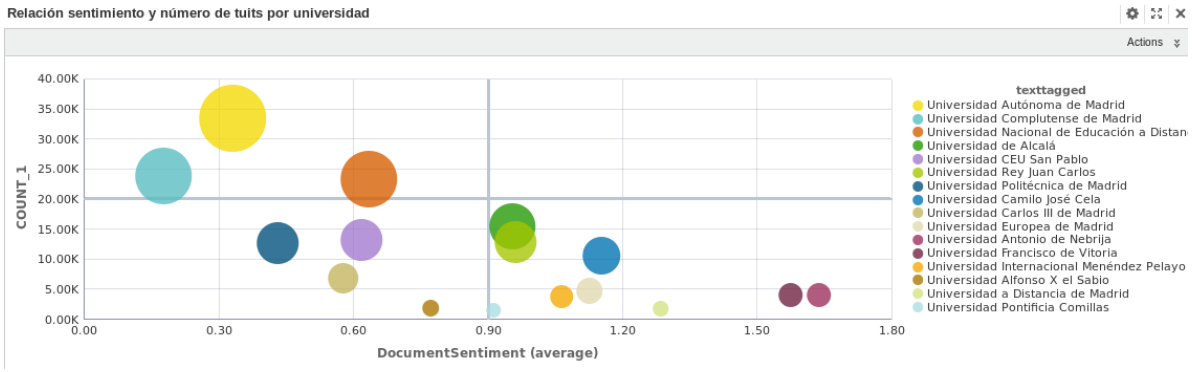


Figura 148. Relación número de tuits con sentimiento generalizado. Fuente: elaboración propia.

Si posicionamos en un mapa de calor los tuits (figura 149) obtenemos la repercusión internacional de las universidades analizadas. En las siguientes secciones se estudiará universidad por universidad su repercusión. Es importante tener en cuenta que según los datos obtenidos, solo el 2% de los tuits se envían con coordenadas de posicionamiento.

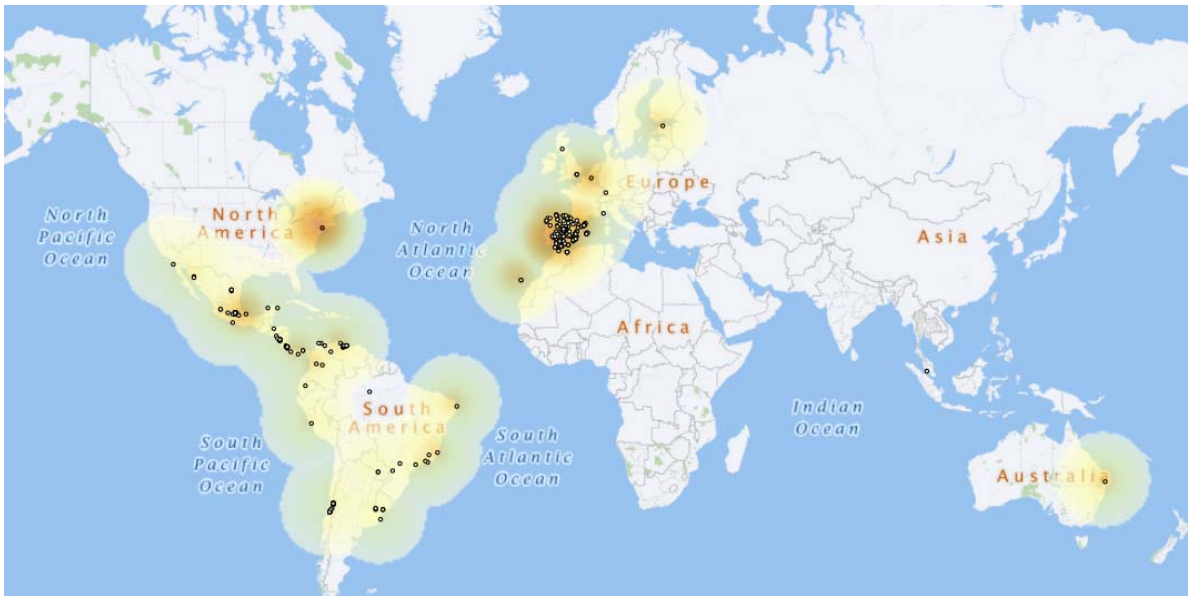


Figura 149. Mapa con los tuits obtenidos. Fuente: elaboración propia.

En la figura 150 se pueden apreciar los mismos datos a nivel nacional. La repercusión de las universidades de la Comunidad de Madrid está concentrada en algunas ciudades como Madrid, Barcelona, Sevilla, Alicante y Almería. También se observa un importante número de tuits en Valencia, Vitoria y Málaga.

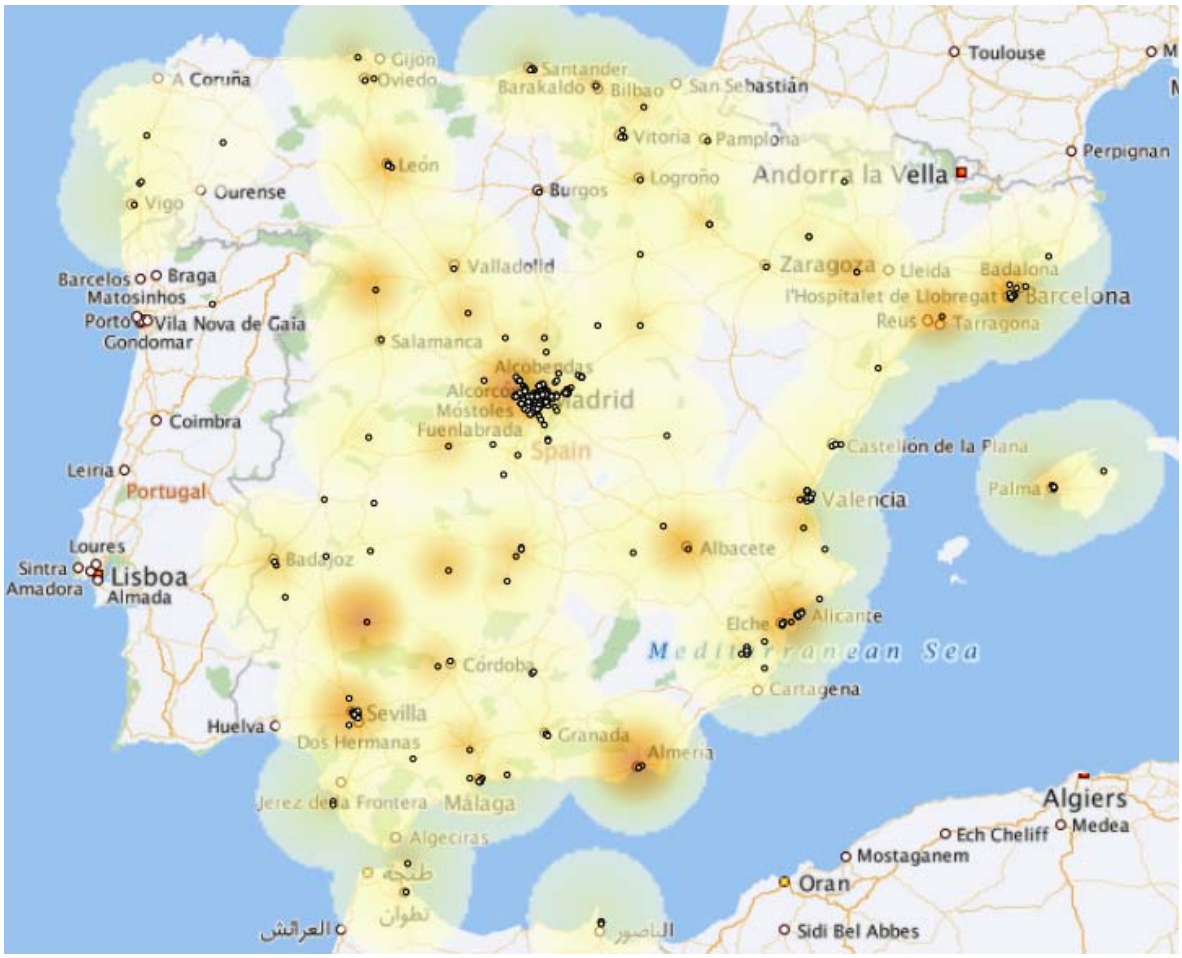


Figura 150. Tuits obtenidos a nivel nacional. Fuente: elaboración propia.

Por último en la figura 151 aparecen los tuits ubicados en Madrid centro. Se puede ver que existe una mayor concentración en Madrid centro y en las zonas en las que se sitúan las universidades como Alcalá, Getafe, Villanueva de la Cañada, Pozuelo y Ciudad Universitaria.

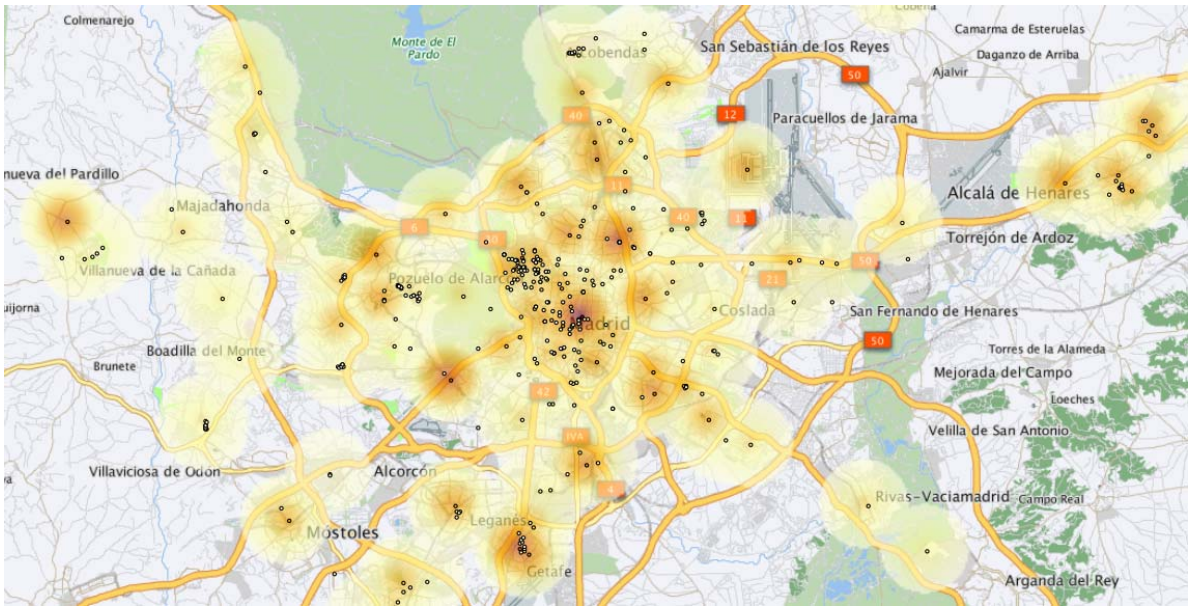


Figura 151. Tuits obtenidos a nivel de Madrid capital. Fuente: elaboración propia.

Los siguientes apartados se procederá a analizar las universidades en función del número de tuits obtenidos.

6.6.2.1 Universidad Autónoma de Madrid

Se han obtenido 33.453 tuits de los cuales el 18,6% eran positivos y el 13,1% negativos. Mediante la tabla 39 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 39. Usuarios más influyentes de la UAM. Fuente: elaboración propia.

	from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
	Univ. Autónoma Mad	324	0.92	28,440	763
	UAM Radio 94.1 FM	232	0.91	12,107	209
	Univ ArturoMichelena	219	0.49	20,045	275
	AlumniUAM	165	1.96	389	247
	UAM Comunidad	126	0.38	10,720	480
	Cosmovalencia	118	0.75	16,630	9,312
	Nancy Avianco	111	0.63	45	9
	Est Sin Censura UAM	93	0.27	2,394	1,023
	UAM Oficina Acogida	72	0.89	332	239
	Rafa valdes daussa	70	0.17	1,818	2,001
	Voz Universitaria	63	0.41	6,527	947
	DesdeLaUAM	61	0.91	974	106
	ONE TV	59	0.59	5,154	730
	Uamistas Mov Est	59	0.52	5,237	1,973
	UAM Xochimilco	56	0.79	3,787	194
	La Salle Campus MAD	55	1.57	1,290	1,947
	Angel Fernández N.	55	0.81	456	533
	FundaciónUAM	54	0.64	922	1,278
	Ligoteos® UAM ❤	54	1.73	5,394	5,712
	Fernando Luis Arraez	53	0.68	2,331	1,916

En la figura 152 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 152. Nube de palabras de la UAM. Fuente: elaboración propia.

La tabla 40 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 40. Tuits más influyentes de la UAM. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @AroomJBF: CUMPLEN 82 DÍAS PRESOS MAÑANA AUDIENCIA, PENDIENTES ESTUDIA...	259	0.00	279
RT @jordievole: Hablar con Mbuyi Kabunda profesor de la @UAM_Madrid es otro lujazo que t...	167	3.00	179
RT @ValenciaLaCalle: #8M Convocatoria #UAM "Vamos al campamento libertad en" #SanDie...	142	6.00	162
RT @Uunidas: la #UAM se pronuncia! http://t.co/an95FbH8k8	134	0.00	178
RT @ParidoNicaragua: Y asi Mayo dejó temblores, perros flechados, varios TT, jinchas jedion...	123	-3.00	132
RT @elalesicumbia: #HoyDaPara SALIR KOM UNA ESKOPETA I MATAR A TODAS LAS JORD...	114	0.00	1,033
RT @trafficVALENCIA: via @AroomJBF: CUMPLEN 82 DÍAS PRESOS MAÑANA AUDIENCIA, ...	103	0.00	114
RT @traffiCARACAS: via @AroomJBF: CUMPLEN 82 DÍAS PRESOS MAÑANA AUDIENCIA, P...	89	0.00	99
RT @HayQueSaberlo: Universidad Autónoma de Madrid: Las mujeres atractivas son más ego...	82	-9.00	146
RT @VOZ_UAM: Actividades que se realizaran mañana #7a en la #UAM. Unete y ayudanos a ...	72	6.00	81
RT @SoplosPAU2014: Los que os examináis en la Universidad Autónoma de Madrid ya podéi...	70	0.00	71
RT @ResistenciaG: Son 8 estudiantes guacareños de la UPEL, UAM, UC, UJAP. Están en el m...	66	0.00	70
RT @EstudiantesUC: Mañana #13M Marcha universitaria en #Carabobo 9:00am en el Carabo...	65	0.00	74
RT @Pajaropolitico: Desde hace tres años la UAM desarrolla un proyecto para convertir la ba...	64	-3.00	26
RT @DALEingenieria: Ya estamos las Universidades Unidas por Venezuela, UC - UJAP - UAM ...	61	0.00	20
RT @laSextaTV: RT @salvadostv: Mbuyi Kabunda es profesor de la @UAM_Madrid e investig...	58	0.00	59
RT @ManceraMiguelMX: Fomentemos en #NuestrosNiñosCDMX el hábito de reciclar. Particip...	57	0.00	60
RT @PabloGuzmanDice: RT @TrapieLLo: "@uam_ve: Autoridades de la UAM y @TeleAragua ...	51	3.00	53
RT @CNNMex: El @IPN_MX es el lugar 26 del @worlduniranking #Latinoamérica; la @IBERO_...	51	0.00	55
RT @OAE_UAM: En 1988 D.Felipe, #FelipeVI tras el anuncio de abdicación de #JuanCarlosI, in...	49	0.00	51

La figura 153 y 154 muestra la posición de los tuits. La principal influencia a nivel nacional es Madrid aunque también se localizan algunos tuits de personas hablando de la UAM en León y Santander.

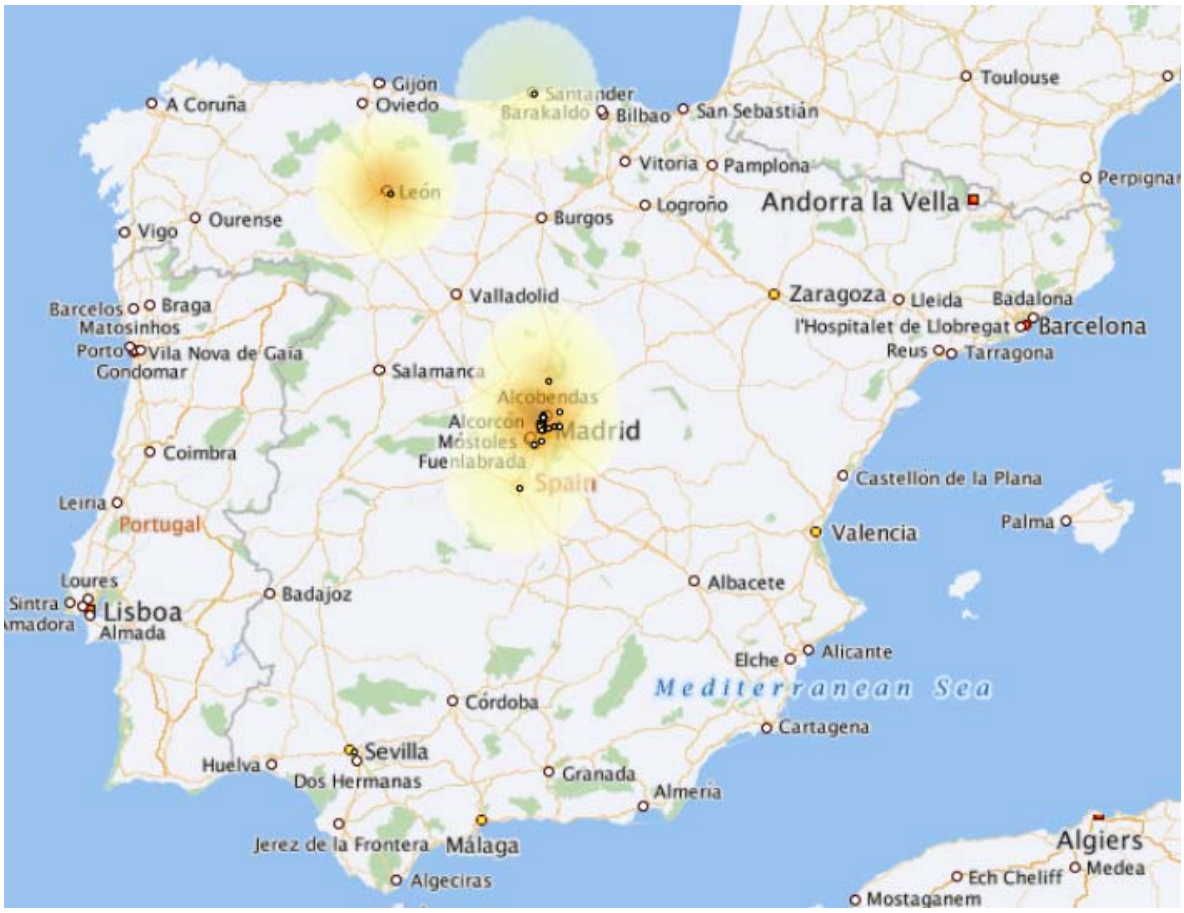


Figura 153. Tuits UAM en España. Fuente: elaboración propia.

En Madrid (figura 154) el foco principal es la zona de Alcobendas, donde se sitúa la universidad. También se observan algunos otros puntos en la capital de personas publicando tuits.

Respecto a sus cuatro centros asociados, el único que tiene impacto en Twitter es el Centro Superior de Estudios Universitarios La Salle.

Por último, el campus de Medicina situado cerca del Hospital La Paz no tiene un número de tuits relevante.

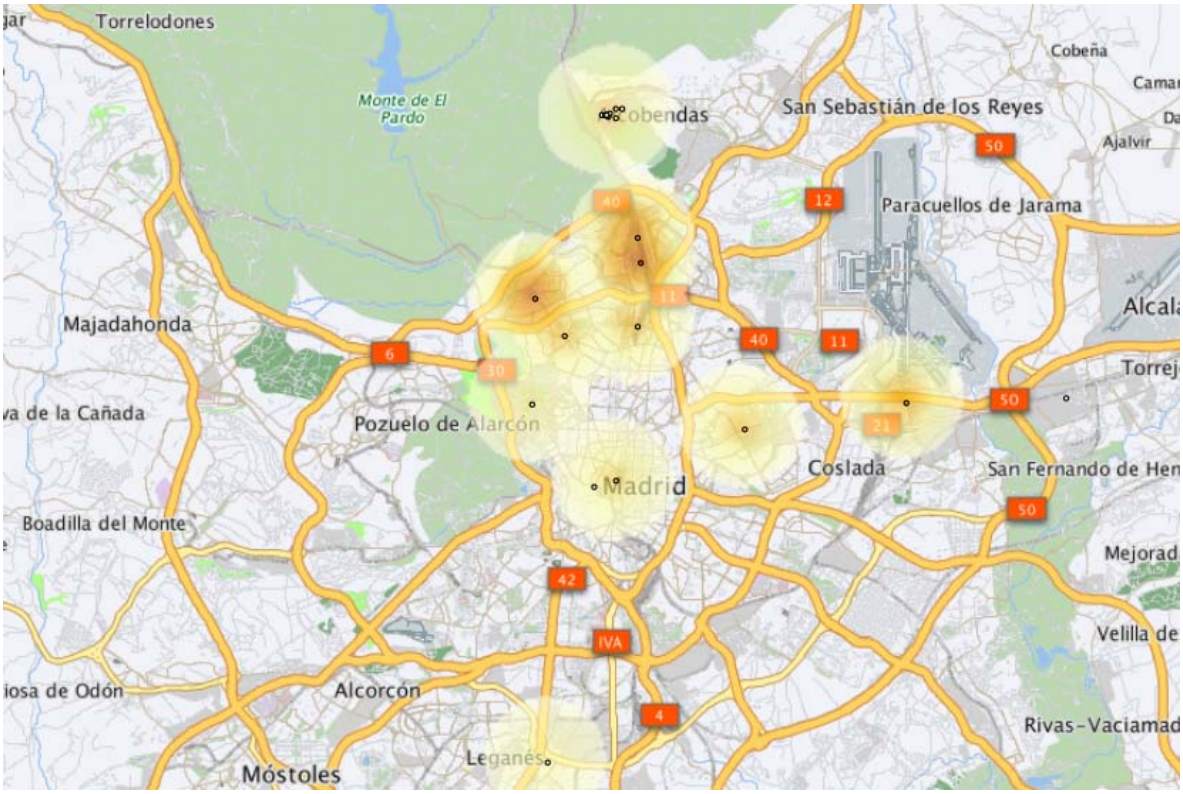


Figura 154. Tuits UAM en Madrid. Fuente: elaboración propia.

6.6.2.2 Universidad Complutense de Madrid

Se han obtenido 23.966 tuits de los cuales el 18,7% eran positivos y el 17,1% negativos. Mediante la tabla 41 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 41 .Usuarios más influyentes de la UCM. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Rebeca Schack	208	6.00	243	203
IEB Alumni	148	1.12	1,033	1,287
IEB	145	0.83	2,719	2,010
La Casa Estudiante	139	1.15	1,745	1,235
ioannis koutsourais	112	-4.00	253	227
WENYARD	99	0.00	25,326	4,074
Extensión UCM	91	0.52	2,205	1,494
ColectivoEstudiantes	78	0.28	2,026	1,010
Alfonso Sánchez	68	0.00	6,258	5,531
Silvia Freire	60	7.27	3,440	17
David B. R.	59	6.00	258	764
Consultor WISHCLUB	56	0.00	2,871	1,618
Osiris S. de Briceño	52	0.55	263	204
Villanueva C.U.	51	0.74	1,287	961
AntonioVChanal	49	0.00	52,662	55,327
Espinosa Unicornio	49	0.37	122	241
UGT-UCM	49	-0.34	381	798
diegoperez	47	-0.42	1,502	1,486
Noticias de Madrid	47	-1.82	2,960	3,007
UCM_Económicas	46	0.88	2,386	1,434

En la figura 152 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 155. Nube de palabras de la UCM. Fuente: elaboración propia.

La tabla 42 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 42. Tuits más influyentes de la UCM. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @Pablo_Iglesias_: Hoy nos encontraremos todos en la UCM para preparar la Asamblea Ci...	131	0.00	144
RT @unicomplutense: El rector de la UCM anuncia que el director del departamento de Anato...	108	-6.00	117
RT @miguelito092: Estoy estudiando en la biblioteca de Geografía-Historia de la UCM y por su...	99	0.00	105
RT @theiwon: Asamblea General de Podemos en la Facultad de Filosofía de la UCM, con Íñigo ...	96	0.00	105
RT @ahorapodemos: Cada vez más gente llegando a los alrededores de la facultad de filosofí...	94	0.00	101
Alfonso Sánchez - La venta directa se estudiará en la UCM http://t.co/a4CY1MVBpe	86	0.00	1
Alfonso Sánchez - La venta directa se estudiará en la UCM http://t.co/PSnRaZOGmn	73	0.00	1
RT @Sr_Dios: Me parece bien que los antidisturbios vayan a la UCM, nunca es tarde para pon...	72	5.00	1,088
RT @PodemosAlicante: Balance y analisis electoral, el paraninfo de la UCM lleno. Hablan @ie...	67	0.00	70
RT @ManuBreakout: Sobre el inesperado cambio del CES Felipe II http://t.co/liDc7tgCGr	57	0.00	60
Alfonso Sánchez - La venta directa se estudiará en la UCM http://t.co/BduJ1E7hNE	56	0.00	1
RT @promero1986: ¿Habéis visto las imágenes del "sotano de los horrores" de la Universida...	54	-7.50	56
RT @J_LoSantos: "Soy profesor en la #UCM porque gané unas oposiciones. ¿Has ganado tú ...	53	6.00	55
RT @CriminologiaUCM: El próximo 28 de Abril venid a la jornada de los alumnos de criminolo...	52	-4.00	53
RT @la_tuerka: "El Rey es el mayor corrupto del reino" Juan Varela (profesor de Filología UC...	51	-6.00	53
RT @ahorapodemos: ¿Te vienes mañana con nosotros al encuentro #Podemos14J? Será a p...	50	0.00	52
RT @Adrimetallica: Si llego a 100 RT salgo corriendo desnudo por el campos de la UCM (????)	49	0.00	52
RT @PabloLolaso: Han cerrado la mítica cancha de UCM sin motivo y hay gente, con razón, in...	47	3.00	52
RT @OscarRG4: Éxito de convocatoria de la asamblea general de @ahorapodemos. Llena la ...	46	0.00	48
RT @Vidal_M_Ostariz: Sótano de los horrores de #JoseCarrillo d casta le viene al galgo. #Dim...	45	-7.50	49

La figura 156 muestra la posición de los tuits a nivel nacional. El principal foco es en Madrid aunque también se observa algún tuit en Valencia, Badajoz, Valladolid y Valencia entre otros de personas hablando sobre la UCM.

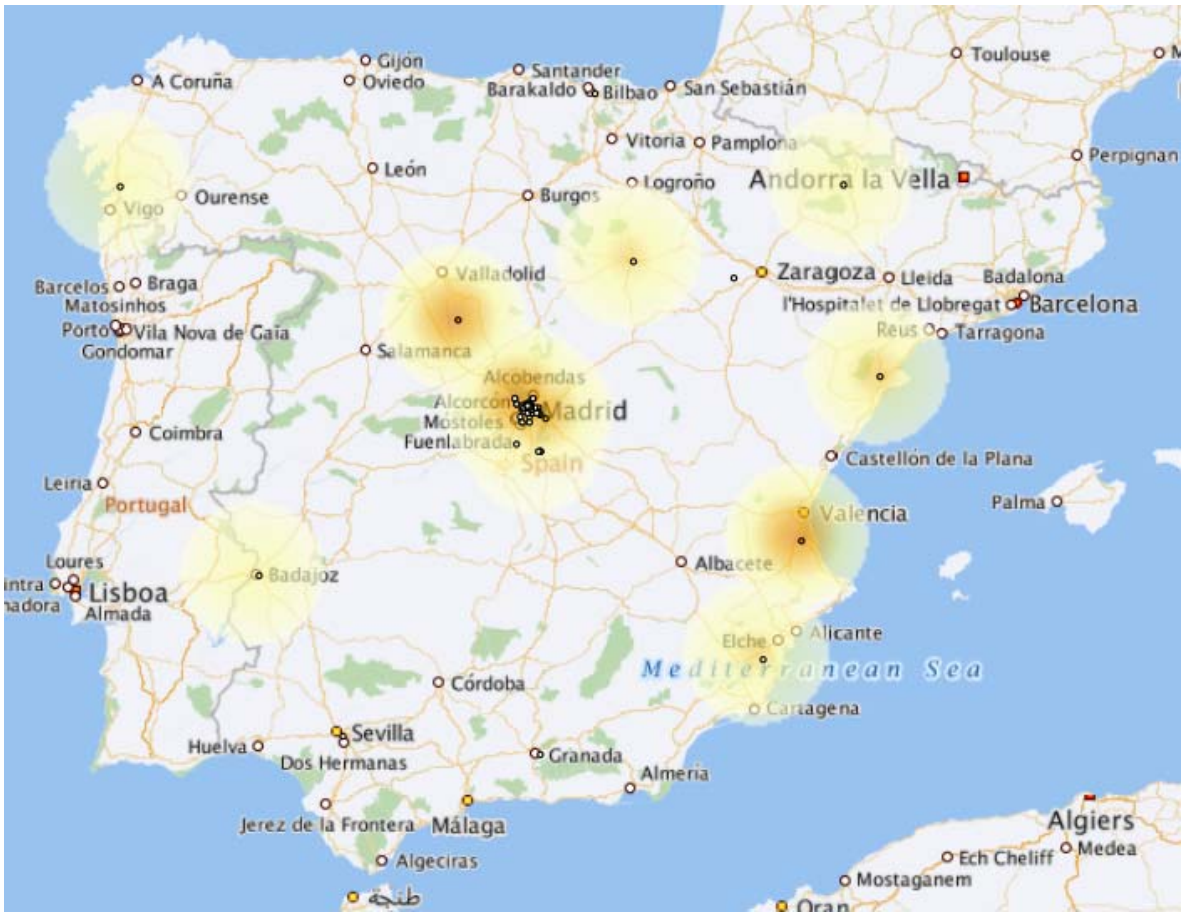


Figura 156. Tuits UCM en España. Fuente: elaboración propia.

En Madrid (figura 157) hay principalmente dos focos, el primero en el área de Ciudad Universitaria donde se encuentra la Universidad Complutense de Madrid y el CES en Humanidades y Ciencias de la Educación Don Bosco. El segundo en el campus de Somosaguas de la UCM.

Respecto a sus centros adscritos no se observa una concentración en torno a ninguno de ellos aunque si aparecen tuits hablando sobre el IEB (zona Retiro), Escuela Universitaria de Magisterio de Madrid (Carabanchel), CES Cardenal Cisneros (Príncipe de Vergara), CES Villanueva (carretera de Colmenar y zona Retiro) y CUNEF (Alonso Martínez).

Por otra parte, existen numerosos tuits publicados desde diferentes zonas de la capital.

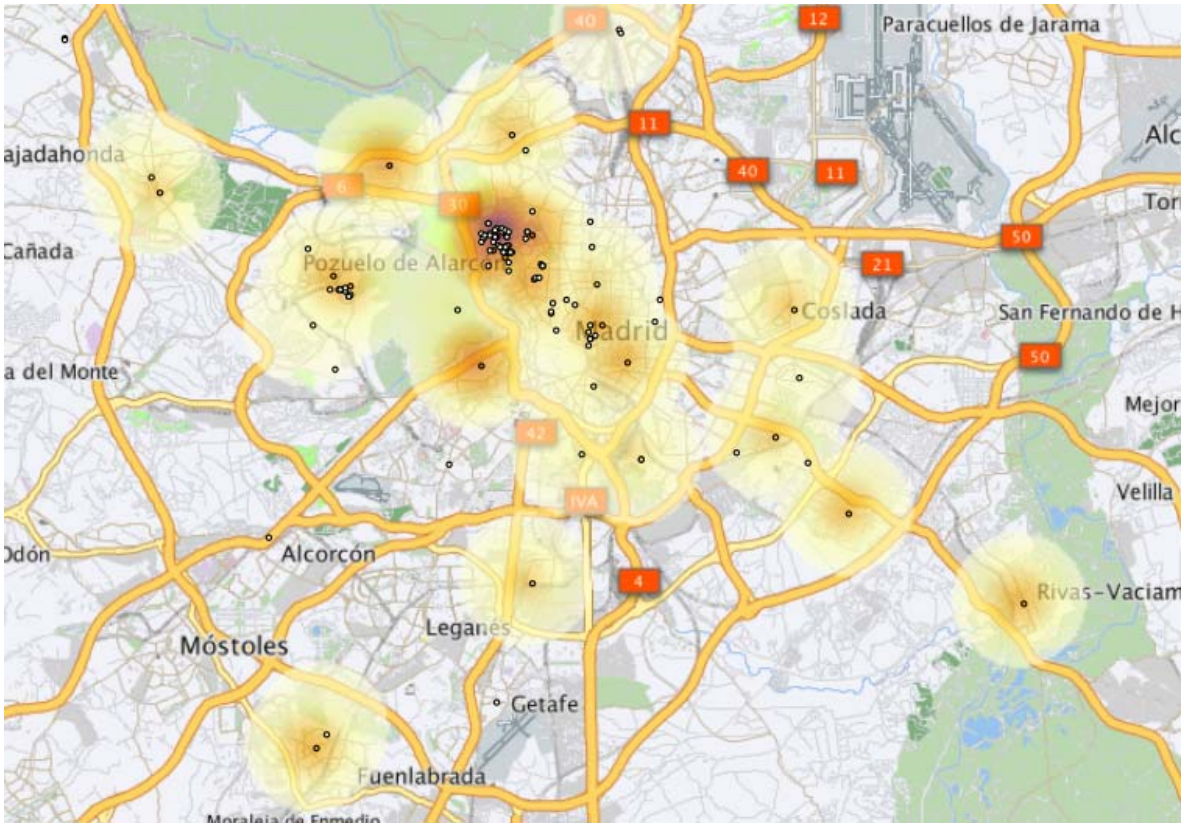


Figura 157. Tuits UCM en Madrid. Fuente: elaboración propia.

6.6.2.3 Universidad Nacional de Educación a Distancia

Se han obtenido 23.390 tuits de los cuales el 20,5% eran positivos y el 6,9% negativos. Mediante la tabla 43 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 43. Usuarios más influyentes de la UNED. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Tiberio Feliz Murias	1,095	0.30	1,936	2,000
ExpertoAnimaLectura	823	0.31	1,126	1,416
Tiberio Feliz	784	0.51	3,388	3,444
Congreso Secundaria	711	0.19	1,344	1,398
Inteligencia Lúdica	688	0.09	834	1,160
Tecnico Infantil	509	-0.06	1,193	1,193
UNED	440	1.29	69,534	838
Promoción de cursos	367	0.02	1,027	1,285
Educación XX1	234	0.48	1,020	1,047
Top 10 iTunes U	186	0.02	94	17
UNED Málaga	167	0.87	1,011	648
Conchita Travesedo	148	1.57	1,433	1,029
Podcast Radio 3	126	0.17	1,311	72
Educa León	121	0.07	3,265	3,157
Biblio UNED Málaga	86	0.76	627	8
Ocio Melilla Now	78	0.15	2,860	898
Marlene Víquez S.	69	0.65	37	79
divulgaUNED	63	-0.11	7,531	686
UNED Barbastro	59	0.19	797	0
INTECCA - UNED	54	0.09	2,899	218

En la figura 158 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 158: Nube de palabras UNED. Fuente: elaboración propia.

La tabla 44 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 44. Tuits más influyentes de la UNED. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @tableroglobal: PP de Melilla quiere cambiar nombre de Federico García Lorca por el de u...	88	0.00	98
RT @UNED: Buenos días. Hoy comienzan los #ExámenesUNED... A nuestra Comunidad Univ...	84	0.00	94
RT @RaquelEcenarro: ¡Ojo! Que llegue a periodistas sin #empleo UNED busca periodista res...	69	0.00	71
RT @_infoLibre: El Gobierno de Melilla prefiere a un político imputado para dar nombre al cent...	68	0.00	45
RT @ramonlobo: El Gobierno de Melilla prefiere a un político imputado para dar nombre al cen...	48	0.00	51
RT @reserva_71: Iba a la biblioteca tengo examenes uned de mi 2a carrera, para que? Para e...	41	0.00	53
RT @UNED: Ven a ver la exposición '100 años con Julio Cortázar' en la @Biblioteca_UNED H...	41	0.00	13
RT @UNED: Asiste a un #CursosdeVerano de la Universidad desde dónde tú quieras. Cursos ...	40	0.00	17
RT @UNED: Buenos días. Fin de semana anterior a los #ExámenesUNED. Deseamos trabajo ...	30	3.00	36
RT @UNED: Tterminando la 1ª semana de #ExámenesUNED no es momento de relajarse. Co...	30	0.00	35
RT @UNED: 48 bibliotecas de museos liberarán en bookcrossing 2.500 vols. Busca el tuyo #D...	27	3.00	29
RT @UNED: Recompensa... #ExámenesUNED #UNED RT @Biblioteca_UNED: "Lo que con m...	27	3.00	28
RT @AngelesMunoz_: La @UNED se implantó en #Marbella hace 30 años y en los últimos año...	26	5.00	26
RT @barkanimelilla: Estoy convencido de que la UNED seguirá avanzando en #Melilla con su ...	25	0.00	27
RT @UNED: Nuestro #FF para los integrantes de la Comunidad Universitaria que participan e...	25	0.00	30
RT @UNED: +1 RT @Karmen_GP En @UNED #Motril una alumna de 69 años me dió la mejor l...	25	-3.00	27
RT @jzamorabonilla: Curso de Verano UNED LOS PÚBLICOS DE LA CIENCIA Incluirá una visit...	24	0.00	26
RT @UNED: Aquí tenéis, Calendario de exámenes por asignaturas de pruebas presenciales d...	24	0.00	26
RT @UNED: ¡Un último esfuerzo al 100%! Nuestro #FF dirigido a toda la Comunidad Universit...	24	0.00	26
RT @RAEinforma: La @UNED organiza un curso de verano en Ávila sobre #RAE300años del ...	23	0.00	14

La figura 159 se muestra la posición de los tuits a nivel nacional. Los principales focos se encuentran en Madrid y Barcelona. También existen otros focos menos numerosos en Valencia, Sevilla, Almería, Zaragoza y Bilbao.

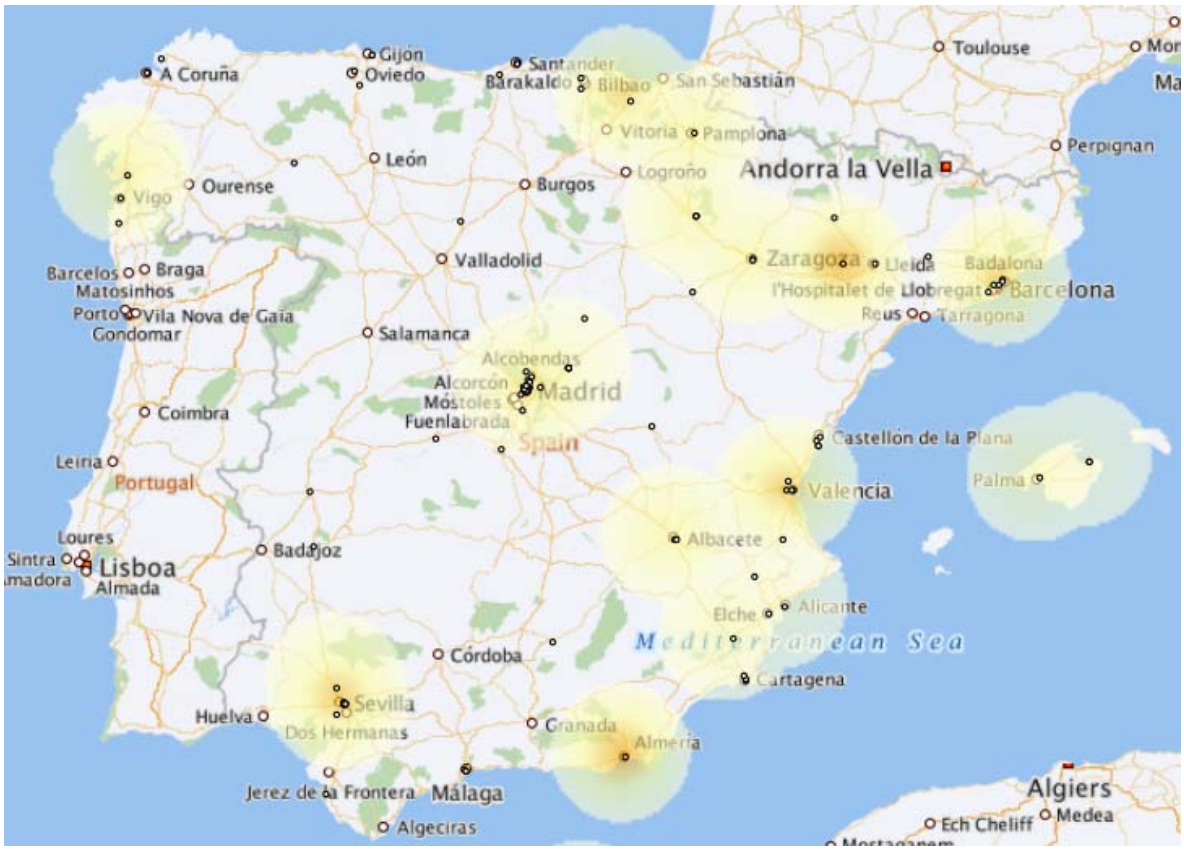


Figura 159. Tuits UNED en España. Fuente: elaboración propia.

En Madrid (figura 160) los tuits se concentran en el centro y en el campus de Ciudad Universitaria.

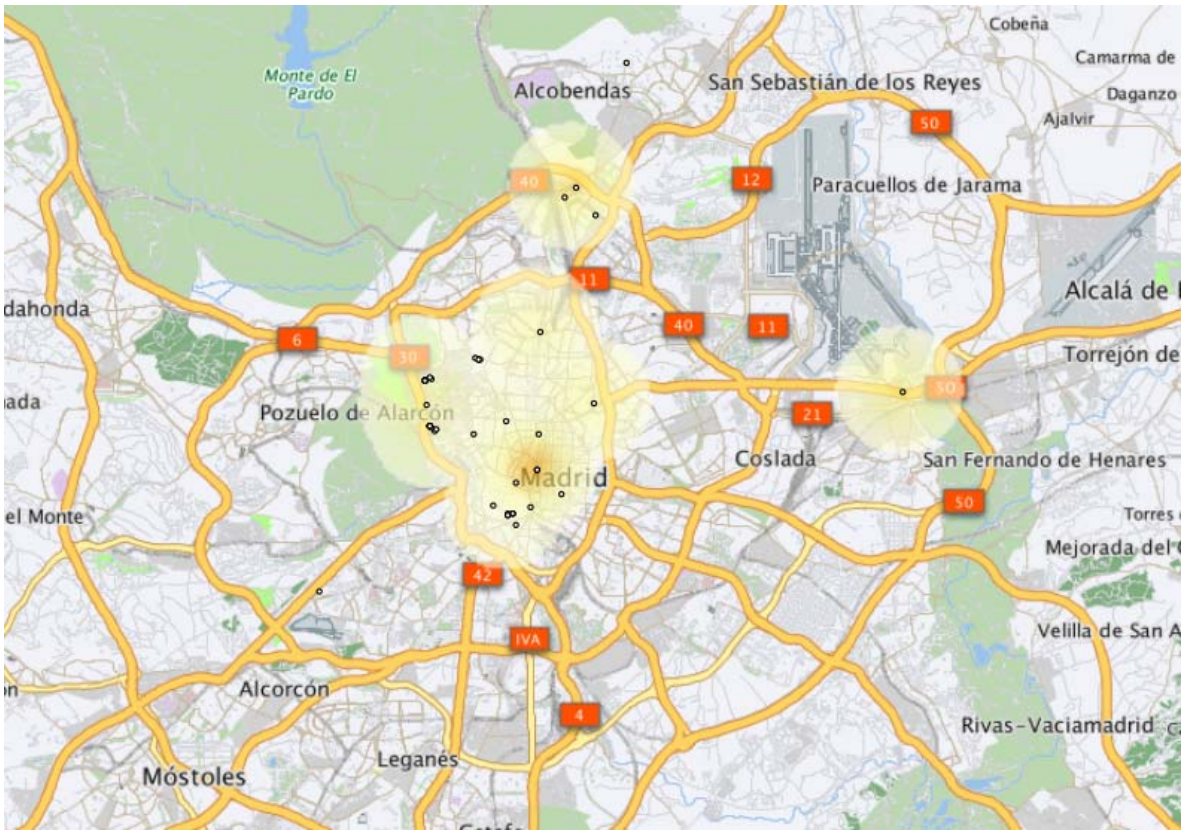


Figura 160. Tuits UNED en Madrid Fuente: elaboración propia.

6.6.2.4 Universidad de Alcalá

Se han obtenido 15.580 tuits de los cuales el 24,3% eran positivos y el 6,5% negativos. Mediante la tabla 45 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 45. Usuarios más influyentes de la UAH. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Universidad Alcalá	734	1.01	16,322	72
Nancy Avianco	341	1.18	44	9
Posgrado UAH	240	1.14	647	89
Rachel Ornay	189	1.11	36	16
Alcalá Turismo	138	0.61	2,183	2,166
Admisión UAH	120	1.56	1,671	1,951
Extensión UAH	85	0.48	733	88
CU Cardenal Cisneros	83	1.35	1,257	1,945
Postgrados UAH	81	1.64	180	90
Estudiantes UAH-Val	74	-0.30	951	1,097
U. Alberto Hurtado	71	1.75	2,172	2,110
Liceus Humanidades	60	0.62	2,474	2,075
CEUAH	53	0.67	1,463	103
David Orden	52	2.18	777	228
InformerUAH	51	0.37	5,714	28
Raquel F	49	0.84	133	163
VoluntariosUAH	46	1.33	3,688	723
Biblio Poli UAH	46	1.02	811	231
AlcalaHoy	45	0.68	469	176
Portal Local	42	1.10	2,418	1,584

En la figura 161 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 161. Nube de palabras UAH. Fuente: elaboración propia.

La tabla 46 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 46. Tuits más influyentes de la UAH. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @alfredoromero: 12pm, 2junio: Con @aliyergallegos estudiante UAH liberado el viernes d...	326	4.50	369
RT @juanflores18: Mañana #12M saldremos a exigir ¡LIBERTAD! #USB #UCV #UCAB #UNIM...	105	12.00	111
RT @vzlapray: Aquí les dejamos las cuentas del Movimiento Estudiantil de #CCS #UCV #UniM...	74	0.00	13
RT @AleRosillon: Hoy los #Humboldtianos nuevamente dieron muestra d su coraje comprom...	70	0.00	75
RT @EstudiantesUC: Mañana #13M Marcha universitaria en #Carabobo 9:00am en el Carabo...	65	0.00	74
RT @gredossandiego: Felicidades a todos nuestros alumnos de @gredossandiego que se gra...	64	0.00	67
RT @DALEingenieria: Ya estamos las Universidades Unidas por Venezuela, UC - UJAP - UAM ...	61	0.00	20
RT @jumastorga: Mañana en #elinformante debatimos sobre la reforma a la educación superi...	57	6.00	59
RT @andreainsunza: Lugar de residencia de estudiantes PUC, UChile, UAI, UDP y UAH en 20...	57	0.00	60
RT @andreainsunza: Lugar de residencia de estudiantes PUC, UChile, UAI, UDP y UAH en 20...	39	0.00	300
RT @victoriajuarezc: Porque la Educación es el futuro de la nación, VOX estuvo en la Universi...	39	0.00	41
RT @AnonymousCcsVE: #UCV #ULA #UDO #LUZ #UC #UCLA #UCAB #UNIMET #USM #URB...	37	0.00	46
RT @EdImpedimenta: Elena Poniatowska en el Paraninfo de la Universidad de Alcalá (@UAHe...	32	0.00	33
RT @AleRosillon: Aliyer Pacheco estudiante de Contaduría UAH sera enviado a Tocoron por 4...	28	0.00	36
RT @24HorasTVN: ESTA NOCHE en #ElInformante: Rectores de la Usach, UAI, PUC y UAH d...	26	0.00	15
RT @UAHes: Discurso de Elena Poniatowska #PremioCervantes en el Paraninfo de la #UAH ...	26	0.00	27
RT @COAMadrid: Por una sociedad con servicios de calidad. Miércoles |13:30| P.Mayor #Noal...	25	4.00	25
RT @24HorasTVN: ESTE MARTES, los rectores de la Usach, PUC, Adolfo Ibáñez y UAH debat...	25	0.00	16
RT @UAHes: Los estudios de Medicina de la #UAH han logrado situarse entre los mejores de ...	23	3.76	26
RT @verdimm: #UAH presente en la UCV #3A http://t.co/lUEWfN4Fsf	23	0.00	30

La figura 162 muestra los tuits a nivel nacional. Los principales focos están bien identificados: Madrid, Barcelona, Valencia, Vitoria y Badajoz.

El Centro Universitario de la Defensa (Murcia) no presenta ningún tuit.

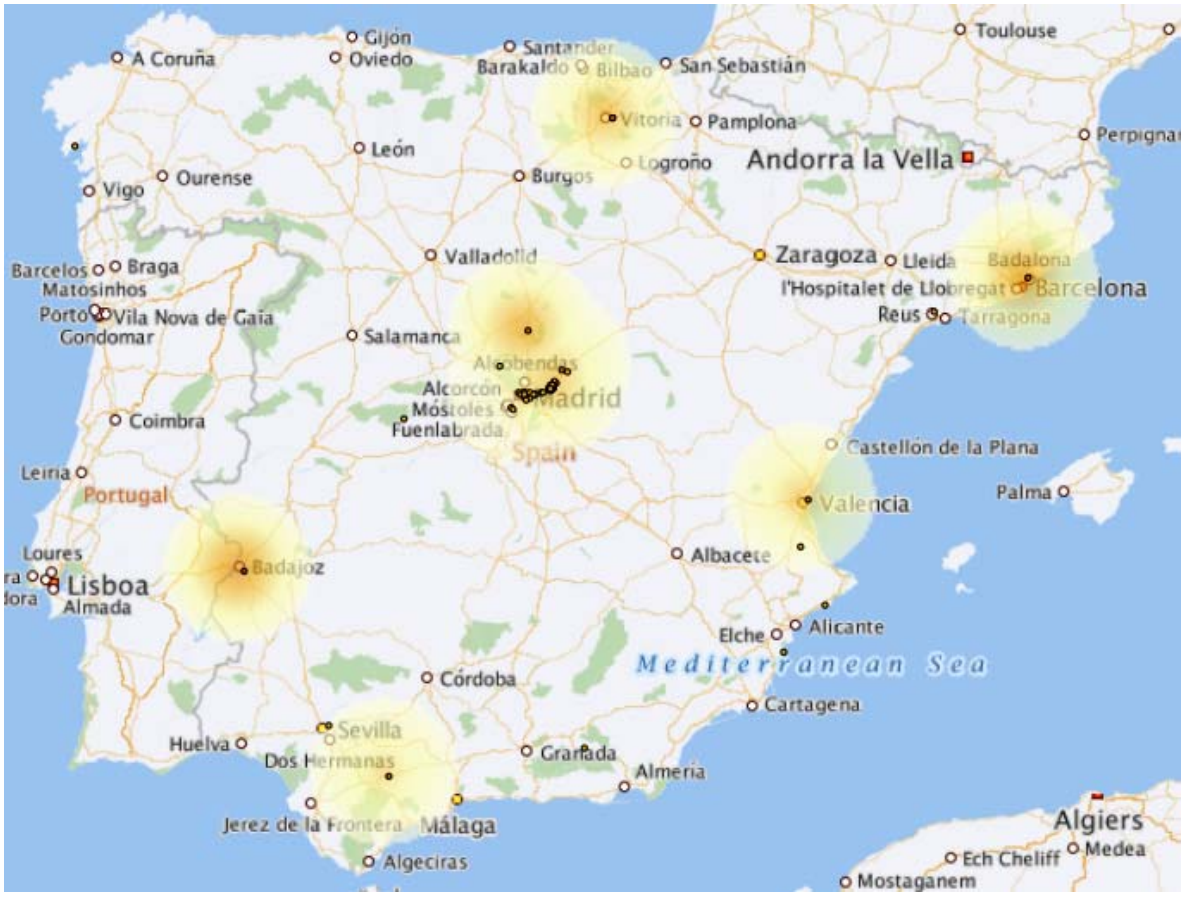


Figura 162. Tuits UAH en España. Fuente: elaboración propia.

En Madrid (figura 163) los principales tuits se encuentran en la ciudad de Alcalá de Henares y en el campus de la Universidad de Alcalá. Otros menores focos en la ciudad de Madrid y en Coslada.

Respecto a su centro asociado en Las Rozas, el Centro Universitario Gredos San Diego, no aparece ningún tuit.

Figura 163. Tuits UAH en Madrid. Fuente: elaboración propia.

6.6.2.5 Universidad CEU San Pablo

Se han obtenido 13.269 tuits de los cuales el 16% eran positivos y el 4,9% negativos. Mediante la tabla 49 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 47. Usuarios más influyentes del CEU. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
CEUMEDIA	121	0.88	787	797
Universidad CEU-UCH	96	1.31	5,515	351
OnCEU Digital	84	1.43	1,300	720
Onda UNED	62	0.06	1,389	418
Fundación CEU	57	0.91	3,699	531
Universidad CEU-USP	56	0.56	4,868	12
Unimel Educación	55	0.00	5,632	4,641
Pechitos McTetis	52	0.00	86	352
onceu	49	0.73	1,249	675
Faul y Cuenta	43	0.12	4,334	333
Inst. Est. Europeos	42	0.51	404	396
SafydeUEx	35	1.74	681	203
Deporte Univ Almeria	33	0.57	1,293	64
Posgrado CEU	31	0.90	382	173
CIC	30	0.00	58,522	389
Miguel Á de Santiago	29	0.91	403	502
Luis G. Carpio M	28	1.85	66	102
ABE AC	28	-0.08	342	13
Sebastian Fournier	27	0.15	521	898
Marlene Víquez S.	27	0.80	35	79

En la figura 164 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 164. Nube de palabras UAH. Fuente: elaboración propia.

La tabla 50 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 48. Tuits más influyentes del CEU. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @ESTDemoscopia: SONDEO pie de urna: PP 18-20 PSOE 15-17 IU 7-8 UPyD 3-4 ERC 2-3 P...	260	0.00	294
RT @atlante83: SONDEO Demoscopia PP 18-20 PSOE 15-17 IU 7-8 UPyD 3-4 ERC 2-3 Podem...	151	0.00	156
RT @elconfidencial: Os recordamos que CIS decía: -PP, 20-21 escaños -PSOE, 18-19 -IU, 5 -C...	94	0.00	759
RT @electionista: #Spain - average of #EP2014 polls: PP 30.5% PSOE 29.5% IU 12.4% UPyD 8...	67	0.00	67
RT @CasaReal: La Infanta Elena, con los alumnos de la Escuela At.Madrid. Universidad del D...	61	0.00	65
RT @pasapelectoral: Sonde pie de urna #España: PP 18-20 PSOE 15-17 IU 7-8 UPyD 3-4 ERC ...	59	0.00	62
RT @electionista: #Spain - Metroscopia #EP2014 poll: PP 32.6% PSOE 32.2% IU 12% CEU 4.7...	56	0.00	57
RT @Sociometro: @ESTDemoscopia SONDEO pie de urna: PP 18-20 PSOE 15-17 IU 7-8 UPyD...	54	0.00	55
RT @elconfidencial: Os recordamos resultados: -PP,16 escaños -PSOE,14 -IU,6 -Podemos,5 -...	51	0.00	54
RT @elconfidencial: Escaños en las europeas según el CIS: PP: 20-21 PSOE:18-19 IU-ICV: 5 C...	42	0.00	43
RT @metroscopia: ¿A qué partido/s no votaría en ningún caso? PP 48% PSOE 23% Izquierda ...	39	0.00	51
RT @plazaro67: Sondeo #España #ep2014 #europees2014 según @ESTDemoscopia:PP 18-2...	35	0.00	36
RT @sanchez_sonia: #CIS Europeas: PP 20-21 escaños / PSOE 18-19 / IU 5 / UPyD 3 / CEU 3 / ...	35	0.00	37
RT @agarzon: Resultados definitivos: PP: 16 PSOE: 14 IU: 6 PODEMOS: 5 UPyD: 4 CEU: 3 EP...	33	0.00	1,553
RT @FranHervias: @jlphispania PP 33,1% 20-21 PSOE 30'2% 18-19 IU 10'6% 6 UPyD 7'2% 4 c...	33	0.00	33
RT @mariaorbegozo: Panorámica del Aula Magna del CEU, donde se está desarrollando el #i...	33	0.00	34
RT @elconfidencial: Intención de voto en las europeas según el CIS: PP: 33,7% PSOE: 31% IU:...	31	0.00	37
RT @ierrejon: Acaba de salir la encuesta del #CIS de las Europeas: PP 20-21 escaños;PSOE ...	29	0.00	29
RT @tuitadyneLPD: Encuesta @Metroscopia para @elpais_politica PP 19 PSOE 19 IU 6 ERC 3...	29	0.00	30
RT @CentroPRONAF: Muy buena imagen de @USPCEU sobre la #obesidad y #sedentarismo ...	29	-1.63	33

La figura 165 se muestra la ubicación de los tuits a nivel nacional. Los principales focos son Madrid donde se encuentra el CEU San Pablo, Sevilla (CES Cardenal Spínola CEU), Barcelona (Abad Oliba CEU), Alicante (CEU Cardenal Herrera) y Valencia (CEU Cardenal Herrera), todos ellos pertenecientes a la fundación CEU San Pablo.

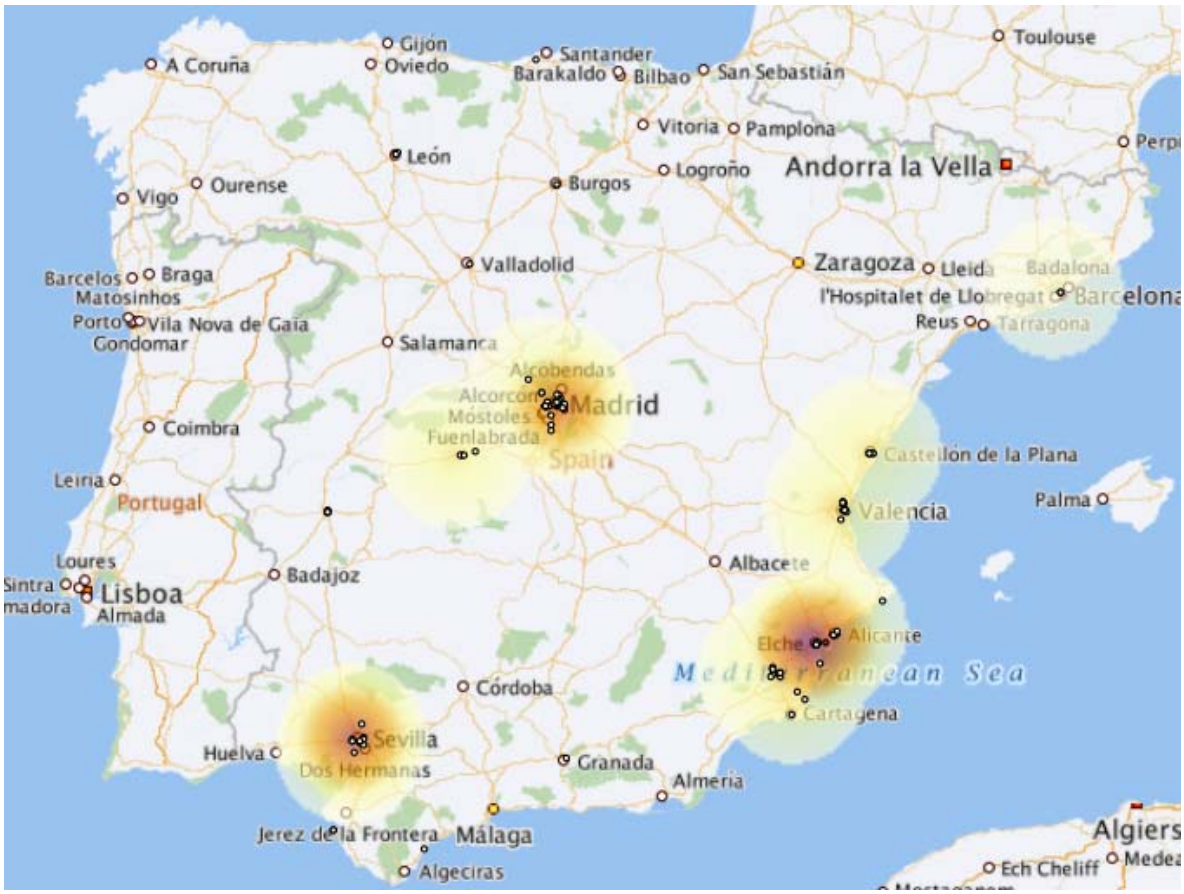


Figura 165. Tuits CEU en España. Fuente: elaboración propia.

En Madrid, figura 166, se localizan los tuits principalmente en el campus de Moncloa (zona Argüelles) y en el área del centro de Madrid. Por otra parte, en Alcorcón se muestran algunos tuits pertenecientes al campus de Montepríncipe.

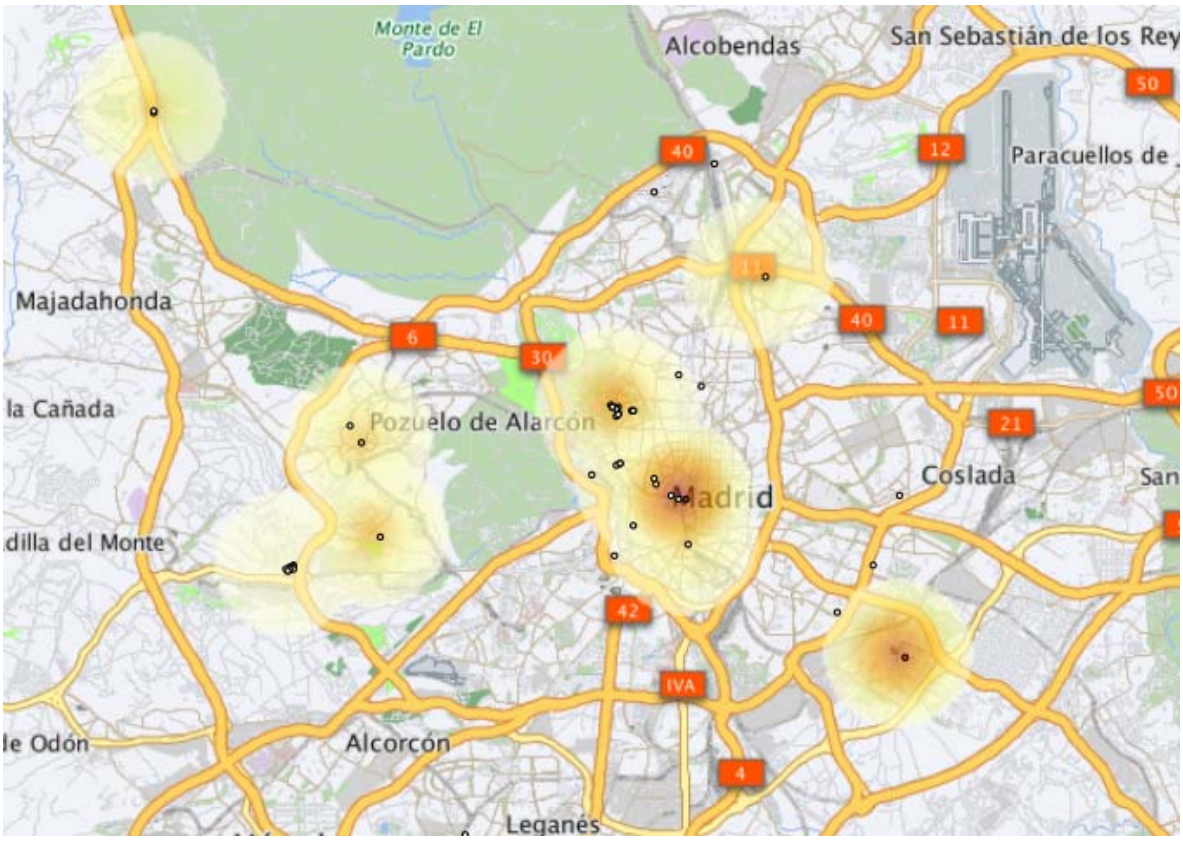


Figura 166. Tuits CEU en Madrid. Fuente: elaboración propia.

6.6.2.6 *Universidad Rey Juan Carlos*

Se han obtenido 12.888 tuits de los cuales el 24% eran positivos y el 5% negativos. Mediante la tabla 53 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 49. Usuarios más influyentes de la URJC. Fuente: elaboración propia.

from_user_name	DocumentSentime...	DocumentSentime...	followers_count (m...	friends_count (max)
ESNE	676	1.14	3,342	1,587
ESERP	123	1.38	10,690	519
ESIC Málaga	119	1.54	2,592	35
librodesignthinking	116	1.33	250	421
ESIC	107	2.00	12,947	419
ESIC Sevilla	94	0.79	984	232
ESIC Madrid	73	2.55	2,635	184
Residencia Santa Ana	69	1.19	130	387
EsicBarcelona	68	1.80	1,000	260
Esne Beltza	56	0.42	9,880	267
Ignacio de la Vega	56	1.00	1,024	692
xabier solano maiza	55	0.04	2,239	143
ESICValencia	52	1.82	1,326	272
Escuela TAI	48	1.57	2,667	845
dparente	44	2.04	1,381	585
Carlos González	42	-0.02	1,911	859
ICEMD	41	1.08	5,848	1,063
Esic Zaragoza	39	3.08	629	43
TOLMOS	38	0.80	1,583	622
BK Spanish	37	0.00	62	109

En la figura 167 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 167. Nube de palabras URJC. Fuente: elaboración propia.

La tabla 54 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 50. Tuits más influyentes de la URJC. Fuente: elaboración propia.

text	Record Count	DocumentSenti...	retweet_cou...
RT @Santi_ABASCAL: Ahora en la Universidad Rey Juan Carlos defendiendo l...	66	0.38	70
RT @metro_madrid: Mañana comienzan las obras en Metrosur entre Hospital ...	54	0.00	60
RT @carloscuestaEM: «¡Terroristas! !Fascistas!», gritan a Edurne Uriarte en la ...	54	0.00	57
RT @vox_es: Da comienzo el acto de Vox en La Universidad Rey Juan Carlos d...	41	0.00	45
RT @ESERP: Felicitamos a Guillermo de los Mozos, antiguo alumno de #ESER...	36	0.00	39
RT @vox_es: @ivanedlm en la Universidad Rey Juan Carlos:"los españoles a t...	35	0.00	35
RT @cocotelemadrid: Otra #Telemadrid es posible @salvemostelema en Unive...	34	4.00	34
RT @SalvaSuay: ¡Ultimo examen superado! Me despido de Esic Valencia hasta...	32	3.56	33
ESNE TV, El Sembrador Estados Unidos - http://t.co/CfF9U4CBNf	31	0.00	1
RT @CristinaiRascon: "Le nombran presidente de la escalera y dice que es Co...	29	0.00	142
RT @iDescubrelo: Investigadores de la Universidad Rey Juan Carlos han dem...	26	-3.75	31
RT @JPelirrojo: Me he colado en una clase de animación de videojuegos en @...	25	0.00	25
RT @JPelirrojo: Esto es lo que me ha dado tiempo a hacer en la clase de dibuj...	24	4.50	25
RT @muguruzafm: Quién manda aquí... nork agintzen du, hemen eta hor... Es...	20	-6.00	22
RT @Tele7Noticias: Mago de Oz,Esne Beltza,Efecto Pasillo también en fiestas ...	20	0.00	23
RT @ESICEducation: 55% de las empresas españolas vinculan su marca a Esp...	20	0.00	20
RT @NYChocolate: Para sobrellevar esta abdicación a los de la Universidad Re...	20	3.55	25
RT @jaretaldea: Mañana 2.Aniversario Iñigo Cabacas . Despues de la mani, ber...	19	0.00	21
Desaparece el único grado en Igualdad de Género: La Universidad Rey Juan C...	18	-4.00	0
RT @lamarea_com: Fuerte presencial policial durante la huelga de limpieza en ...	18	-4.38	18

La figura 168 se muestra la localización de los tuits a nivel nacional. Se puede observar que aparecen pocos tuits en numerosas localizaciones. El principal foco es Madrid. Aparecen algunos tutis concentrados en Barcelona y Bilbao. Por otra parte, numerosos tuits repartidos por el territorio nacional como en Sevilla, Valencia y Málaga.

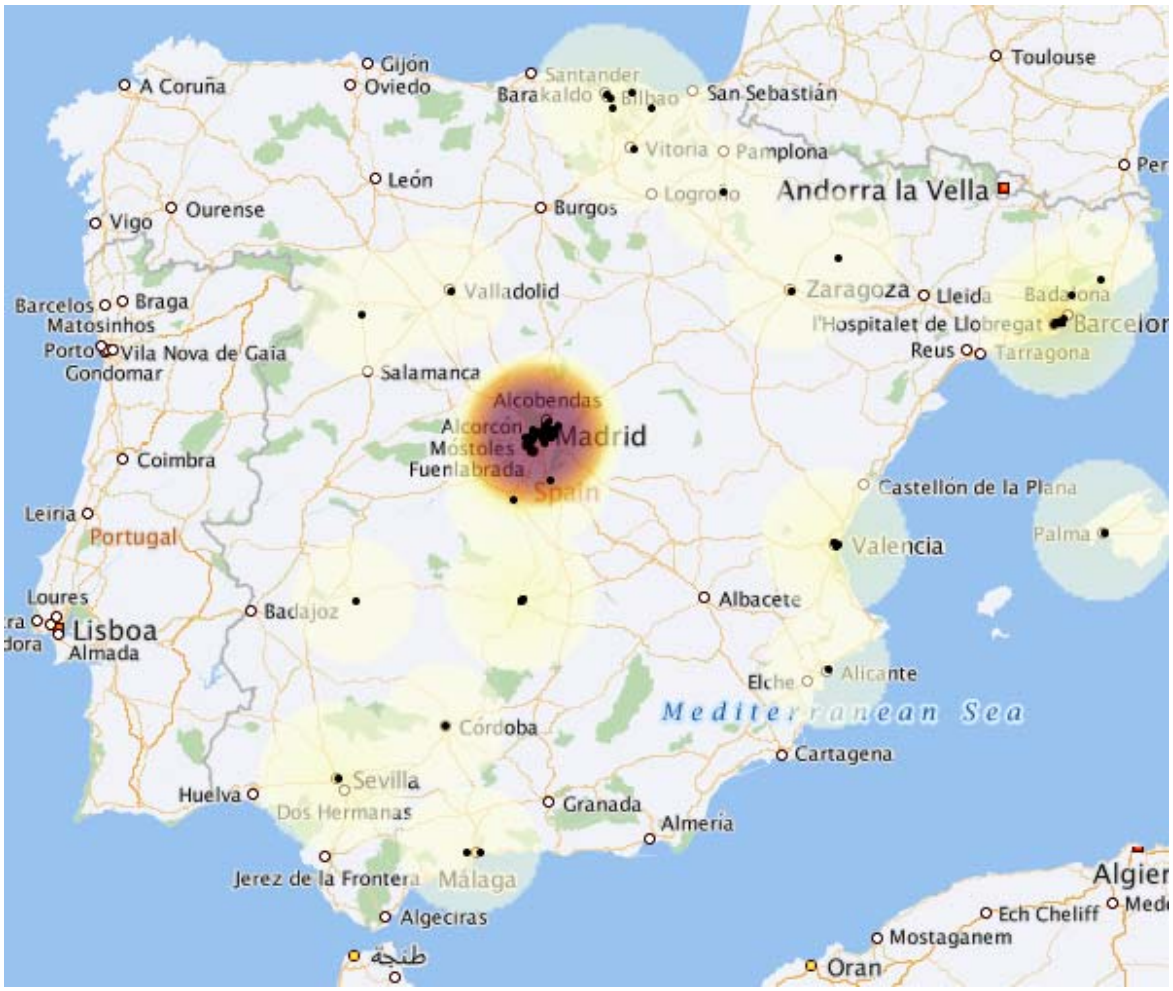


Figura 168. Tuits URJC en España. Fuente: elaboración propia.

En Madrid (figura 169) se muestran muchos tuits repartidos en diferentes localizaciones. La principal fuente proceden del ESIC, Pozuelo. Además aparecen tuits en Fuenlabrada (campus de Fuenlabrada), en Móstoles (campus de Móstoles) y en Vicálvaro (campus de Vicálvaro). Por otra parte, en Madrid capital se observa por la Avenida de Alfonso XIII una pequeña concentración de tuits debido al centro asociado ESNE y un gran número de tuits dispersos por el área metropolitana.

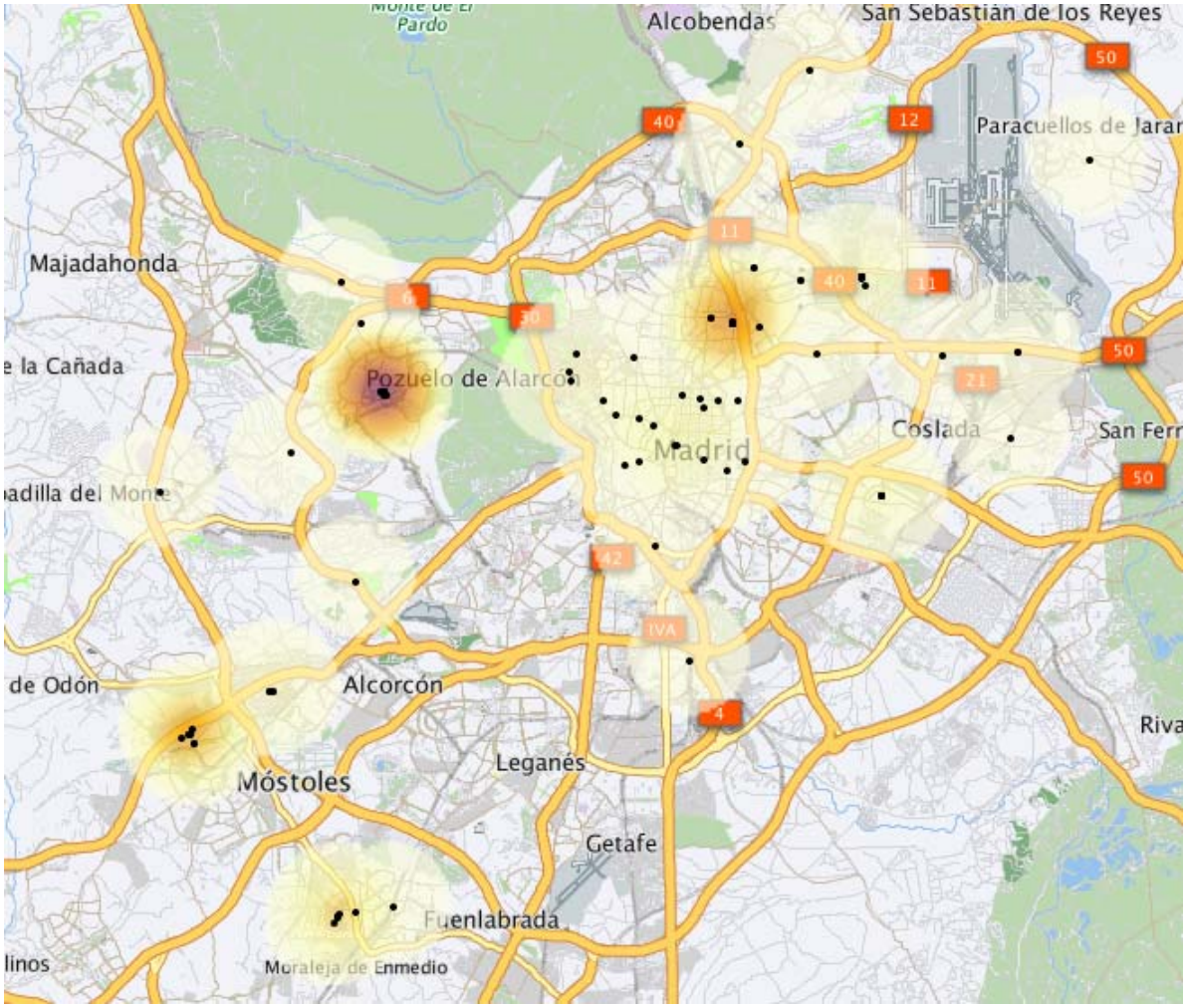


Figura 169. Tuits URJC en Madrid. Fuente: elaboración propia.

6.6.2.7 *Universidad Politécnica de Madrid*

Se han obtenido 12.745 tuits de los cuales el 22,3% eran positivos y el 12,9 % negativos. Mediante la tabla 51 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 51. Usuarios más influyentes de la UPM. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Politecnica Madrid	155	1.21	18,067	98
; ~Ƹtɛяя '♥	103	1.08	542	315
GIA-UPM	99	-0.31	1,322	1,701
José Miguel Atienza	83	1.60	200	104
ETSI MINAS Y ENERGÍA	59	1.16	1,557	1,888
OTRI UPM	53	1.07	920	405
MDRGS	50	1.15	60	202
GSI-UPM	46	1.15	120	12
redcei	43	1.13	606	2,001
Amigos Ingeniería	41	0.83	6,456	5,175
ETSI Informáticos	35	1.21	748	28
Lo Mejor de Madrid	33	1.74	2,036	566
bibcaminosupm	33	1.10	215	307
i.	32	-4.00	10	36
Biblioteca UPM	32	0.31	1,313	386
rott wailer	32	-2.53	60	129
Shama . =)	30	0.50	211	114
A.U. La Siega	29	-2.39	310	416
El Metro de Panamá	29	1.05	38,503	1,217
ETSEMoficial	29	1.99	1,126	62

En la figura 170 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 170. Nube de palabras UPM. Fuente: elaboración propia.

La tabla 52 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 52. Tuits más influyentes de la UPM. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @germanin: Gracias UPM por empezar examenes 1 mes despues que CIII y URJC, acabar...	62	0.00	69
RT @La_UPM: La UPM, la la más valorada de las universidades politécnicas españolas según...	49	6.00	58
RT @educaINTEF: Aprendizaje Basado en Problemas, Guías rápidas sobre nuevas metodolog...	35	1.00	39
RT @SamsungEspana: Haz de tu adicción al móvil tu modo de vida: fórmate en el #SamsungT...	34	-6.00	36
RT @Asambleaagronom: Un compañero de la escuela habla para @TeleK_tv sobre la huelga, ...	32	-4.38	32
RT @hermajoan: Después de 8 h de mediación UPM sigue queriendo cargarse el Conv metal ...	31	4.50	31
RT @TuUPM: Hay dos tipos de universitarios: los que acaban la universidad en nada y los que...	30	0.00	31
RT @Sr_Hobbs: El T-Rex de la UPM que la custiodia para que solo entren mentes superiores l...	26	3.00	27
RT @elbosquep: PRÓXIMO CAPÍTULO Sábado 12 de abril 15,05 en @la2_tve FAUNA AMENAZ...	25	0.00	33
RT @elpaisuy: ÚLTIMO MOMENTO - Argentina llevará a La Haya a Uruguay por aumento de pr...	25	0.00	25
RT @cadenachori: En el segundo tiempo los de Bosnia se cambian el nombre a UPM y lo dan ...	24	0.00	24
RT @La_UPM: La UPM, la universidad española que más patentes presentó en 2013 http://t.co...	24	0.00	27
RT @bomberosvina: Unidades CJ2-CJ3-T1-U13-U31-U51-U61-U71-U73-U81-U92-UPM del CB...	23	0.00	23
RT @misaqui: Según Linkedin, somos 85K titulados de @La_UPM en la red. Así estamos distr...	22	4.50	30
RT @Acaf77Arias: 400 millones de dólares y contando,nos costo esto. Ni el mundial ni UPM lo...	22	0.00	22
RT @camusbeto: @cristin04787201 No. No hubo absolutamente nada con el loro barranquero...	22	0.00	24
RT @martaoleac: En @La_UPM el día 6 vamos a arrasar. http://t.co/p3hJcNfBpx ... @eui_upm...	22	0.00	24
RT @elbosquep: PRÓXIMO CAPÍTULO Sábado 26 de abril 15,05 en @la2_tve FAUNA AMENAZ...	21	0.00	21
RT @BomberosdeChile: CB VIÑA: Unidades CJ2-CJ3-T1-U13-U31-U51-U61-U71-U73-U81-U9...	19	0.00	19
RT @jesuspalop: Muy currada la estructura de los de diseño de la Etsidi este año en cartón, si...	19	0.00	20

La figura 171 se muestra los tuits a nivel nacional, ubicados en Madrid principalmente. Aparecen 2 focos menores en Barcelona y Alicante. Además un tuit en Jaén y dos en Vitoria.

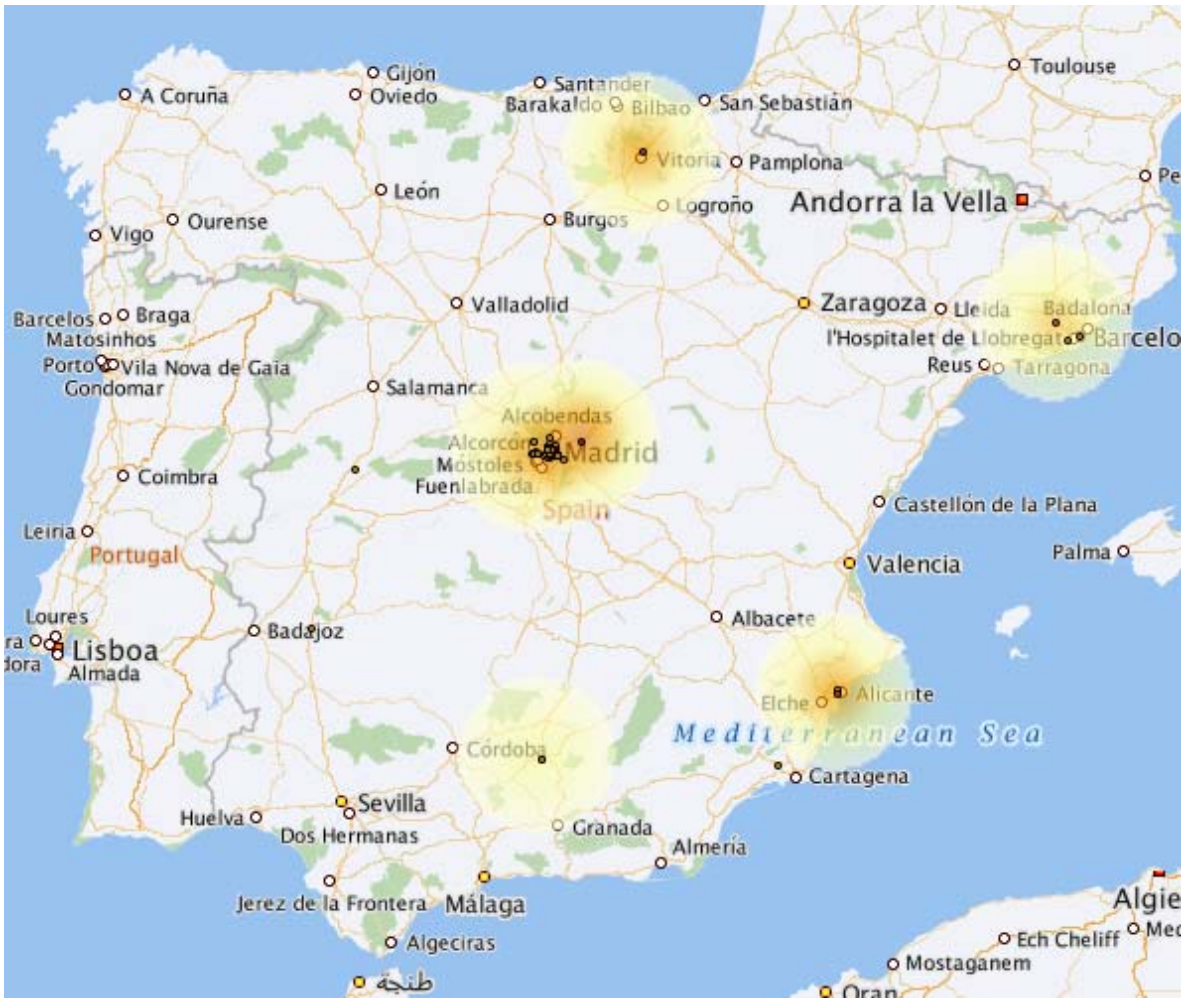


Figura 171. Tuits UPM en España. Fuente: elaboración propia.

En Madrid (figura 172) se muestran tuis en la zona de Campus Universitario y en la zona de Vallecas (campus Sur de Vallecas). Por otra parte, aparecen algunos tuits en Boadilla del Monte, en el campus de Montegancedo. En la zona metropolitana de Madrid aparecen tuits dispersos.

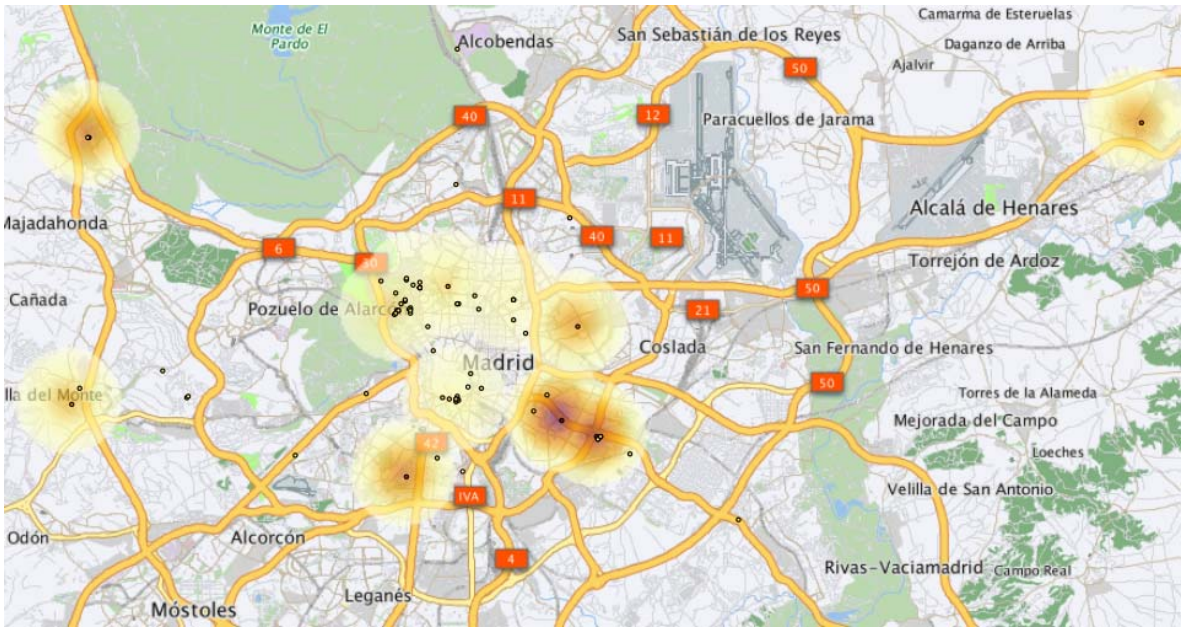


Figura 172. Tuits UPM en Madrid. Fuente: elaboración propia.

6.6.2.8 Universidad Camilo José Cela

Se han obtenido 10.626 tuits de los cuales el 27% eran positivos y el 4% negativos. Mediante la tabla 47 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 53. Usuarios más influyentes de la UCJC. Fuente: elaboración propia.

	from_user_name	DocumentSentime...	DocumentSentime...	followers_count (m...	friends_count (max)
	U-tad	968	1.23	2,808	1,841
	Carmen Magaña	301	1.33	167	131
	UCJC	282	1.45	9,283	1,687
	Carlos Soriano	276	1.39	75	114
	Carlos Fuente	221	1.40	1,953	1,300
	Medina Media	186	1.02	1,495	402
	DigitalBusinessU_tad	174	1.22	709	862
	Silvia Medrano	162	1.22	129	188
	Alma de las Empresas	151	1.45	1,259	58
	isPE	148	1.21	2,353	1,373
	Marieta	111	1.46	236	334
	Laura Raya	94	1.23	85	79
	¡¡Blanca¡¡	84	-0.02	308	252
	Gloria Campos	83	1.13	1,437	737
	Rafael Ramiro	78	1.65	244	146
	Domingo Lopez	75	1.44	728	54
	Gema P	66	1.37	547	560
	Gamaliel Martinez	60	1.41	254	240
	Francisco Gallego	58	0.97	854	541
	Jose Maria Font	56	1.52	58	62

En la figura 173 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 173. Nube de palabras UCJC. Fuente: elaboración propia.

La tabla 48 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 54. Tuits más influyentes de la UCJC. Fuente: elaboración propia.

RT @cabezacuco: #ElAlmadeU_tad @U_tad gracias a lo que U-tad me ha dado...	208	0.00	292
RT @Caroli636: #ElAlmadeU_tad @U_tad nos enseña que las pequeñas ideas, ...	194	3.00	284
RT @SirSeanHD: En menos de un año estar ya programando videojuegos en ...	134	0.00	141
RT @x1sara1x: Por tener un futuro en lo que me gusta @U_tad #ElAlmadeU_tad	66	4.00	67
RT @cabezacuco: #ElAlmadeU_tad gracias a lo que U-tad me ha dado, soy Co...	61	0.00	66
RT @aviudaserrano: Que te despidan de la Universidad Camilo José Cela, rast...	54	-4.50	61
RT @emilio_mas: En el stand de @U_tad del InnGames con @damiconeme htt...	38	0.00	43
RT @UCAMMurciaCF: UCAAAAAMPEONES! La selección de la @UCAM ha ga...	30	6.00	30
RT @Caroli636: #ElAlmadeU_tad! El nivel tan alto de u-tad nos enseña a realiz...	24	0.00	33
RT @dieghobonilla: Gracias a @U_tad @InnGamesOficial puedo disfrutar de ...	20	5.50	26
RT @CarlosFuente1: Según diario El Mundo, el Grado de Protocolo y Eventos ...	18	0.00	18
RT @UcjcEnfurecida: Despiden a 3 magníficos profesores durante exámenes. ...	15	0.00	28
RT @ipdolset: Alumnos y empleados de @U_tad serán los protagonistas de ...	15	0.00	15
RT @DeportesUCJC: ¡Final del partido! La UCJC vuelve a una final del Campe...	15	0.00	15
RT @U_tad: ¿Te gustaría pertenecer a una comunidad universitaria única? Par...	15	0.00	2
RT @MadridEmprende: De la idea a la empresa, Cursos de Verano en UCJC: Cr...	14	0.00	12
RT @AlmaEmpresas: ¿Sabes para que sirven las gafas oculus rift? ¿Y en qué ...	14	0.00	14
RT @MedinaMediaTV: @U_tad cuenta con más de 250 profesores #docencia #I...	14	0.00	15
RT @U_tad: 330 empresas y el 67,4% se concentra en Madrid, Barcelona y Vale...	13	6.00	13
RT @CarlosFuente1: @ispeprotocolo @universidadcjc Análisis del protocolo ...	12	0.00	13

La figura 174 se muestra los tuits a nivel nacional. El principal foco es en Madrid. Además aparecen algunos tuits dispersos por el territorio como en Vigo, Oviedo, Zaragoza, Barcelona, Valencia, Badajoz, Córdoba y Málaga.

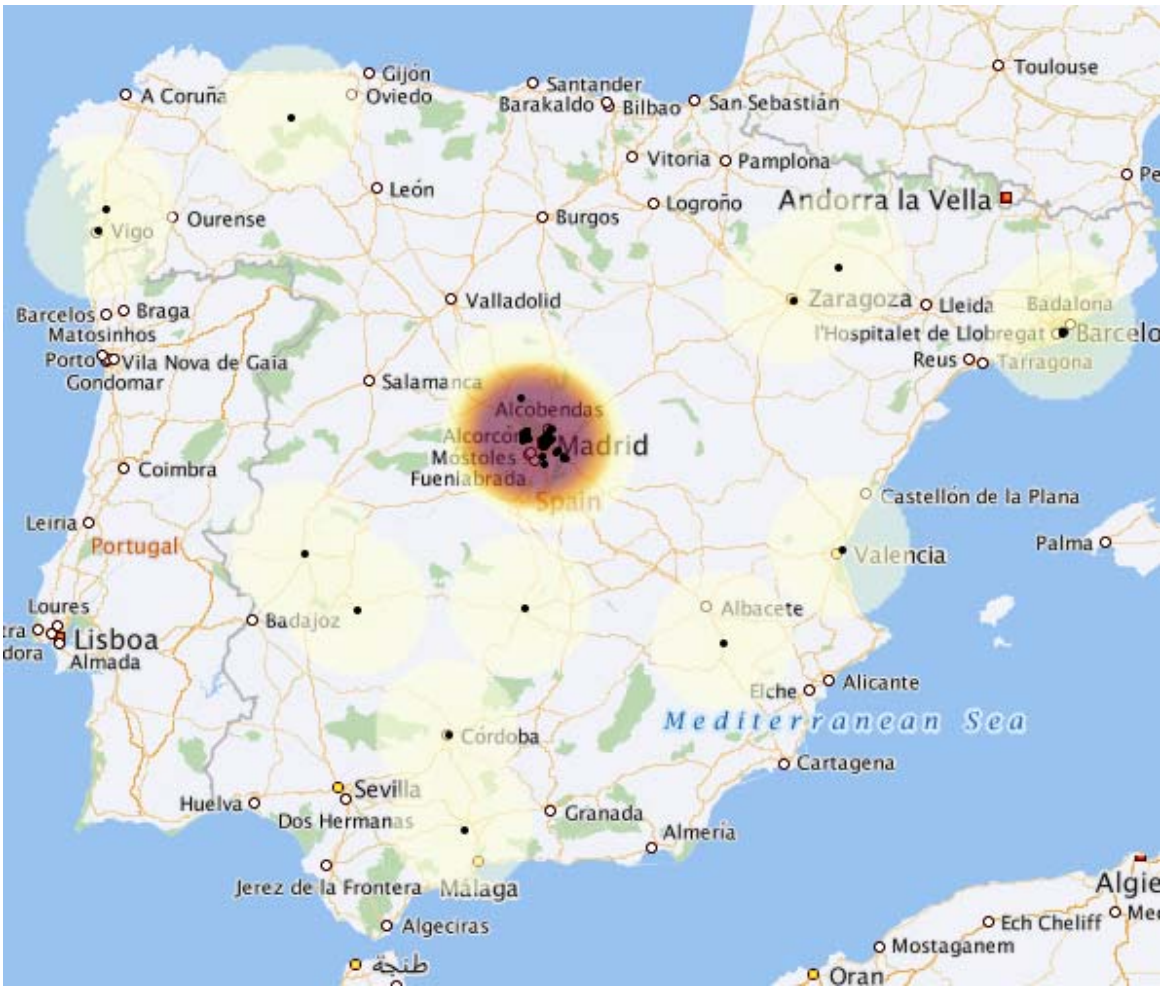


Figura 174. Tuits UCJC en España. Fuente: elaboración propia.

En Madrid (figura 175) se observan un gran número de tuits en el centro, en la sede de Ferraz, en la sede de Mar de Cristal y en el campus de Villafranca del Castillo. Además existen otros focos como en Las Rozas (centro U-TAD) y numerosos repartidos por el área metropolitana de Madrid.

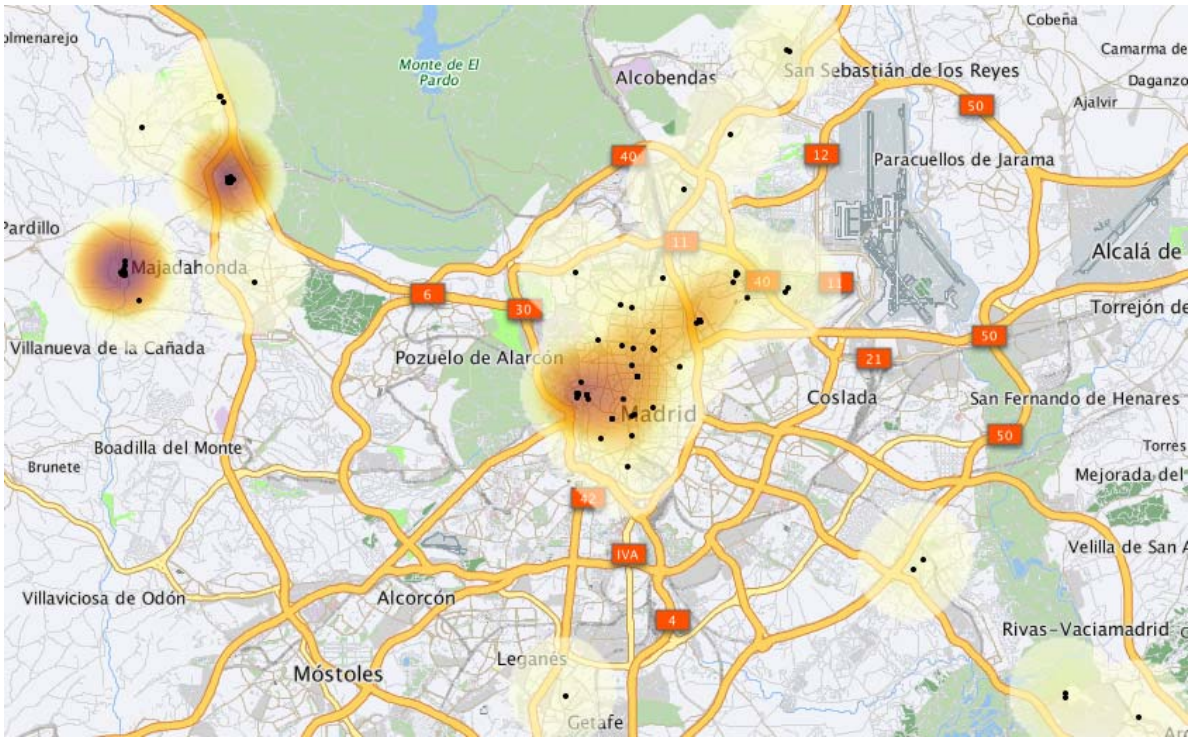


Figura 175. Tuits UCJC en Madrid. Fuente: elaboración propia.

6.6.2.9 *Universidad Carlos III de Madrid*

Se han obtenido 7.076 tuits de los cuales el 19,2% eran positivos y el 7,7% negativos. Mediante la tabla 55 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 55. Usuarios más influyentes de la UC3M. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Aula de las Artes	79	0.82	1,307	484
PIC Leganés uc3m	76	0.88	1,855	173
biblioteca_uc3m	61	0.94	6,850	441
Sección Sindical CGT	44	0.27	716	275
FármacoEconomía UC3M	44	0.20	277	269
Delegación HCD UC3M	40	1.42	718	664
El Club de Guapo	39	0.00	4,254	4,605
Delegación FCSJ UC3M	38	1.59	465	340
UC3M	31	1.33	17,138	137
Divulga UC3M	29	1.06	2,562	480
Fiesta Charlie's	25	1.74	2,556	1,810
Emilio Olías Ruiz	23	0.43	1,796	2,001
Postgrado UC3M	21	0.68	639	565
The Observer	20	-0.06	1,590	485
Derechos Humanos	19	0.26	2,827	2,337
Damos Voz	19	1.52	68	24
Crossing Stages	19	0.58	63	139
redcei	18	0.60	605	2,001
Sr. Presidente	18	0.33	500	567
DColaborativoMadrid	18	1.81	112	141

En la figura 176 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 176. Nube de palabras UC3M. Fuente: elaboración propia.

La tabla 56 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 56. Tuits más influyentes de la UC3M. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @charlies_uc3m: RT si el Lunes 2 de junio en la Fiesta Charlie's UC3M en la Sala Marco Al...	163	0.00	176
RT @QueenCarlosIII: El Madrid le está haciendo al Bayer lo mismo que nos hace la #UC3M a n...	86	0.00	94
RT @20m: Un centenar de estudiantes se encierra en la Universidad Carlos III de Madrid http:/...	80	0.00	86
RT @charlies_uc3m: LUNES 2 DE JUNIO: FIESTA CHARLIE'S UC3M DE FIN DE EXÁMENES E...	53	0.00	56
RT @caroolgonzalezz: Hoy en Cuarto Milenio el alumno que aprobó un final de 2014 en la UC3M	46	0.00	50
RT @ObserverUC3M: ¿Y tú, por qué decidiste estudiar en la UC3M y no en otra universidad m...	36	0.00	126
RT @QueenCarlosIII: En la #UC3M lo más normal es dar temario que entra en el examen dura...	33	0.00	37
RT @ConCienciaUC3M: En #uc3m Leganés @Partido_X @Equo @iunida @escanosenblanco ...	32	0.00	33
RT @ObserverUC3M: ¿Quiénes somos? ¿De dónde venimos? ¿A dónde vamos? ¿En qué mo...	29	0.00	32
RT @charlies_uc3m: Ya tenemos.. EL FLYER DE LA FIESTA! LUNES 2 DE JUNIO, FIESTA CHA...	23	0.00	23
RT @JustNaitmer: Un alumno cualquiera de la UC3M http://t.co/YMHDCfhfBu	23	0.00	26
RT @QueenCarlosIII: El palo que pegó Neymar en el minuto 90 fue para el Madrid como un 5.0 ...	23	0.00	24
RT @ObserverUC3M: Lo bueno de estudiar en la UC3M es que como mucho te tiras sin follar ...	22	4.00	22
RT @MalekJandali: Viva Free #Syria from Carlos III University #Madrid #Spain @leila_na @uc...	21	0.00	23
RT @SaraDeMiguel: el nivel de los finales de este año no es ni medio normal #uc3m	21	0.00	25
RT @ConEduMadina: Hoy @EduMadina compartirá ideas y reflexiones en la Universidad Carl...	20	0.00	23
RT @jmc_nz: ¡Ven a conocer el proyecto de @UPyDEuropa para Europa de la mano de Sosa ...	19	0.00	19
RT @charlies_uc3m: RT SI QUIERES SABER FECHA Y LUGAR DE LA FIESTA CHARLIES UC3...	17	0.00	18
RT @uc3m: Hoy la @UC3M cumple #25años. Gracias a todos los que habéis recorrido este c...	16	0.00	80
RT @uc3m: Los Rectores piden la finalización de las medidas excepcionales al estudio, la acti...	15	7.50	16

La figura 177 se muestra la localización de los tuits de la Universidad Carlos III. El principal foco de concentración aparece en Madrid.

Figura 177. Tuits UC3M en España. Fuente: elaboración propia.

En Madrid (figura 178) la mayoría de los tuits se concentran en el campus de Getafe y Leganés. Además aparece un número disperso de tuits por Ciudad Universitaria y por el área metropolitana.

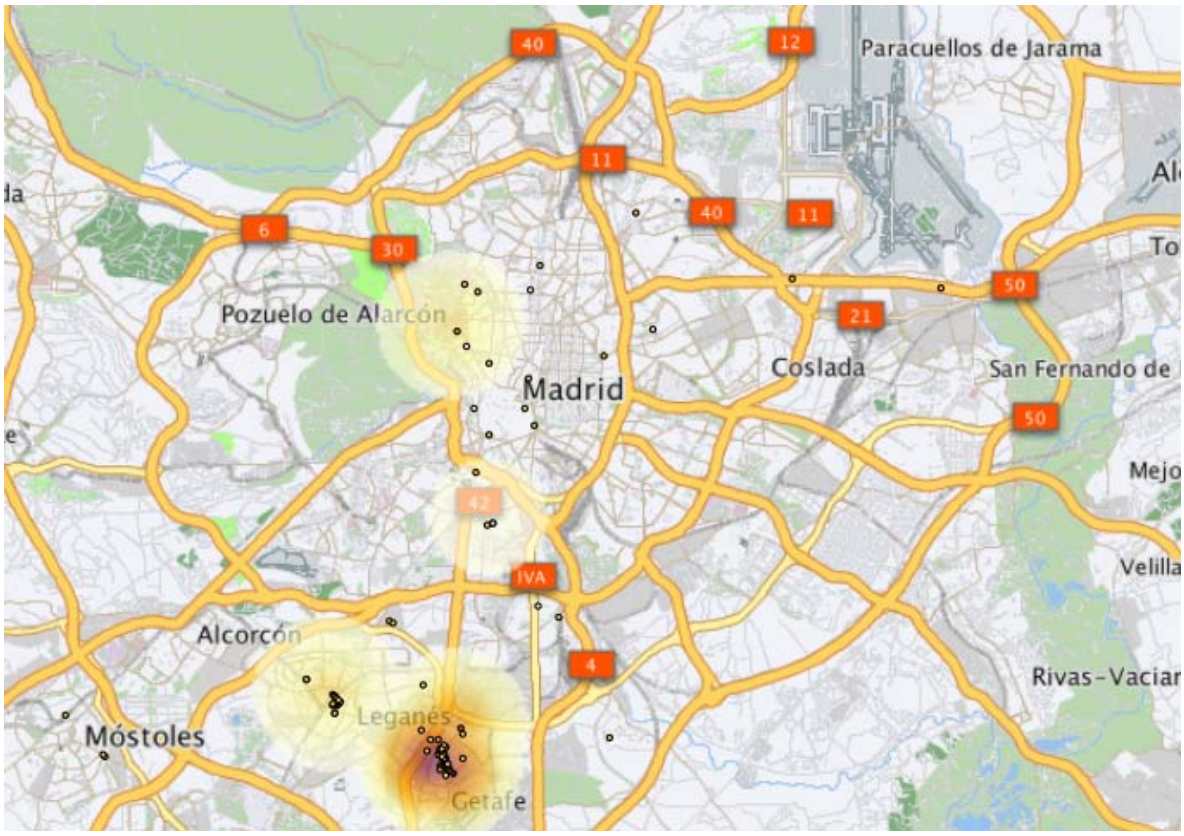


Figura 178. Tuits UC3M en Madrid. Fuente: elaboración propia.

6.6.2.10 Universidad Europea de Madrid

Se han obtenido 4.925 tuits de los cuales el 26,5% eran positivos y el 5,4% negativos. Mediante la tabla 57 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 57. Usuarios más influyentes de la UEM. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Nancy Avianco	104	1.39	45	9
Universidad Europea	88	1.69	28,022	4,345
Jorge Ramiro Pérez	50	0.97	1,350	882
Rachel Ornay	49	1.61	37	16
Derecho UE	41	0.34	339	170
alexadra	41	0.00	2,104	1,768
Miríada X	32	2.46	27,784	173
Escuela I. Protocolo	32	1.07	3,717	2,099
Cristina Alvarez	30	0.08	584	669
UECanarias	30	2.54	421	267
Biblioteca CRAI UEM	27	1.80	97	71
Beatriz Martínez	25	1.12	97	142
Pedro Lara	24	0.70	36	14
Compromiso Social UE	22	2.31	1,087	922
rafael fontán tirado	20	0.20	1,419	473
***C. Gloria	19	0.86	225	932
Ciudadanos y Salud	19	-0.18	645	2,001
D@nicr@ck	18	1.25	53	112
Real Madrid UE	18	0.18	109	28
Premruethai S.	17	0.00	83	85

En la figura 179 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

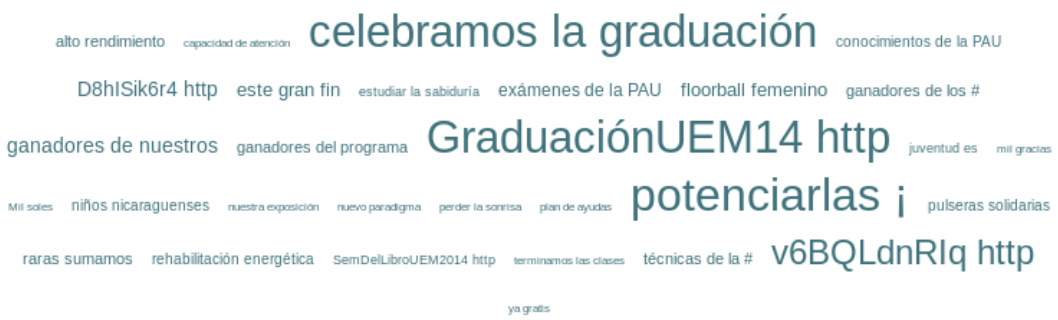


Figura 179. Nube de palabras UEM. Fuente: elaboración propia.

La tabla 58 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 58. Tuits más influyentes de la UEM. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @UEuropea: Comprueba si estás preparado para la #PAU con la #QuestionsPAU de Andro...	499	0.00	795
RT @UEuropea: ¿Estás preparado para la #PAU? Reta a tus amigos con la App #QuestionsPA...	45	0.00	111
RT @UEuropea: ¿Conoces tus cualidades? En la Universidad Europea te ayudamos a potenci...	39	3.00	71
RT @UEuropea: ¿Sabías que celebramos la graduación en el campo del @RealMadrid? ➡ http...	39	3.00	44
RT @UEuropea: Descubre la Universidad en la que alcanzar tu mejor yo. ¡Síguenos! http://t.co...	35	0.00	48
RT @Defensagob: II Simposio internacional sobre entrenamiento para ambientes extremos, ...	29	6.00	30
RT @alvaromerino: El poder de abrazarnos @LiderHazGO @Retrocycle @UEuropea http://t.c...	27	0.00	31
RT @UEuropea: ¿Qué alimentos te ayudan a concentrarte? #Salud Vía @LaureateConnect htt...	27	0.00	30
RT @UEuropea: Así se prepararon para la PAU los 50 ganadores de nuestro #CallOfFuture! ➡...	23	12.00	23
RT @UEuropea: Los 50 alumnos del #CallOfFuture 2014 ya durmieron hoy en nuestro campus...	22	0.00	23
RT @UEuropea: El acto de #GraduaciónUEM14 de este año será épico. No importa qué hayas ...	20	3.00	22
RT @UEuropea: ¿Añadirías algún consejo a estos para afrontar la #PAU? http://t.co/Bcdl32dk...	20	0.00	21
RT @sosFLOORBALLfem: La @UEuropea acogerá el 17 y 18 de mayo el Campeonato de Esp...	18	0.00	19
RT @UEuropea: ¿Quieres conocer los 10 ganadores de nuestros #PremiosJES de 2014? ➡ ht...	16	6.00	16
RT @UEuropea: ¡Reta a tus amigos con nuestra App! qUEuestionsPAU (Android) ➡ http://t.co/...	15	0.00	30
RT @UEuropea: Nuestros 50 elegidos de @CallOfFuture simulan hoy los exámenes de la PAU...	14	3.50	14
RT @MariaGallego95: Voy a la universidad más bonita del mundo @UEuropea http://t.co/GlVH...	13	6.00	13
RT @patri_judo: Soy la única que me muero por ver el video que está montando Hugo? #CallO...	12	0.00	12
RT @UEuropea: ¿Una tesis doctoral sobre el carnaval gaditano? Sí, es de @NSacaluga ➡ http...	12	0.00	13
RT @UEuropea: #CallOfFuture Aquí ➡ http://t.co/kcte2ZtHcf los 50 ganadores del programa d...	11	5.25	11

La figura 180 se muestra la localización de los tuits a nivel nacional. El principal foco se sitúa en Madrid. Además en mucha menor cantidad aparecen algunos tuits en Salamanca hablando sobre la Universidad Europea de Madrid.

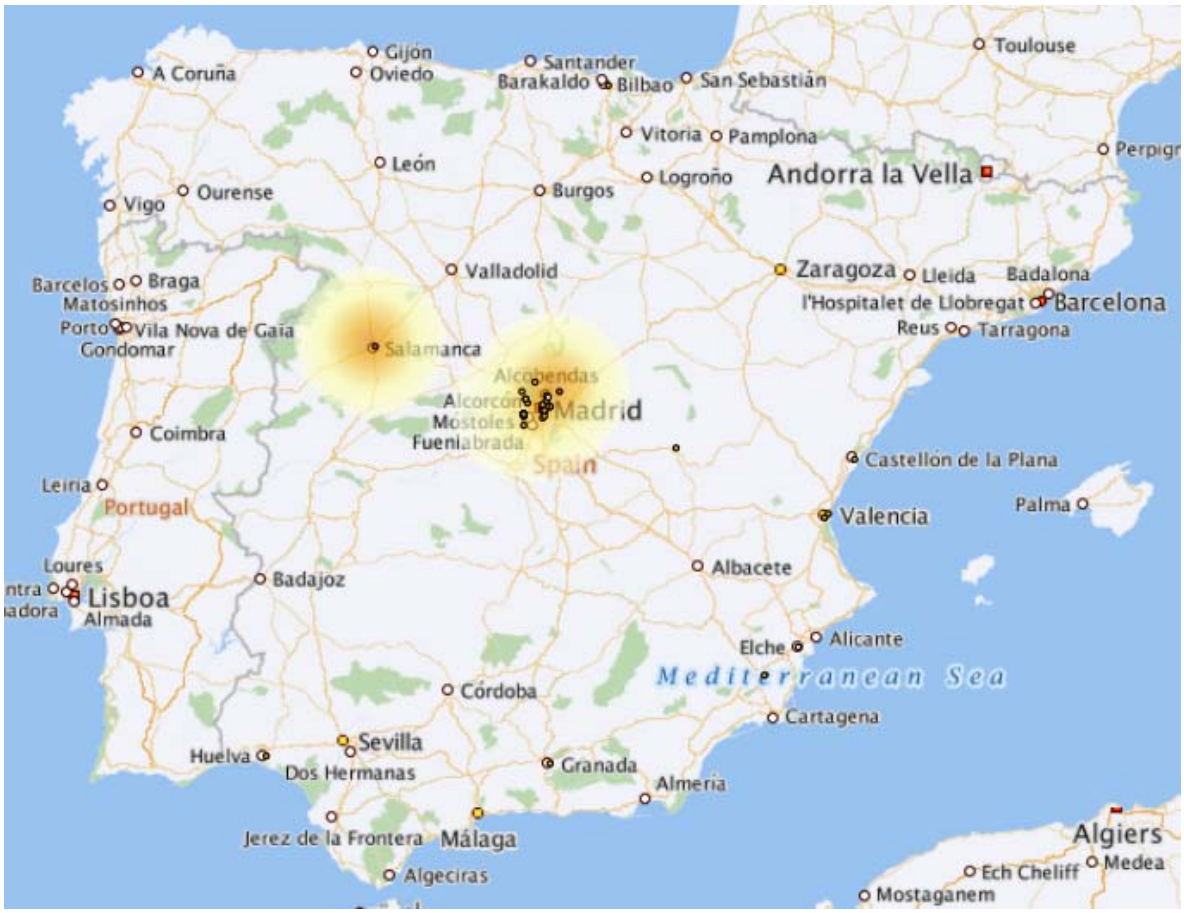


Figura 180. Tuits UEM en España. Fuente: elaboración propia.

En Madrid (figura 181) se ubican principalmente en el campus de Villaviciosa de Odón. En el área metropolitana se observa una importante concentración en algunas de sus escuelas de postgrado, IEDE (Alcobendas) y PROY3CTA (zona Retiro). Se muestran tuits dispersos por la zona noreste de Madrid.

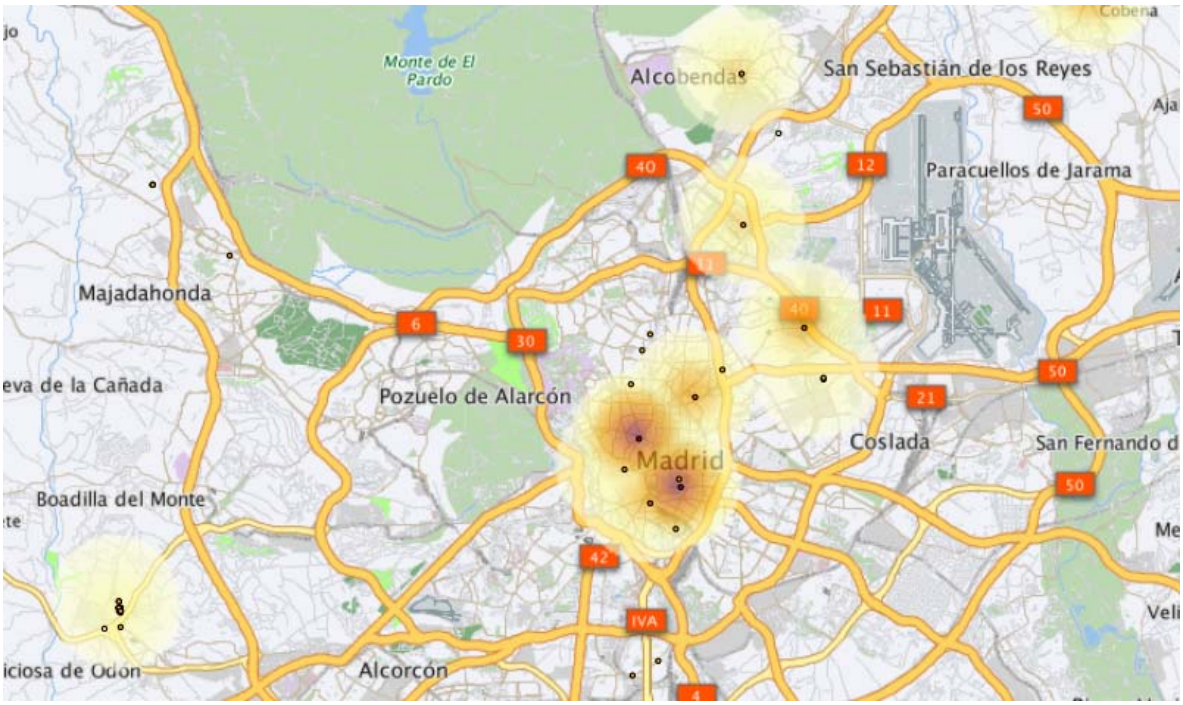


Figura 181. Tuits UEM en Madrid. Fuente: elaboración propia.

6.6.2.11 Universidad Antonio de Nebrija

Se han obtenido 4.280 tuits de los cuales el 31,9% eran positivos y el 2,3% negativos. Mediante la tabla 59 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 59. Usuarios más influyentes de la Nebrija. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Nebrija BS	229	1.50	1,333	252
Nebrija Universidad	203	2.44	7,574	1,899
Nancy Avianco	171	2.20	44	9
Rachel Ornay	98	2.32	36	16
tecnicaf1	48	1.23	1,454	170
Guideo	36	3.42	2,087	2,144
EUTIP	33	1.07	181	326
Romina Re	31	2.11	815	130
電子書籍セール情報	30	0.00	21,876	21,881
FernandoGomezBlanco	28	0.90	1,971	711
nebrijasolidaria	27	1.35	144	501
Kvothe	26	0.49	237	264
María García terrón	26	1.63	187	210
Nebrija LA	25	1.64	664	366
Nebrija RACING	25	1.19	146	122
El Sory Soriano	25	0.59	371	335
Centímetros Cúbicos	24	0.25	5,408	479
電子書籍情報	24	0.00	22,187	22,193
•Irene®©•	21	0.76	327	304
Marga Víctor Morales	21	1.61	244	518

En la figura 182 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 182. Nube de palabras de la Nebrija. Fuente: elaboración propia.

La tabla 60 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 60. Tuits más influyentes de la Nebrija. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
Universidad Nebrija en la Fórmula SAE: Los futuros ingenieros de la Universidad Nebrija parti...	23	0.00	0
RT @Nebrija: Y los flamantes vencedores, del Colegio Montpellier @debate_MONTPE #nebrija...	22	4.50	22
Estrenamos nueva web @Nebrijabs y @Nebrija #nuevawebNBS #somosnbs ¿no nos conoces...	22	0.00	1
Ya está disponible la nueva web de @Nebrijabs y @Nebrija #nuevawebNBS ¿Habéis entrado y...	18	0.00	1
RT @Nebrija: Enhorabuena al colegio Santa Teresa de Jesús! Ganadores de este XVI Premio ...	14	0.00	14
RT @TurismoAsturias: Tras Premios Alimara y Nebrija Tourism, hoy recogemos premio @atr...	13	10.00	15
RT @CCubicos: La @FsaeNebrija implica a jóvenes ingenieros en el diseño, construcción y p...	12	0.00	14
RT @puntalproduc: El domingo a las 10:00h en @CCubicos hablaremos con los alumnos de ...	12	0.00	14
RT @GobEx_Fomento: La Ruta del Jamón Ibérico gana el premio nacional Nebrija Tourism Ex...	11	5.00	14
¿Quieres conocer la nueva web de Nebrija Business School? Ven a visitarnos @Nebrija @Neb...	11	0.00	1
RT @NebrijaLenguas: A menos de un mes para la celebración del II #congresolenguas en la ...	10	4.00	10
RT @EduardBueno: Ya estoy en @Nebrija para hablar d ingeniería de competición con un de l...	10	0.00	10
RT @martamontero: Atención noticia!!Finalistas Nebrija @ErreQr y @Diverdentix !!Enhorabu...	10	0.00	10
RT @Extremadura_tur: La @RutaJamnIbrico, el mejor producto turístico del país http://t.co/e2...	9	6.00	10
Estamos preparando el XIX Foro de Empleo "Nebrija Profesional" para el próximo 29 de abril ...	9	0.00	1
La Fiesta del Deporte del Club Nebrija y del Club de Deportes ha llegado a @nebrija y @nebrij...	9	0.00	0
RT @albacarrascu: En que momento le tuve que dar yo mi número a la universidad de nebrija....	9	0.00	9
RT @CCubicos: El domingo a las 10:50h en @antena3com @EduardBueno hablará con @emil...	9	0.00	9
RT @CCubicos: El Máster en Competición de @Nebrija te da acceso al mundo de las carreras...	9	0.00	9
¿Todavía no formas parte de la comunidad Nebrija en Linkedin? ¡Te estamos esperando! http:/...	9	0.00	0

La figura 183 se muestra a nivel nacional la localización de los tuits. El principal foco se encuentra en Madrid. Repartidos por la geografía se encuentran algunos tuits que hablan sobre la Universidad Antonio de Nebrija en Almería, Ceuta, Cádiz, Oviedo y Logroño.

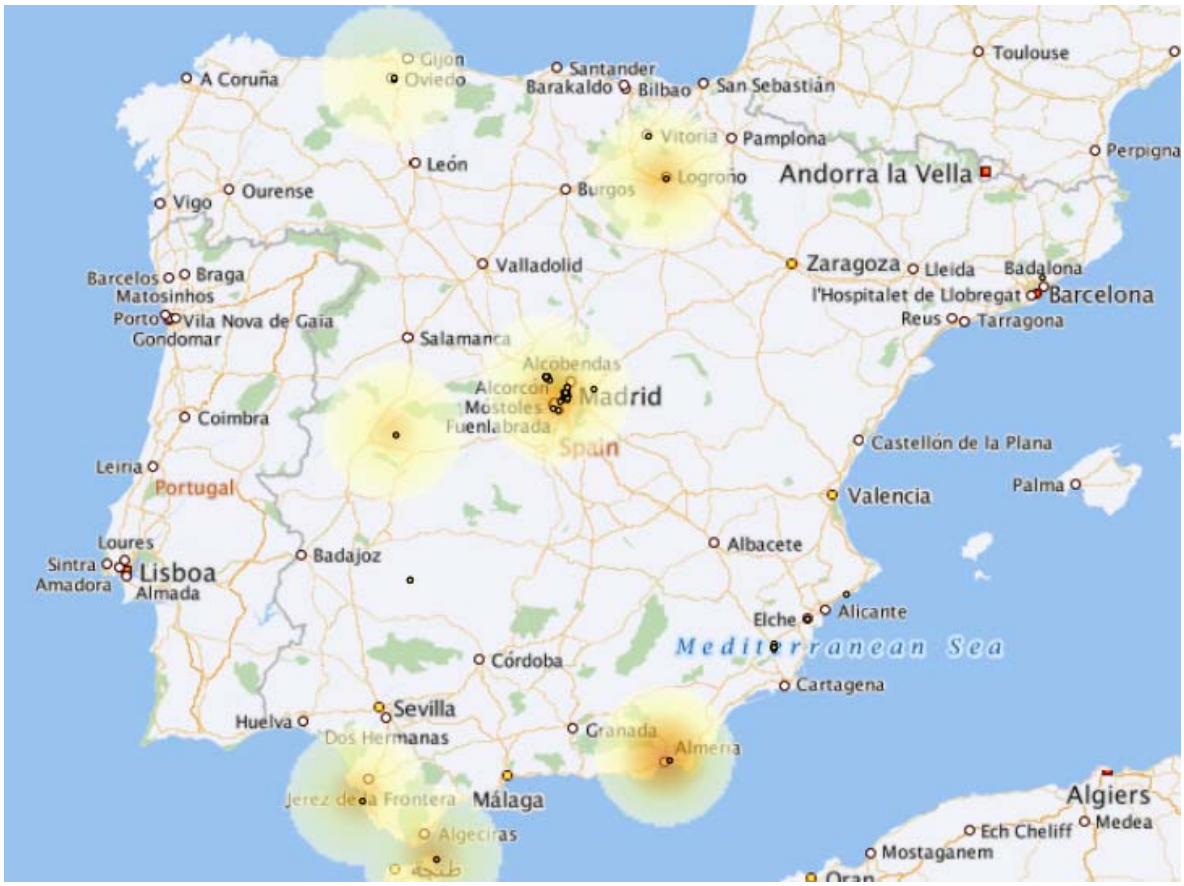


Figura 183. Tuits Nebrija en España. Fuente: elaboración propia.

En Madrid, figura 184, se sitúan algunos tuits en el campus de la Berzosa (noroeste de Madrid), por la zona del Centro de Estudios Garrigues (Recoletos) y en el campus de la Dehesa de la Villa (Moncloa).

La concentración de tuits del Paseo del Extremadura se trata de algunas personas que ha publicado sobre “los nervios de la prueba de admisión en la Nebrija”.

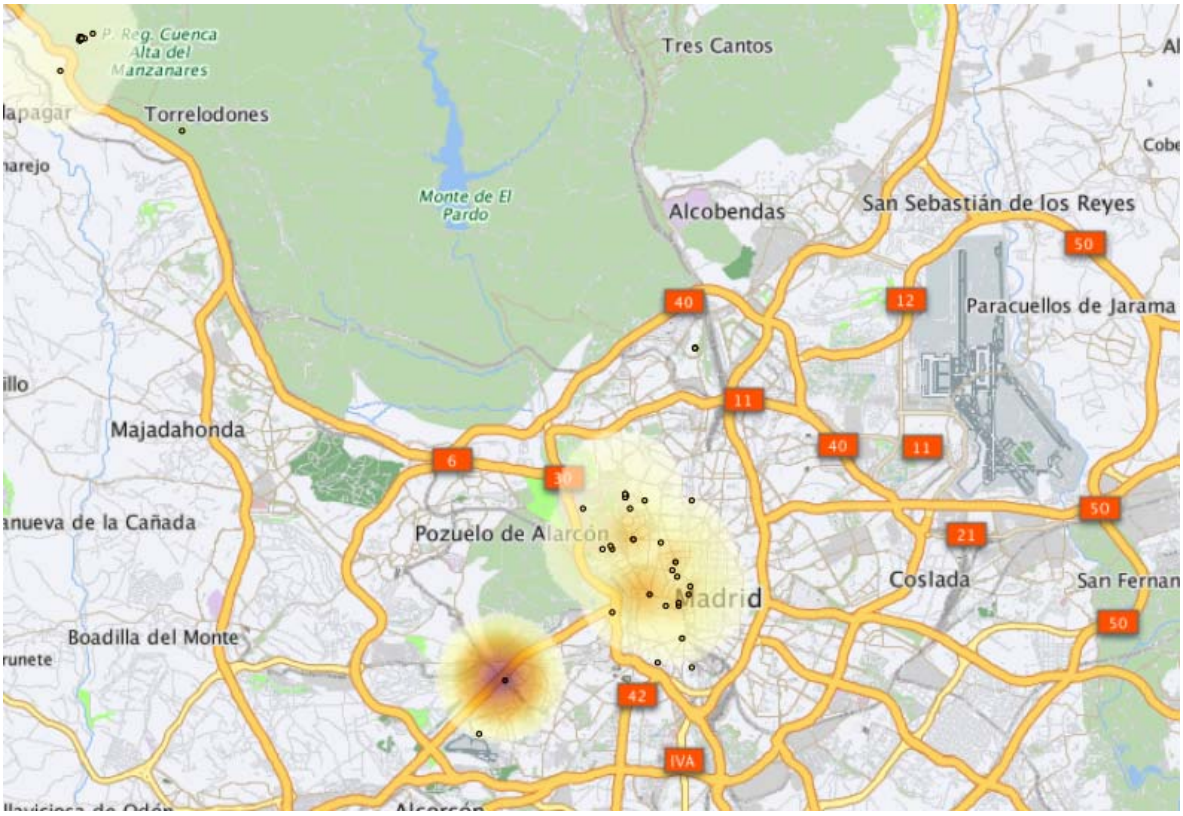


Figura 184. Tuits Nebrija en Madrid. Fuente: elaboración propia.

6.6.2.12 Universidad Francisco de Vitoria

Se han obtenido 4.200 tuits de los cuales el 30,1% eran positivos y el 3,6% negativos. Mediante la tabla 61 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 61. Usuarios más influyentes de la UFV. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Francisco de Vitoria	278	1.96	5,439	707
UFV Business	132	1.48	653	1,158
Mirada21.es	73	1.23	836	722
Comunicación UFV	67	1.12	1,993	897
Mirada21TV	63	0.77	337	253
Precio UFV	62	0.00	2	4
Marta_UFV	49	1.67	243	370
Corresponsales d Paz	42	1.11	856	1,008
ti3CEIEC	38	0.54	179	413
Comenta UFV	36	1.81	139	266
Jane del Tronco	35	3.70	4,886	2,696
Javier_UFV	34	1.10	208	292
Esteban_UFV	32	0.80	81	200
Inst. Robert Schuman	29	0.74	208	391
Onda Universitaria	27	0.46	594	462
Nacho Gamma	25	2.53	225	116
Paulina Nuñez	25	1.99	1,849	1,244
Álvaro_UFV	23	1.13	69	122
Juana_UFV	22	1.32	87	214
RegnumChristi España	22	1.60	343	295

En la figura 185 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 185. Nube de palabras UFV. Fuente: elaboración propia.

La tabla 62 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 62. Tuits más influyentes de la UFV. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @_Traveo_: RT @Unievento:Al grupo de #fisioterapia de la #ufv esperamos que tengáis u...	51	7.87	54
RT @josecuervo_es: Y el ganador del fiestón de #UniversityPartyAnimals es...@ufvmadrid ¡E...	30	12.00	37
RT @ufvmadrid: ¡Atención! Cambio de señalización de entrada al Campus desde Majadahond...	20	0.00	21
RT @ufvmadrid: ¡Ya tenemos fecha! La fiesta de #UniversityPartyAnimals será el 26 de junio, ...	19	0.00	22
RT @C_dPaz: Una alumna de periodismo de la @ufvmadrid acompaña a la @policia en la reci...	16	1.50	20
RT @josecuervo_es: ¡ATENCIÓN! El fiestón de #UniversityPartyAnimals cambia de fecha al 2...	14	0.00	16
RT @DebatesUFV: Muchas gracias a los que habéis hecho posible esto: jefes, equipos, juece...	13	0.00	13
RT @Fedgolfmadrid: Homenaje de la UFV a nuestros golfistas http://t.co/yE4jsHWOS3 @Seta_...	13	0.00	13
RT @BecasEuropa: Ya van llegando todos los candidatos de Madrid para X edición en @ufvm...	12	0.00	13
RT @ufvmadrid: La Marcha Imperial suena en @El_Hormiguero con #RickyMartinEH y la disq...	12	0.00	12
RT @NoTeCortes40: Hoy tenemos a alumnos de la @ufvmadrid y @feijoose nos cuenta su ide...	11	5.25	11
RT @ufvmadrid: El día de #Selectividad es muy posible que los nervios te traicionen. Sigue es...	10	5.00	10
RT @BecasEuropa: Haciendo nuevas amistades gracias a #BecasEuropa en @ufvmadrid htt...	10	4.00	11
RT @ufvmadrid: ¡Atención! Ya podéis descargaros vuestra entrada para la fiesta exclusiva d...	10	0.00	10
RT @ufvmadrid: ¡El campus de la Universidad Francisco de Vitoria desde el cielo! http://t.co/p...	10	0.00	10
RT @ufvmadrid: ¡Te esperamos este sábado para celebrar contigo la Jornada de Puertas Abi...	10	0.00	10
RT @Naukas_com: Nuevo post en La Lista de la vergüenza: La Universidad Francisco de Vito...	10	-5.40	13
RT @SusanaNavaln: Atención #medios! Nuevos #periodistas sobrados de #talento, #potencia...	9	3.75	9
RT @colegio_orvalle: Alumnas de 4to ESO participan en Torneo Intermunicipal de Debate #TI...	9	0.00	10
RT @ufvmadrid: ¡Buenos días! Abierto el plazo hasta 30 sept para la solicitud de Becas y Ayu...	8	9.00	8

La figura 186 muestra los tuits que hablan sobre la Universidad Francisco de Vitoria. El principal foco se encuentra en Madrid. Existen tres ubicaciones más con un número muy pequeño de tuits. En Valencia comentan sobre la universidad, en Ciudad Real sobre las Becas Europa y en Sevilla sobre el nuevo Grado en Gastronomía de la universidad.

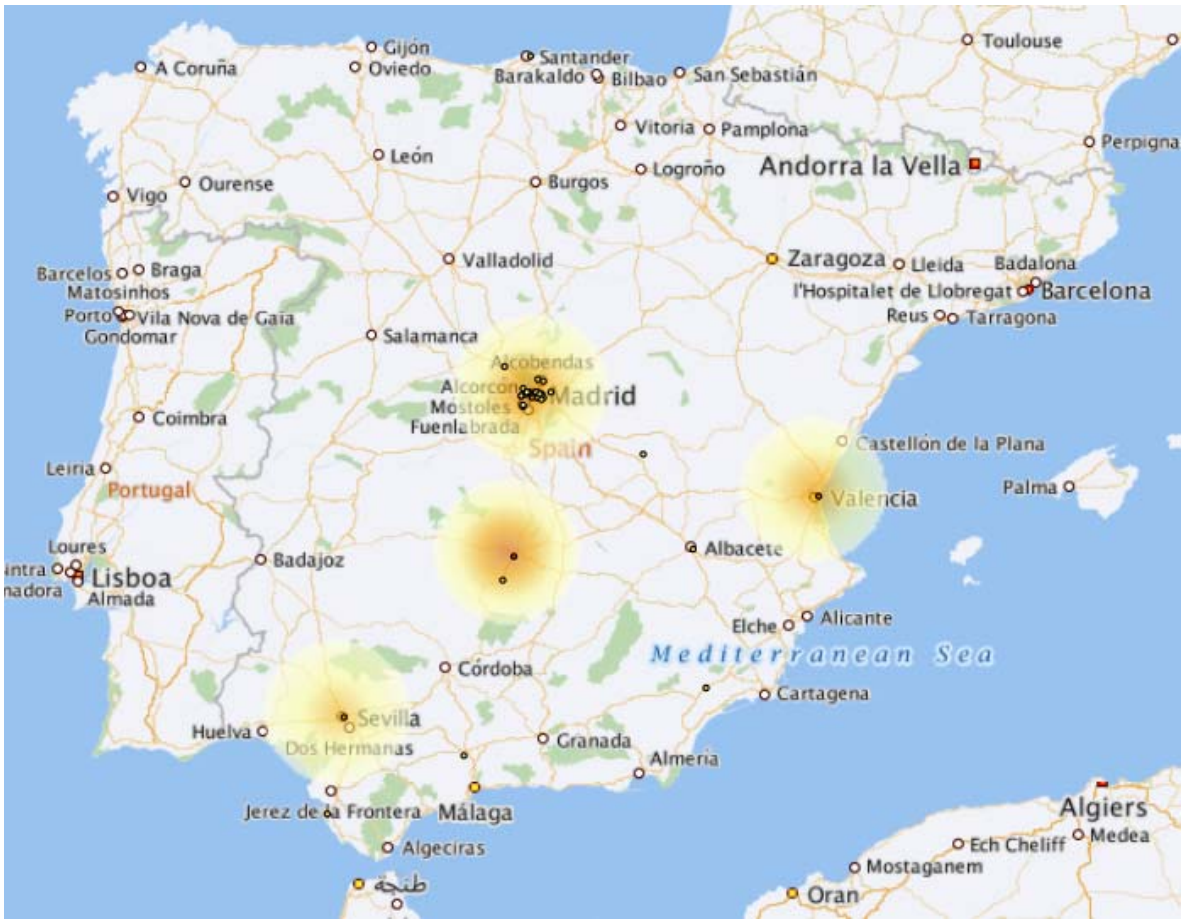


Figura 186. Tuits UFV en España. Fuente: elaboración propia.

En Madrid (figura 187), el principal foco está en Pozuelo donde se ubica la universidad. Además se aparecen otros tuits dispersos hablando sobre las pruebas de admisión de la UFV.

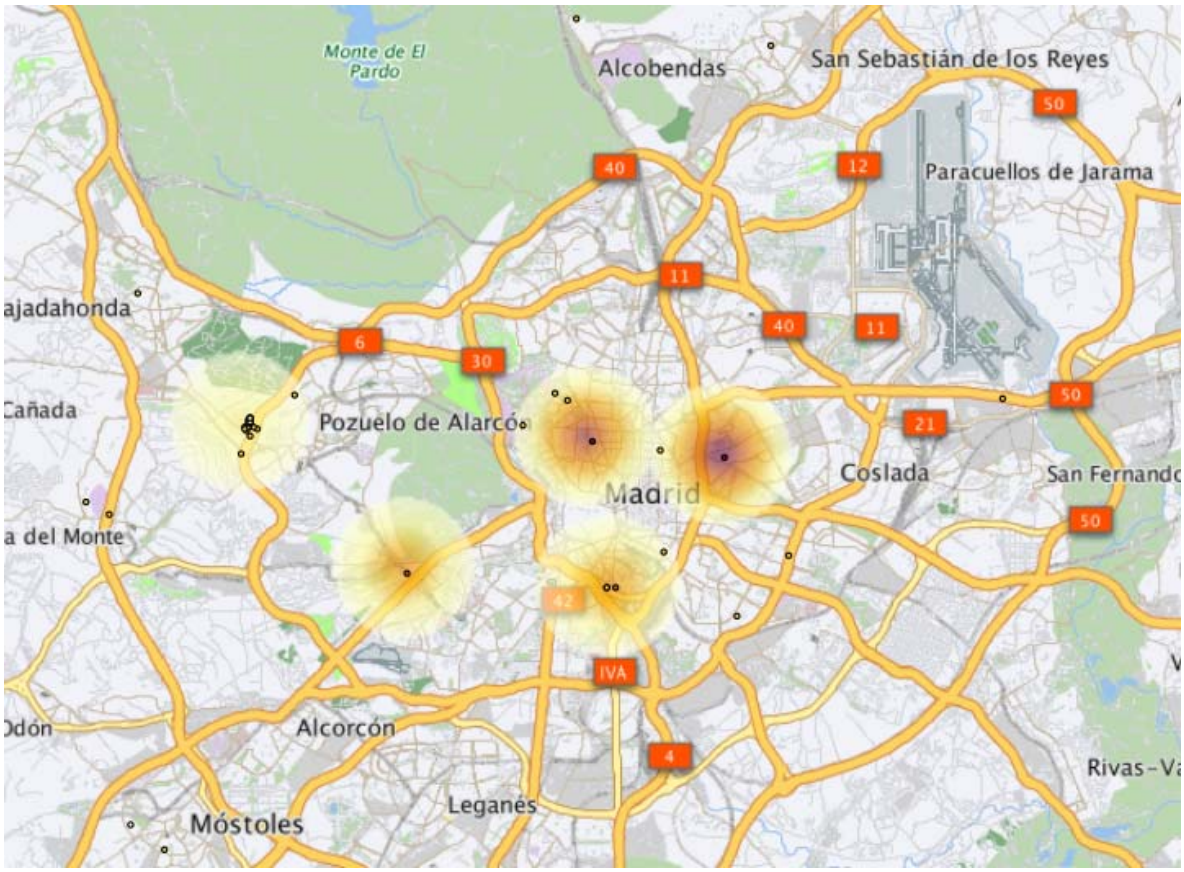


Figura 187. Tuits UFV en Madrid. Fuente: elaboración propia.

6.6.2.13 Universidad Internacional Menéndez Pelayo

Se han obtenido 3.787 tuits de los cuales el 28,2% eran positivos y el 4,6% negativos. Mediante la tabla 63 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 63. Usuarios más influyentes de la UIMP. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
UIMP de Valencia	514	1.12	7,551	3,367
UIMP Sevilla	121	1.32	556	279
UIMP	70	1.13	10,352	166
Cursos de Verano	52	1.35	1,246	42
César Nombela Cano	50	1.06	667	100
Elena.V Segura Trigo	30	0.23	927	427
Pablo Mugüerza	24	0.46	1,581	2,002
Guía educativa	23	0.00	63,798	67,843
Belén Elisa Díaz	23	0.35	909	2,001
Encarna Aguilar	20	0.73	17	14
Máster Eco. Creativa	19	0.00	731	1,437
Fundación Incorpora	19	0.47	70	301
david_busta	17	0.82	888	832
APIE_es	15	1.70	731	297
Daniel Pérez Fdez	15	0.20	1,122	1,348
Europa Press	14	0.43	3,735	162
Jon Caballero	13	0.98	514	644
Juan Carlos Garcia	12	6.33	1,120	904
Pedro Ybarra	12	0.00	813	354
Crece Empleo	11	3.00	79	117

En la figura 188 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

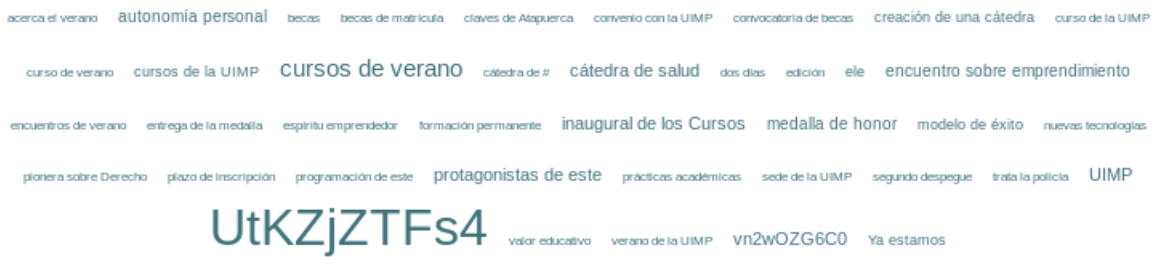


Figura 188. Nube de palabras UIMP. Fuente: elaboración propia.

La tabla 64 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 64. Tuits más influyentes de la UIMP. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @UIMP: 1932-2014. La primera Universidad de Verano, la UIMP. #CursosdeVerano ¿Te los ...	169	0.00	261
RT @educaciongob: Convocadas plazas #cursosverano #formaciónpermanente #profesorad...	51	3.00	60
RT @CursosUIMP: 1932-2014. La primera Universidad de Verano...La UIMP. #CursosdeVerano...	32	0.00	32
RT @UIMP: Nuestros Cursos de Verano. @CursosUIMP http://t.co/6wwzzQApEO	20	0.00	27
RT @CursosUIMP: ¡Ya estamos aquí! ¡Os esperamos! #Santander #Cantabria #UIMP #Cursos...	18	0.00	20
RT @CursosUIMP: Abierta la convocatoria de #becas para los Cursos Avanzados de Verano ...	17	3.00	17
RT @BarrosoEU: En #Santander hoy. Inauguración curso de verano @UIMP. Emoción al recibi...	15	4.50	17
RT @educaciongob: Quedan dos días de plazo #cursosverano #formaciónpermanente #profe...	14	3.00	14
RT @cncano: Dentro de diez días presentamos en Santander la programación de este verano ...	14	0.00	14
RT @CJC_Cantabria: Así trata la policía a quien protesta en la entrega de la medalla a Durao ...	12	-4.00	13
La UIMP convoca el XXVIII Premio Internacional Menéndez Pelayo: El Boletín Oficial del Estad...	11	0.00	0
RT @CursosUIMP: ¿Sabías que la primera universidad en implantar los Cursos de Verano fue ...	11	0.00	11
RT @UIMP: Ya somos más de 10.000. Muchas gracias a todos. Os esperamos este verano en ...	11	0.00	26
#Becas La Universidad Internacional Menéndez Pelayo convoca becas para todos sus centro...	10	3.00	1
RT @LuisngeldeBenit: XII Curso de Análisis Musical, Simposio "Música, significado y comunic...	10	0.00	10
RT @uimpsevilla: Programa del encuentro sobre emprendimiento y espíritu emprendedor en l...	9	4.50	3
RT @UIMP: Prácticas para el Gabinete de prensa en los @CursosUIMP. ¿Estás interesado? #...	9	3.00	9
RT @AndreinaSeijas: Interesado en #ciudades? @BID_Ciudades y @UIMP te invitan al curso ...	9	0.00	9
RT @CursosUIMP: La UIMP presenta en Santander sus Cursos Avanzados de Verano 2014 co...	9	0.00	9
RT @DidacTEK: #Formación Profesores de idiomas en @UIMP "LA ENSEÑANZA DE LENGU...	9	0.00	11

La figura 189 muestra los tuits a nivel nacional. La UIMP cuenta con numerosos centros repartidos por la geografía, entre los cuales se encuentran Santander y Madrid. Los principales focos se localizan en Badajoz y Almería desde donde los usuarios han comentado que han sido admitidos.

También aparecen algunos tuits repartidos de manera heterogénea por todo el territorio. No se encuentra ninguna relación.

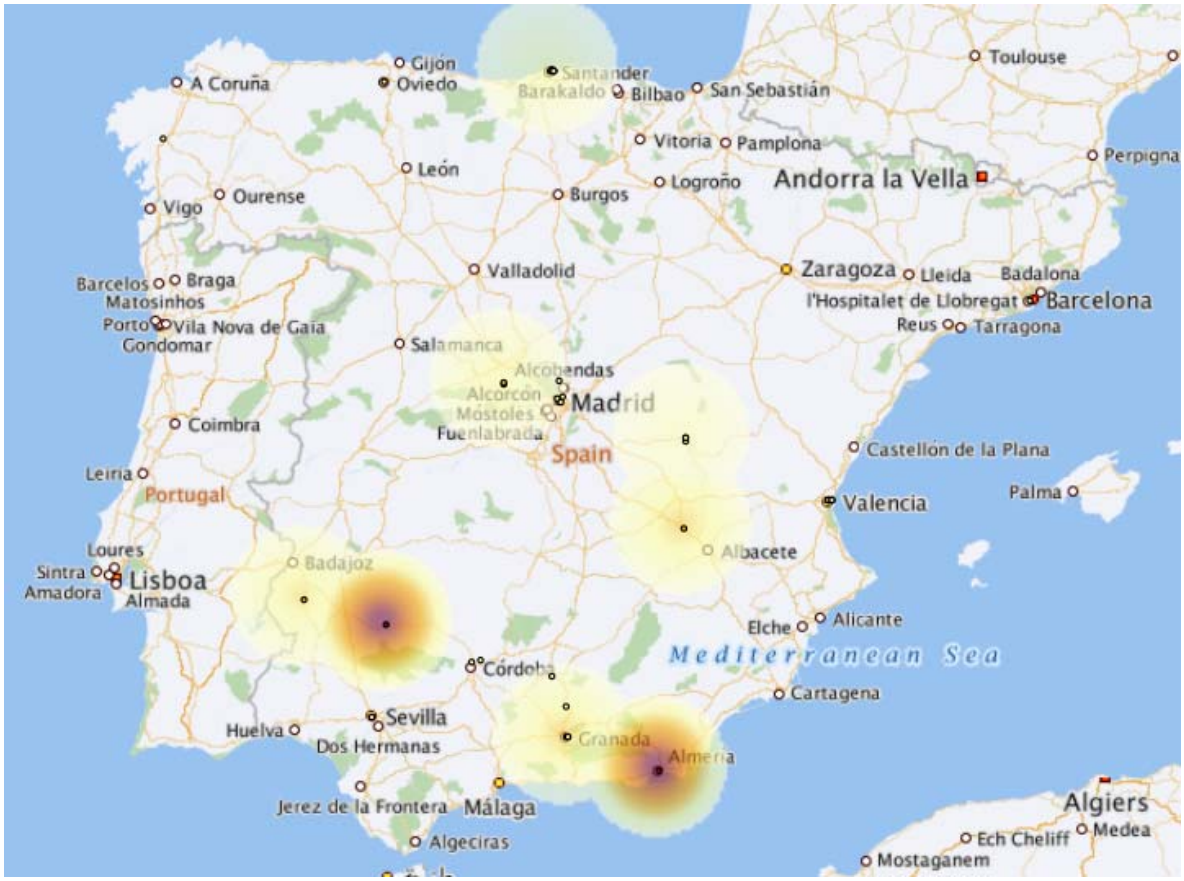


Figura 189. Tuits UIMP en España. Fuente: elaboración propia.

En Madrid (figura 190) aparecen los tuits concentrados en la zona noroeste de Madrid hablando sobre los cursos que imparten.

Figura 190. Tuits UIMP en Madrid Fuente: elaboración propia.

6.6.2.14 Universidad Alfonso X el Sabio

Se han obtenido 2.023 tuits de los cuales el 19,3% eran positivos y el 4,5% negativos. Mediante la 65 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 65. Usuarios más influyentes de la UAX. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
UAX	120	1.34	3,562	839
CAPEA UAX	45	0.00	19	211
Iñigo De Juana	35	0.26	15,325	16,786
Indignados UAX	31	0.40	397	726
Hotel don Angel	28	1.34	152	317
Angel Sampedro	23	1.85	543	712
ᵈᵘᵐᵉ ✖α†γ™ ☜=㋡	20	0.00	4,174	600
Noticias de Madrid	19	0.76	455	44
Noticias Noroeste	14	0.00	28,559	23,871
Iván Miñano	12	0.63	211	229
ingenieríacarreteras	11	2.06	1,326	100
Susana Álvarez	10	0.74	368	488
ani	10	0.09	900	1,179
Anáhuac Xalapa	9	0.00	1,540	67
Daniela Torea Olego	9	-0.50	122	213
Francisco BarbosaDDS	8	0.84	2,159	2,363
Javier Darkholme	8	0.00	142	323
Madridiario.es	7	2.43	37,662	3,291
CéSaR MaRs	7	-0.81	323	318
I ♥ StreetDanceWorld	7	0.00	1,414	789

En la figura 191 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 191. Nube de palabras. Fuente: elaboración propia.

La tabla 66 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 66. Tuits más influyentes de la UAX. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
Para dar clase en la UAX solo te piden la ESO. Es lo imprescindible para buscar en google y le...	33	0.00	0
#MercurioOo Una cátedra de la Real Academia de Gastronomía en la Universidad Alfonso X el...	27	0.00	1
RT @UAXenfurecida: EN LA UAX HEMOS VIVIDO MUY DE CERCA EL GRAN PREMIO DE ESP...	25	0.00	30
RT @CrisWlls: Última semana en la UAX. http://t.co/cHJwQ5IPnv	14	0.00	15
RT @velezmeca: Final de curso de farmacología Medicina de la UAX. Buen trabajo y esfuerzo....	13	0.00	15
RT @universidad_uax: Nace la primera cría en cautividad de buitre leonado en la Granja Expe...	12	0.00	13
RT @VictorParedes78: Creación de un TRX con materiales de escalada por @PeT3R_XCM (U...	11	3.00	11
RT @universidad_uax: Todos los estudiantes de la UAX podrán beneficiarse de un 50%de des...	11	0.00	12
RT @universidad_uax: La UAX adopta 3 yeguas que se incorporarán al Hospital Clínico Veteri...	10	0.00	12
RT @universidad_uax: Ya queda poco para el Acto de Graduación 2014. ¡Entérate de todo aqu...	10	0.00	13
RT @jardelrivero: Mis Alumnos q Terminaron con Excelencia el Semestre @UJAT @Joseman...	8	0.00	8
RT @JJ_Arevalillo: Cartel del Experto en Tradumática, Localización y Traducción Audiovisual,...	8	0.00	8
RT @pgmoinante11: Grandees mis uax! Campeones del torneo de futbol 7 siendo de primer a...	8	0.00	8
RT @Madridiario: El #premiosmadrid educativo es para @universidad_uax, por su modelo inn...	7	4.50	7
RT @universidad_uax: La UAX como uno de los mejores centros para cursar Odontología y V...	7	3.76	9
RT @universidad_uax: El proyecto "Sport4U" realiza con éxito el torneo de fútbol sala para pe...	7	1.44	7
RT @NoticiaNoroeste: RT.- Alumnos #UAX organizan "MomentoInclugol", un torneo de #FUTB...	7	0.00	5
RT @paulatorres50: "Jesús Núñez Velázquez, dueño de U. Alfonso X El Sabio (UAX) y el coleg...	7	0.00	8
RT @paulatorres50: Los rallys de la UAX. http://t.co/ev6YjikbeG	7	0.00	8
RT @universidad_uax: Ya puedes ver el Spot Publicitario de la UAX ¡¡Gracias a todos por parti...	7	0.00	7

La figura 192 muestra los tuits a nivel nacional. El principal foco de concentración se sitúa en Madrid, donde se encuentra la universidad, y otros dos en Oviedo y Cádiz. En estas dos últimas ubicaciones se habla sobre un anuncio de televisión de la UAX.

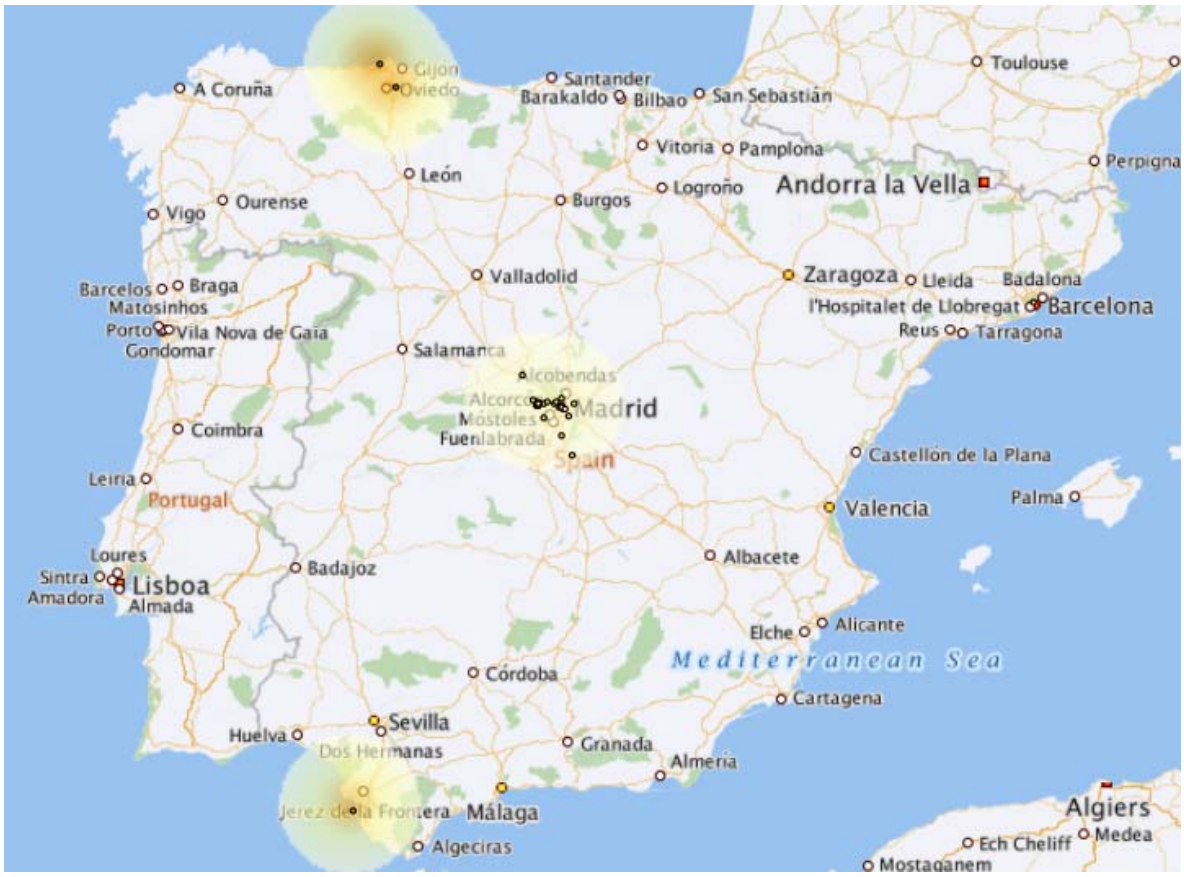


Figura 192. Tuits UAX en España. Fuente: elaboración propia.

En Madrid (figura 193) aparecen claramente dos ubicaciones. La primera en Villanueva de la Cañada, donde se encuentra la universidad. La segunda en el área metropolitana de Madrid, desde donde se habla de la universidad en general.

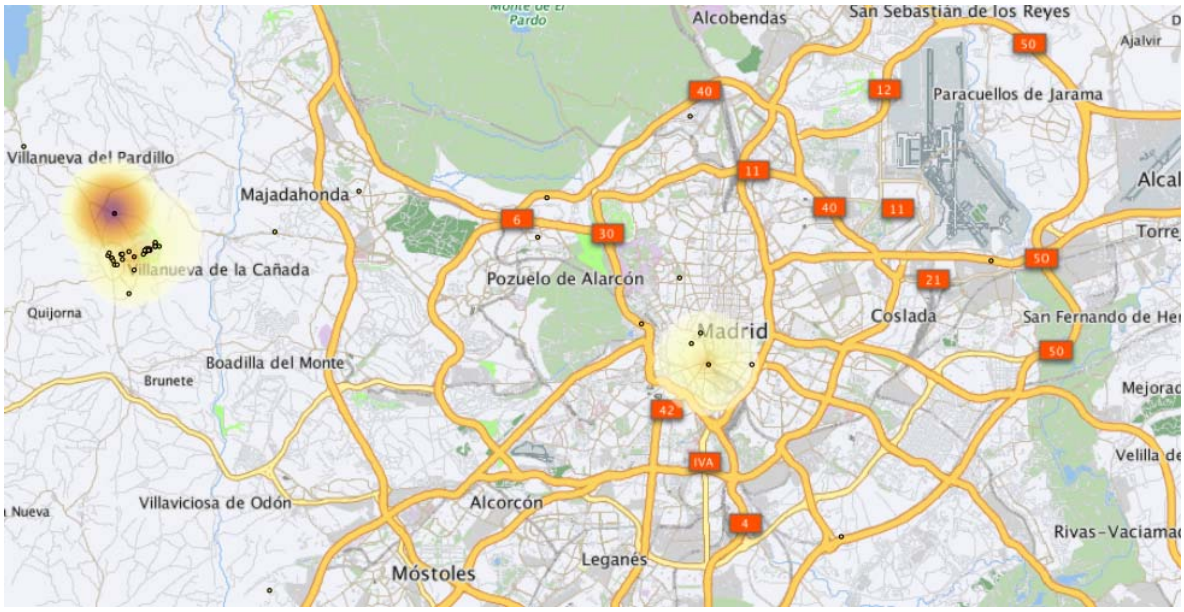


Figura 193. Tuits UAX en Madrid. Fuente: elaboración propia.

6.6.2.15 Universidad a Distancia de Madrid

Se han obtenido 1.900 tuits de los cuales el 29,4% eran positivos y el 3,5% negativos. Mediante la tabla 67 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 67. Usuarios más influyentes de la UDIMA. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
UDIMA	133	1.60	5,155	2,179
Silvia Prieto	69	1.60	2,500	2,514
Arturo de las Heras	56	1.95	3,699	3,261
CEF.-	43	1.88	5,959	1,097
Elena Curbelo	41	0.66	35	117
Liceus Humanidades	34	1.04	2,464	2,075
Luis Miguel Belda	29	0.82	659	554
analandeta	29	1.42	670	232
Antonio Rguez Ruibal	24	2.68	1,975	1,689
Alberto Joven	22	1.44	1,491	1,229
OH! MY PLAN	21	3.11	2,934	2,606
Abel González	19	0.44	1,104	1,533
Lucas Castro	19	1.94	378	755
María Abajo	18	1.88	475	658
Formazion	16	1.29	667	1,795
Liceus-Ensino de ELE	14	1.32	154	178
enClave-ELE	14	1.68	1,501	723
Manuel Herrador	13	1.29	117	239
Pilar Moreno Collado	13	1.23	413	554
Inst. Idiomas UDIMA	13	1.30	195	177

En la figura 194 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 194. Nube de palabras UDIMA. Fuente: elaboración propia.

La tabla 68 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 68. Tuits más influyentes de la UDIMA. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @UDIMA: Las I Jornadas gratuitas de Enseñanza de Español en la #EraDigital @UDIMA co...	13	0.00	13
La universidad a distancia UDIMA impartirá el primer Doctorado online el próximo curso http:/...	12	0.00	1
RT @pedro_borrego: ¿Cómo aprendemos? @estudiosCEF @UDIMA @IdiomasUDIMA @artur...	12	0.00	16
RT @arturocef: Con el equipo CEF.-UDIMA en la final del concurso de talentos de marketing @...	11	7.00	12
RT @arturocef: Quiero compartir una buena noticia, ya tenemos los dos primeros Másteres d...	11	4.00	11
RT @arturocef: Convocadas #Becas Internacionales 2014 @MESCyT_1 http://t.co/apfstIT4sh ...	11	0.00	11
RT @arturocef: Ya está aquí el trofeo de #TalentosMarketingPeugeot Enhorabuena! @UDIMA ...	10	6.00	10
RT @arturocef: Ganamos!!!! Edificio @udima premios ASPRIMA 2014 @simaexpo http://t.co/Z...	10	5.00	10
RT @PuertoLocal092: Seguimos realizando encuestas a comerciantes gracias a la colaboraci...	10	3.00	10
RT @arturocef: ¿Conoces el edificio @UDIMA? Miguel Ángel López, mejor atleta español 2013...	10	0.00	10
RT @inGameEXP: Gracias a @estudiosCEF hablamos con el profesor Francisco Vacas de la ...	10	0.00	11
RT @UDIMA: ¿Por qué tus hijos prefieren #YouTube a la televisión? con Francisco Vacas prof...	10	0.00	12
RT @UDIMA: Estaremos en las Jornadas @X1RedMasSegura, para el uso seguro de Internet ...	9	4.00	9
RT @AbelGlezG: Muy interesante el repaso por los riesgos de internet para menores de @_A...	9	0.63	10
RT @rodriguezruibal: Voy a dar el máximo en la asignatura de Medios y Redes Sociales de @...	9	0.00	12
RT @UDIMA: "Adolescentes y RRSS: Claves para no perderte como padre" María García Qui...	9	0.00	9
RT @AbelGlezG: Comenzamos el taller para educadores en @UDIMA #x1redmassegura http:/...	8	0.00	8
RT @jobandtalent_es: "La prioridad del @estudiosCEF y la @UDIMA es continuar por la send...	8	0.00	8
RT @UDIMA: La @UDIMA sigue creando empleo: #convocatoria de plazas curso académico 1...	8	0.00	9
RT @X1RedMasSegura: @HackingMom aporta el punto de vista de los padres en el taller de e...	8	0.00	8

La figura 195 muestra la ubicación de los tuits a nivel nacional. Claramente se diferencian dos ubicaciones, la primera en Madrid, donde se encuentra la universidad. La segunda en Palma de Mallorca, donde una persona comenta sobre el trabajo de fin de máster que realizará en la UDIMA.

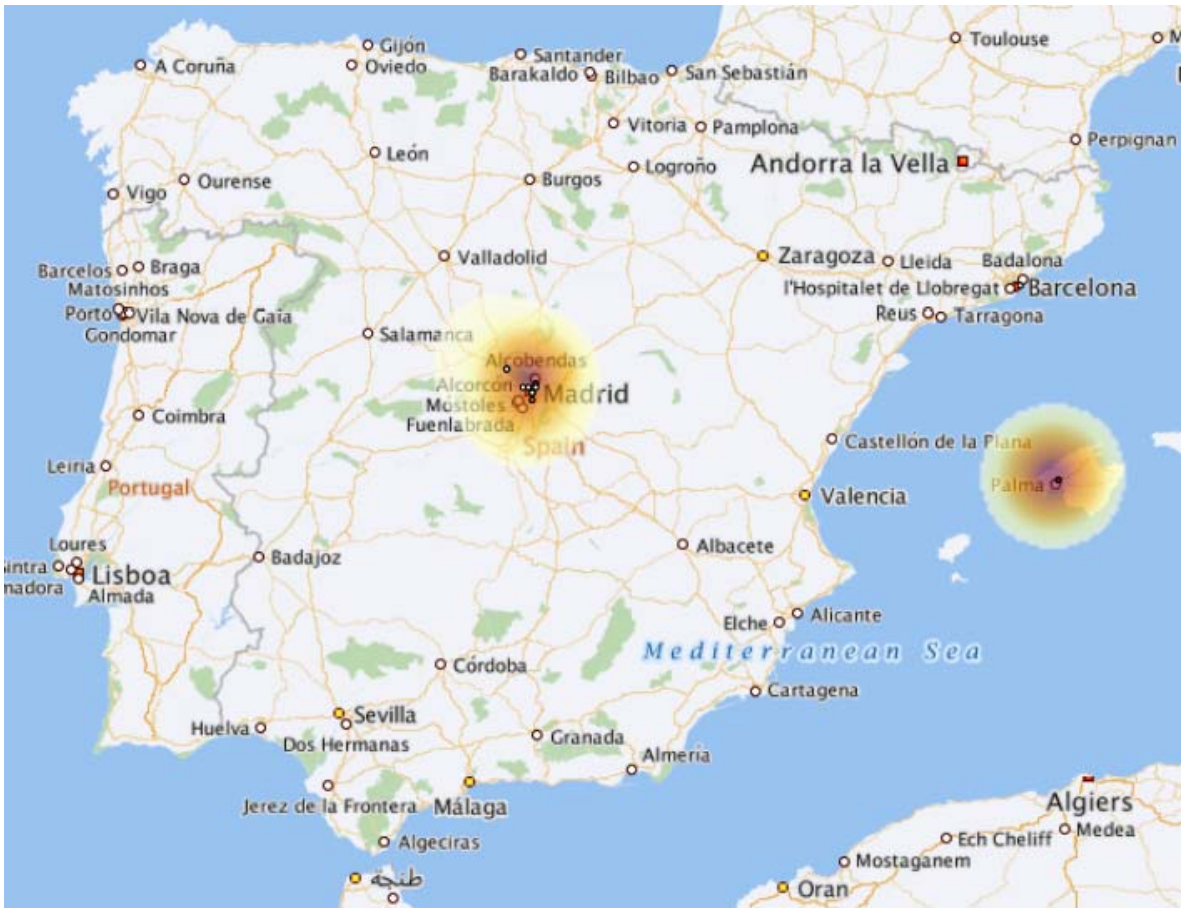


Figura 195. Tuits UDIMA en España. Fuente: elaboración propia.

En Madrid (figura 196) se ubican tuits repartidos por la zona. Las zonas más calientes del mapa indican una mayor concentración de tuits (en este caso, 2 tuits), pero no se encuentra una relación entre lo que se dice y donde se ubica. Además la UDIMA no tiene centros asociados.

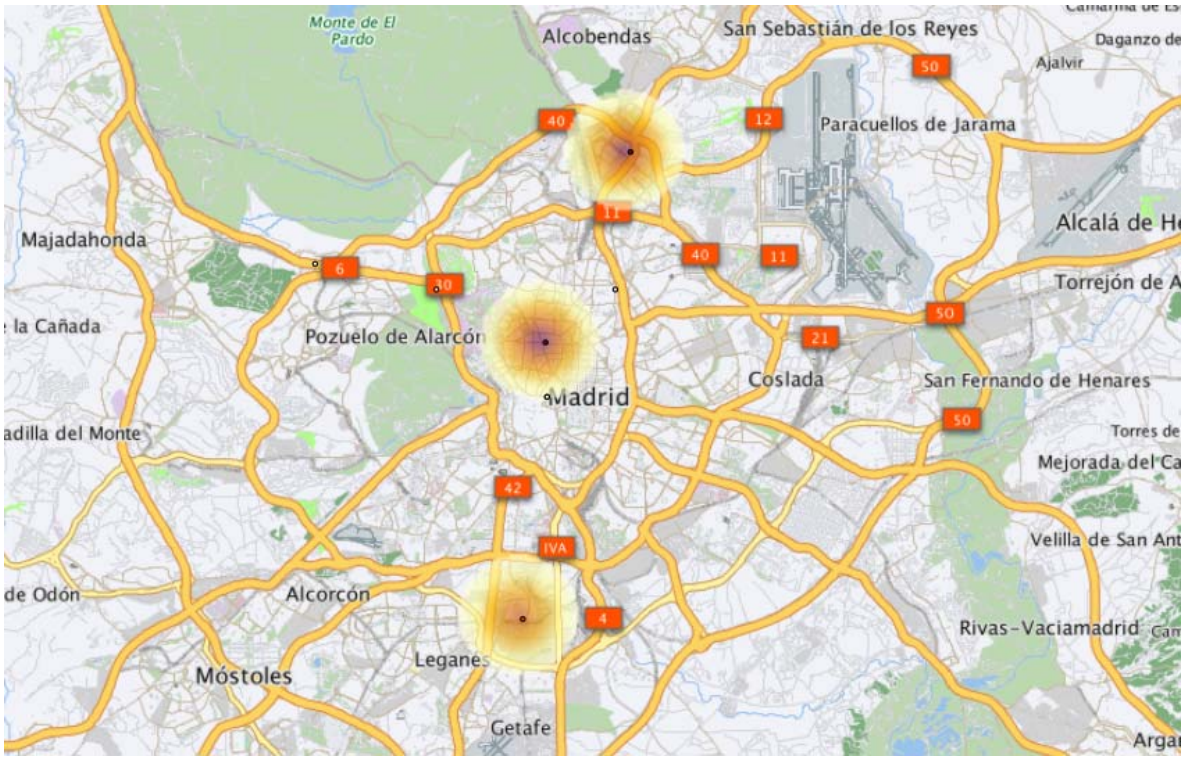


Figura 196. Tuits UDIMA en Madrid. Fuente: elaboración propia.

6.6.2.16 Universidad Pontificia de Comillas

Se han obtenido 1.612 tuits de los cuales el 23,6% eran positivos y el 9,2% negativos. Mediante la tabla 69 se han ordenado los usuarios en orden decreciente por el número de tuits (segunda columna). En la tercera columna (DocumentSentiment) la media obtenida del análisis de sentimiento de los tuits que cada usuario. La cuarta columna (followers_count) y la quinta columna (friends_count) son los seguidores del usuario y el número de amigos respectivamente.

Tabla 69. Usuarios más influyentes de la UPCOMILLAS. Fuente: elaboración propia.

from_user_name	DocumentSentimen...	DocumentSentimen...	followers_count (m...	friends_count (max)
Comillas O.Promoción	92	0.25	531	832
Comillas ICAI-ICADE	82	0.92	2,805	1,438
Emilio Sáenz-Francés	60	1.49	659	515
MAS Consulting	29	0.86	5,832	5,195
Amigos de Nyumbani	26	0.75	248	308
Education Trend	24	-0.71	39	139
Rodrigo Pérez Perela	23	2.41	57	252
Comillas_Rectorado	23	2.50	248	59
Federica Lupi	21	2.02	119	347
Rafael Ortega	20	0.50	60	116
Carmen Quiñones	18	2.92	168	477
Comillas_FacultadCHS	18	1.00	455	459
Fabiola García	18	1.63	136	302
Carlos Ballesteros	17	0.94	778	760
EducaManagement	14	-0.02	55	388
Vanessa Muñoz	13	1.64	206	397
Comillas_biblioteca	13	2.54	369	51
Magis Radio sj	11	0.00	392	260
LuisFranBlanco	11	1.25	419	244
maria jesus baguena	9	3.42	911	2,006

En la figura 197 se muestra una nube de palabras en función de los temas encontrados y su frecuencia.

Figura 197. Nube de palabras UPCOMILLAS. Fuente: elaboración propia.

La tabla 70 muestra los tuits de la universidad ordenados por número de apariciones. En la segunda y tercera columna se muestra el sentimiento del tuit y el número de retuits.

Tabla 70. Tuits más influyentes de la UPCOMILLAS. Fuente: elaboración propia.

text	Record Count	DocumentSentimen...	retweet_count (max)
RT @UCOMILLAS: Hoy, Día Internacional de la Enfermería, nos acordamos de los enfermeros...	25	0.00	28
RT @RLDieguezAES: Desde @Impulso_Social no paramos proximo martes 22 estaré a las 17...	16	0.00	16
RT @garcia_jordi: En un mundo tan competitivo, quien no corre vuela! Felicidades @UComilla...	14	6.00	15
RT @Fundacion_ONCE: La primera promoción de personas con #discapacidad intelectual se ...	14	-0.06	14
RT @Ecojesuit: Universidad Jesuitas: Pensamiento Social Cristiano en el s. XXI. http://t.co/k0...	11	0.00	11
RT @comunicarLe: Todas las grandes cosas se gestan en el silencio. #Ejercicios @UCOMILL...	10	3.00	11
RT @BraulioPareja: "Educar en competencia emprendedora sirve para saber enfrentarse a lo...	9	0.67	9
RT @Comillas_marca: Ven a conocer nuestros grados en Criminologia,Psicologia, TrabajoSoc...	9	0.00	9
RT @Entreculturas: Hoy a las 11h en la @UCOMILLAS el Servicio Jesuita a Migrantes España...	8	4.00	8
RT @CarlaSancor: Es verla, y sentir alegría y nostalgia al mismo tiempo. Graduados de Enfer...	8	1.44	9
RT @cristinasancho: Presentación del Estudio #comunicacioninterna en Empresas Cotizadas...	8	0.00	8
RT @ignacioaguado: En @UCOMILLAS acompañando a @luissalvador en su presentación so...	8	0.00	9
RT @UCOMILLAS: Si quieres estudiar ingeniería en @Comillas_ICAI, ven a las sesiones de or...	8	0.00	8
RT @ESFrances: Don Fabrizio Salina desea a los alumnos de E6 en @UCOMILLAS mucho éxit...	7	12.25	10
RT @Accem_ong: Hoy participamos en la jornada "Visibilizando lo oculto" sobre #trata de per...	7	0.00	7
RT @comunicarLe: Hay palabras que nunca se llevará el viento. #Ejercicios Día 2 @UCOMILL...	7	0.00	8
RT @RAISFundacion: .@UCOMILLAS y @RAISFundacion lanzan una residencia para estudian...	7	0.00	7
RT @RLDieguezAES: Desde @Impulso_Social no paramos mañana martes estaré a las 17:30...	7	0.00	8
RT @RLDieguezAES: Gracias @periodistadigit por cubrir el debate de ayer por @JMP_Oficial ...	7	0.00	7
RT @RLDieguezAES: No lo olvides hoy a las 17:30 estaré en @UCOMILLAS con @vox_es @...	7	0.00	7

La figura 198 muestra los tuits a nivel nacional. Se puede observar que solamente aparecen en Madrid.

Figura 198. Tuits UPCOMILLAS en España. Fuente: elaboración propia.

En Madrid (figura 199) aparecen tres concentraciones de tuits claramente diferenciadas, al norte el campus de Cantoblanco, en el centro ICADE e ICAI (zona Argüelles) y al sur el campus de Ciempozuelos.

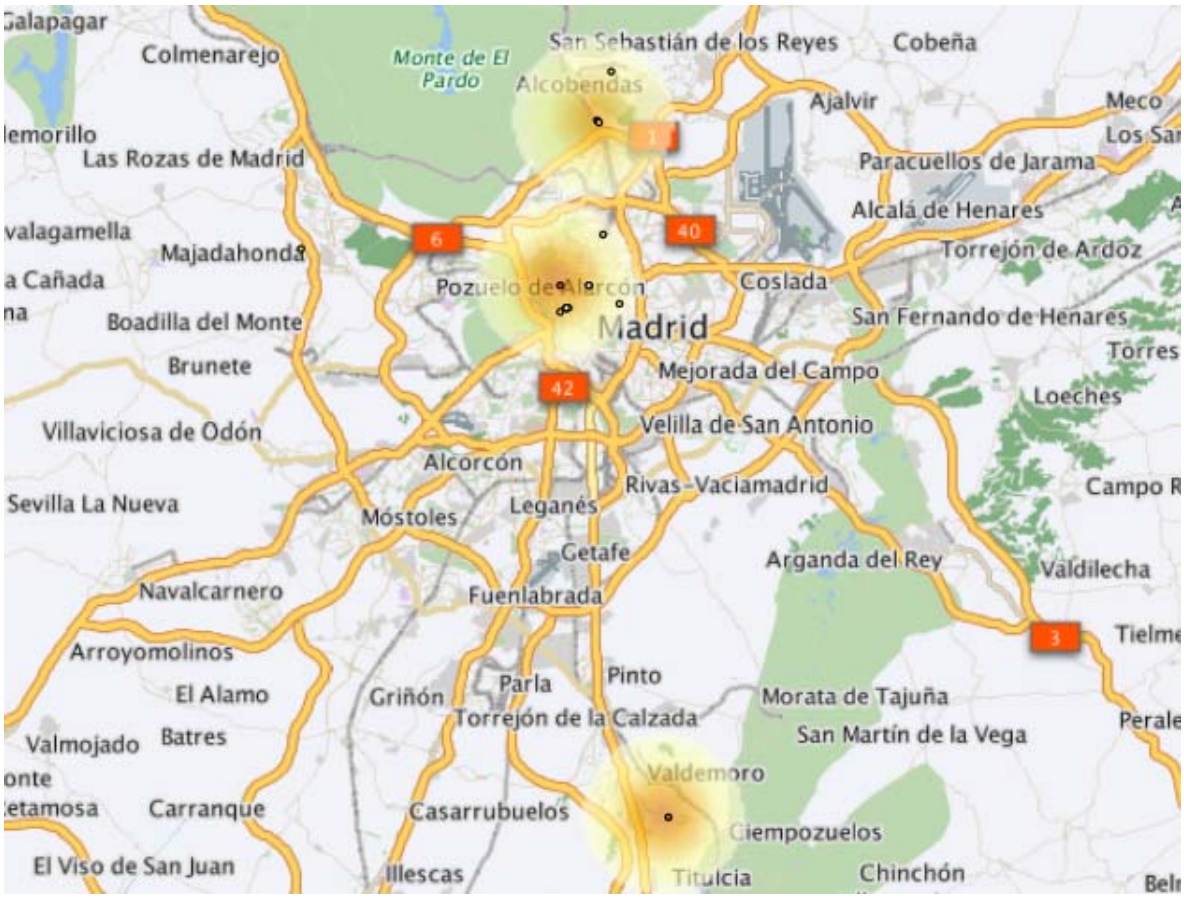


Figura 199. Tuits UPCOMILLAS en Madrid. Fuente: elaboración propia.

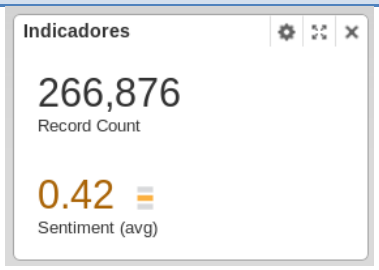
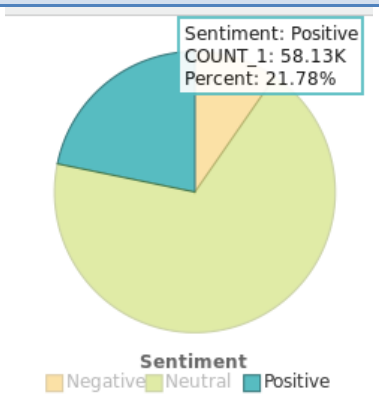
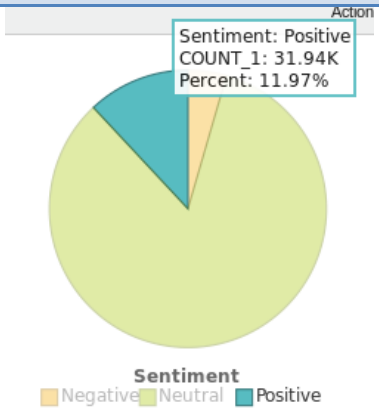
6.6.2.17 *Resultado análisis de sentimiento*

Según la metodología descrita en el apartado 6.4.3.2, se han clasificado manualmente 3.254 palabras utilizando los siguientes criterios:

- Muy positiva: +0,6
- Positiva: +0,3
- Negativa: -0,3
- Muy negativa: -0,6

Los resultados son los mostrados en la figura 200. Con un conjunto de datos de 266.876 tuits la valoración general con el diccionario original es de 0,42 frente a 0,64 con el modificado.

Con el diccionario original el número de tuits clasificados positivamente son casi 32 mil y negativos 12 mil. Con las mejoras realizadas el número de tuits calificados positivamente son más de 58 mil y negativos cerca de 10 mil.

Diccionario mejorado	**Diccionario original**
Indicadores: número de registros y valoración del sentimiento general de los tuits	
	
Sentimiento positivo	
	

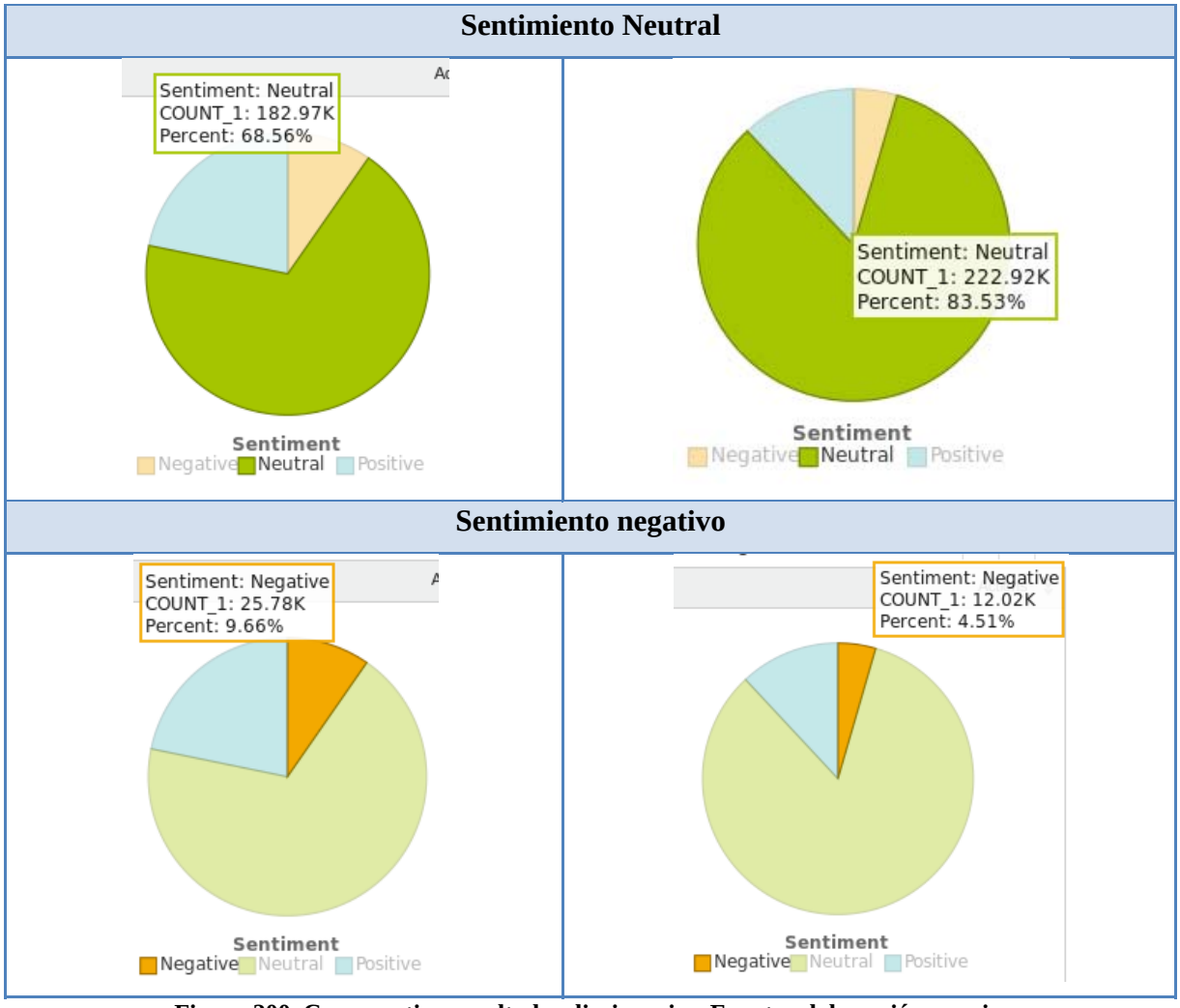


Figura 200. Comparativa resultados diccionarios. Fuente: elaboración propia.

La mejora realizada en cuanto a la valoración de tuits también mejora la clasificación de entidades de manera significativa. En la figura 201 se puede observar que con los cambios realizados en el diccionario respecto al original (figura 202), se clasifican en torno al doble de tuits por tema detectado.

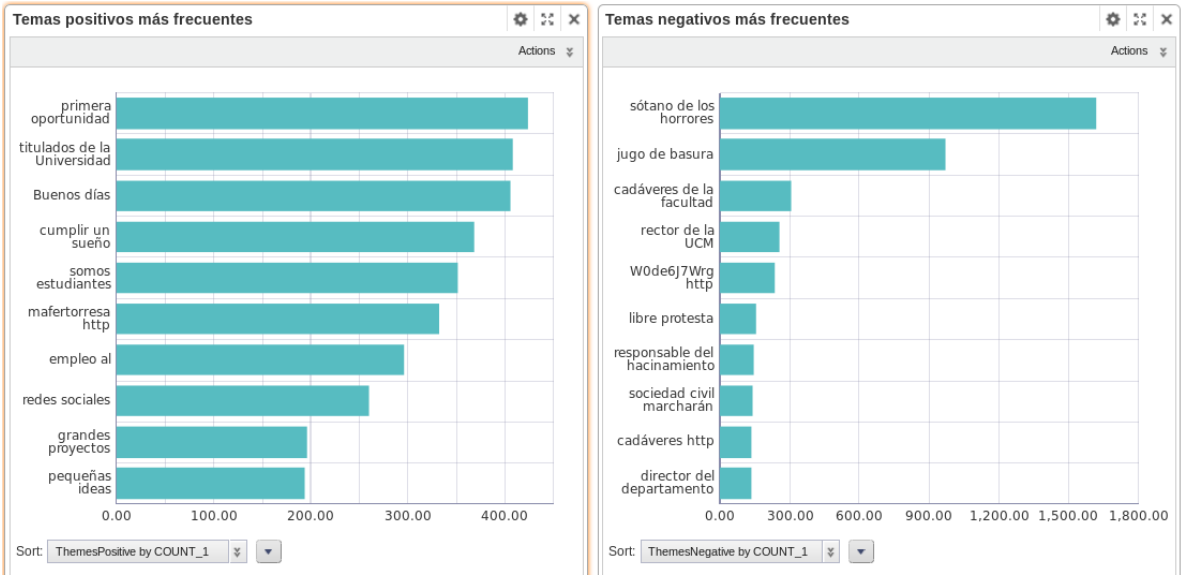


Figura 201: Temas positivos y negativos con las mejoras realizadas. Fuente: elaboración propia.

Figura 202. Temas positivos y negativos con el diccionario original. Fuente: elaboración propia.

6.6.2.18 *Resultado de los datos de las universidades con datos normalizados*

Para realizar una comparativa de las universidades, se va a normalizar los resultados en función del número de alumnos y del tiempo. El proceso de adquisición comenzó el 26 de marzo y terminó el 21 de junio, es decir, 95 días.

Se han considerado los siguientes datos sobre las universidades según el último informe accesible del Ministerio de Educación y cultura, "Avance de la Estadística de estudiantes universitarios del curso 2012-13" (59).

Normalizando los resultados obtenidos a lo largo del proyecto se obtiene la tabla 71.

Tabla 71. Tuits por alumno y mes. Fuente: elaboración con datos propios y (59).

Nombre de la universidad	Número de tuits	% tuits positivos	% tuits negativos	Número de alumnos	Tuits por alumno/ mes	Tuits positivos por alumno/mes	Tuits negativos por alumno/mes
UAM	33.453	18,6	13,1	27.181	0,388	0,060	0,041
UCM	23.966	18,7	17,1	76.497	0,098	0,060	0,054
UNED	23.390	20,5	6,9	179.741	0,041	0,066	0,022
UAH	15.580	24,3	6,5	17.922	0,275	0,076	0,022
UCJC	14.685	25,4	4,3	7.884	0,587	0,079	0,013
CEU	13.269	16	4,9	11.623	0,360	0,051	0,016
UPM	12.745	22,3	12,9	36.619	0,111	0,069	0,041
URJC	8.829	25,6	5,6	34.042	0,082	0,082	0,019
UC3M	7.076	19,2	7,7	16.675	0,133	0,060	0,025
UEM	4.925	26,5	5,4	13.251	0,117	0,085	0,016
NEBRIJA	4.280	31,9	2,3	3.340	0,404	0,101	0,006
UFV	4.200	30,1	3,6	4.342	0,306	0,095	0,013
UIMP	3.787	28,2	4,6	927	1,292	0,088	0,016
UAX	2.023	19,3	4,5	8.264	0,076	0,060	0,016
UDIMA	1.900	29,4	3,5	5.750	0,104	0,092	0,013
UPCOMILLAS	1.612	23,6	9,2	7.707	0,066	0,076	0,028

En la figura 203 se muestra la relación entre el número de tuits por alumno (eje x) frente al número de tuits positivos por alumno. Se ha eliminado de la gráfica la UIMP ya que distorsiona el resto de los resultados. La Universidad Internacional Menéndez Pelayo se encontraría arriba a la derecha en la gráfica, con 4 tuits por alumno y 1 tuit positivo por alumno.

La gráfica de la figura indica, cuanto mayor sea en el eje x, que los usuarios comentan más en proporción al número de alumnos por universidad. Cuanto mayor sea el número de tuits positivos por alumno, en el eje y, indica que los usuarios comentan más positivamente sobre la universidad.

Hay que tener en cuenta que para poder identificar los usuarios de Twitter que hablan de la universidad con alumnos de la universidad, se necesitarían agregar datos de los alumnos al sistema.

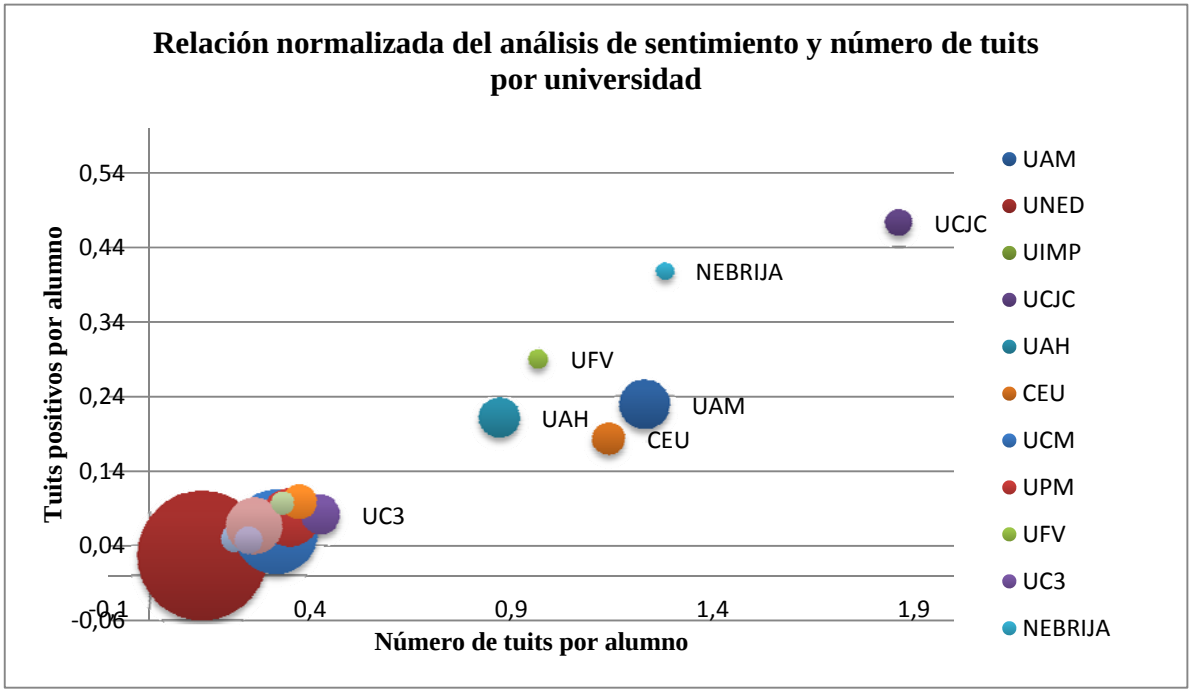


Figura 203. Número de tuits por alumno frente a tuits positivos por alumnos por universidad. Fuente: elaboración propia.

7 Conclusiones y propuestas

Cada vez es más necesario en las empresas saber qué opinan de ti y qué opinan de la competencia, ya sea para lanzar un nuevo producto al mercado o para mejorar los existentes. Este proceso a menudo es complicado, ya que con los datos que se generan hoy en día es difícil identificar los que aportan valor de los que no lo aportan.

Para realizar este análisis, se necesitan nuevas herramientas que permitan integrar datos de diferentes sitios, en diferentes formatos y que puedan ser accesibles en el momento. Estas herramientas tienen que estar orientadas a los usuarios que realmente conocen el negocio y muchas veces, no tienen perfiles técnicos.

A lo largo de este proyecto se ha desarrollado un sistema de descubrimiento de información, que permite encontrar relaciones entre los datos. Se han analizado las universidades de la Comunidad Autónoma de Madrid a través de la red social de Twitter identificando no solo los usuarios que más hablan de las universidades, sino aquellos que mejor o peor hablan de las mismas. Se ha podido identificar cuáles son las universidades mejor valoradas, quién está hablando y desde dónde se habla.

A continuación se detallan las conclusiones de cada universidad.

- **Universidad a Distancia de Madrid**
  Se obtuvieron 1.900 tuits siendo el 29,4% positivos y el 3,5% negativos. Los usuarios más influyentes son “UDIMA”, “Silvia Prieto” y “CEF.-”. El tuit positivo más comentado fue “el concurso de talentos de marketing”.
  Los usuarios que hablan de la UDIMA están dispersos por Madrid.
- **Universidad Alfonso X el Sabio**
  Se obtuvieron 2.023 tuits siendo el 19,3% positivos y el 4,5% negativos. Los usuarios más influyentes son “UAX”, “Iñigo De Juana” y “Noticias Noroeste”. El tema positivo más comentado fue “la creación de un TRX con materiales de escalada”.
  Los usuarios que hablan de la UAX están concentrados principalmente en Villanueva de la Cañada (Madrid).
- **Universidad Antonio de Nebrija**
  Se obtuvieron 4.280 tuits siendo el 31,9% positivos y el 2,3% negativos. Los usuarios más influyentes son “Nebrija BS”, “Nebrija Universidad” y “Guideo”. El tuit positivo más comentado fue “el torneo de debates de Nebrija Versus”.
  Los usuarios que hablan de la Universidad Antonio de Nebrija están concentrados en el campus de la Berzosa (Madrid) y en el área metropolitana de Madrid (zona Ciudad Universitaria y zona centro). Se observan algunos usuarios en Almería, Ceuta, Cádiz, Oviedo y Logroño.
- **Universidad Autónoma de Madrid**
  Se obtuvieron 33.453 tuits siendo el 18,6% positivos y el 13,1% negativos. Los usuarios más influyentes son “Univ. Autónoma Mad”, “Ligoteos® UAM ❤” y “AlumniUam”. El tema positivo más comentado fue “la creación de un TRX con materiales de escalada”.

Los usuarios que hablan de la UAM están concentrados principalmente en campus de Cantoblanco (Madrid) y en la zona norte de Madrid. Se observan algunos usuarios en León y Santander.

- **Universidad Camilo José Cela**
  Se obtuvieron 10.626 tuits siendo el 27% positivos y el 4% negativos. Los usuarios más influyentes son "U-tad", "UCJC" y "Carlos Fuente". El tema positivo más comentado fue sobre la educación en el centro U-TAD.
  Los usuarios que hablan de la UCJC están concentrados en centro Ferraz, Villafranca del Castillo, Las Rozas y dispersos en el área metropolitana de Madrid. Se observan algunos usuarios dispersos por el territorio nacional (Vigo, Oviedo, Zaragoza, Barcelona, Valencia, Badajoz, Córdoba y Málaga.).
- **Universidad Carlos III de Madrid**
  Se obtuvieron 7.076 tuits siendo el 19,2% positivos y el 7,7% negativos. Los usuarios más influyentes son "UC3M", "biblioteca_uc3m" y "PIC Leganés uc3m". El tema más frecuente que aparece en los comentarios es la dificultad de la universidad.
  Los usuarios que hablan de la UC3M están concentrados principalmente en el campus de Getafe y Leganés de Madrid. Además se muestra un número disperso de tuits por la Ciudad Universitaria.
- **Universidad CEU San Pablo**
  Se obtuvieron 13.269 tuits siendo el 16% positivos y el 4,9% negativos. Los usuarios más influyentes son "Universidad CEU-UCH", "Unimel Educación" y "Universidad CEU-USP".
  Los usuarios que hablan del CEU se ubican en el centro de Madrid, en la zona de Argüelles y en el campus de Montepríncipe. Además se observan numerosos tuits en Sevilla (CES Cardenal Spínola CEU), Barcelona (Abad Oliba CEU), Alicante (CEU Cardenal Herrera) y Valencia (CEU Cardenal Herrera), todos ellos pertenecientes a la fundación CEU San Pablo.
- **Universidad Complutense de Madrid**
  Se obtuvieron 23.966 tuits siendo el 18,7% positivos y el 17,1% negativos. Los usuarios más influyentes son "IEB Alumni", "IEB" y "AntonioVChanal". El tema más frecuente que aparece en los comentarios es el escándalo del "sótano de los horrores" (58).
  Los usuarios que hablan de la UCM están concentrados principalmente en la Ciudad Universitaria de Madrid. Además se observa un gran número de tuits distribuidos por toda la capital, en la zona del Retiro (donde se encuentra el IEB), en Carabanchel (Escuela Universitaria de Magisterio de Madrid), en Príncipe de Vergara (CES Cardenal Cisneros), en Alonso Martínez (CUNEF) y en el norte donde se encuentra el CES Villanueva. A nivel nacional también se registran tuits en Valencia, Badajoz, Valladolid y Valencia entre otras ubicaciones.
- **Universidad de Alcalá**
  Se obtuvieron 15.580 tuits siendo el 24,3% positivos y el 6,5% negativos. Los usuarios más influyentes son "Universidad Alcalá", "Alcalá Turismo" e "Informer UAH". El tuit más retuiteado fue sobre la graduación del Centro Universitario Gredos San Diego.

Los usuarios que hablan de la UAH se sitúan principalmente en Madrid, en la ciudad de Alcalá de Henares y en el campus de la Universidad de Alcalá. Además, a nivel nacional se han observado algunos tuits en Barcelona, Valencia, Vitoria y Badajoz.

- **Universidad Europea de Madrid**
  Se obtuvieron 4.925 tuits siendo el 26,5% positivos y el 5,4% negativos. Los usuarios más influyentes son "Universidad Europea", "Miríada X" y "Escuela I. Protoclo". El tema más frecuente fue sobre una aplicación para móvil con preguntas para el acceso a la PAU.
  Los usuarios que hablan de la UEM se ubican principalmente en el área metropolitana de Madrid, en el campus de Villaviciosa, Alcobendas (donde se encuentra el IEDE) y en la zona del Retiro (donde se encuentra PROY3CTA).
- **Universidad Francisco de Vitoria**
  Se obtuvieron 4.200 tuits siendo el 30,1% positivos y el 3,6% negativos. Los usuarios más influyentes son "Francisco de Vitoria", "Jane del Tronco" y "Comunicación UFV". El tuit más repetido fue sobre los alumnos de Fisioterapia.
  Los usuarios que hablan de la UFV se ubican principalmente en el campus de Pozuelo (Madrid) y dispersos por la capital. Además de manera puntual se encuentran algunos tuits en Valencia, Sevilla y Ciudad Real.
- **Universidad Internacional Menéndez Pelayo**
  Se obtuvieron 3.787 tuits siendo el 28,2% positivos y el 4,6% negativos. Los usuarios más influyentes son "UIMP de Valencia", "UIMP Sevilla" y "UIMP". El tuit más repetido fue sobre los cursos de verano ofrecidos por la UIMP.
  Los usuarios que hablan de la UIMP se encuentran repartidos de manera heterogénea por el territorio nacional. Aparecen algunos tuits en Badajoz, Almería, Madrid y Santander.
- **Universidad Nacional de Educación a Distancia**
  Se obtuvieron 23.390 tuits siendo el 20,5% positivos y el 6,,9% negativos. Los usuarios más influyentes son "UNED", "Tiberio Feliz" y "Tiberio Feliz Murais". El tuit más repetido fue de la UNED sobre el comienzo de los exámenes.
  Los usuarios que hablan de la UNED se encuentran totalmente repartidos de manera por el territorio nacional. Algunas de las ciudades son Barcelona, Valencia, Sevilla, Almería, Zaragoza, Mallorca y Bilbao.
  Además se observa en el área metropolitana de Madrid un importante número de tuits en la zona centro y la Ciudad Universitaria.
- **Universidad Politécnica de Madrid**
  Se obtuvieron 12.745 tuits siendo el 22,3% positivos y el 12,9% negativos. Los usuarios más influyentes son "Politécnica de Madrid", "GIA-UPM" y "Amigos Ingeniería". El tuit positivo más repetido fue publicado por la UPM, sobre un estudio del periódico El Mundo, que afirma que UPM es la más valorada de las universidades politécnicas españolas.
  Los usuarios que hablan de la UPM en Madrid se encuentran ubicados en el Campus Sur en Vallecas, en Boadilla del Monte (Campus de Montegancedo), en Ciudad

Universitaria y repartidos por toda la zona metropolitana. A nivel nacional aparecen algunos tuits en Barcelona, Vitoria, Alicante y Jaén.

- **Universidad Pontificia Comillas**
  Se obtuvieron 1.612 tuits siendo el 23,6% positivos y el 9,2% negativos. Los usuarios más influyentes son "Comillas O.Promoción", "Comillas ICAI-ICADE" y "MAS Consulting". El tuit más repetido fue sobre el Día Internacional de la Enfermería.
  Los usuarios que hablan de la Universidad de Comillas ubicados en Madrid, repartidos en tres áreas muy diferenciadas. Al norte el campus de Cantoblanco, en el centro ICADE e ICAI (zona Argüelles) y al sur el campus de Ciempozuelos.
- **Universidad Rey Juan Carlos**
  Se obtuvieron 12.888 tuits siendo el 25,6% positivos y el 5,6% negativos. Los usuarios más influyentes son "ESNE", "ESERP", y "ESIC". Uno de los tuits más repetido fue sobre el comienzo del acto de Vox en la URJC.
  Los usuarios que hablan de la URJC en Madrid proceden del ESIC (Pozuelo), del camups de Fuenlabrada, del campus de Móstoles, del campus de Vicálvaro y numeros tuits dispersos por el área metropolitana. A nivel nacional, en Bilbao y Barcelona se encuentras las principales concetraciones, aunque también se observan muchos tuits dispersos por el país.

Los principales objetivos que se han cumplido y que han sido necesarios para el desarrollo del proyecto son los siguientes:

- Para la realización de este proyecto se han utilizado diversos procesos ETL. Procesos para combinar datos estructurados y no estructurados, para la adquisición de datos, para su transformación, depuración y posterior aprovisionamiento en una base de datos analítica.
- También ha sido necesario enriquecer el texto no estructurado mediante el análisis de sentimiento y como los resultados no eran suficientes, se han conseguido mejorar.
- Se han diseñado e implementado diferentes arquitecturas para el desarrollo y las pruebas del sistema, desde servicios web de procesamiento en la nube, hasta la integración de diversos componentes empresariales para el funcionamiento del mismo.
- Además se han diseñado cuadros de mando, orientados a KPIs adecuados, sobre las universidades, sobre los usuarios y sobre dónde se encuentran. Permitiendo analizar cientos de miles de comentarios sin tener que leerlos.
- Todo ello utilizando una metodología de desarrollo en espiral, analizando los riesgos inherentes al proyecto, a lo largo de cuatro fases.

Este desarrollo puede ser utilizado para conocer mejor a los usuarios, para identificar nuevos intereses o incluso para identificar y monitorizar qué están haciendo bien y mal otras universidades.

Posibles trabajos futuros

- **Integrar datos sobre alumnos de la universidad:** relacionando los alumnos de la universidad con sus cuentas de Twitter se identifican los amigos de los alumnos, sus influencias y se conocen sus intereses. Esto centraría el estudio en lo que piensan los alumnos de la universidad sobre sus universidades en Twitter.
- **Identificar las carreras y másteres por universidad:** para realizar esta identificación habría que añadir datos sobre qué carreras y qué másteres se imparten en cada universidad. Además habría que buscar dentro del tuit alguna manera para asociarlo a una carrera, por ejemplo, “enfermería” o “derecho”.
- **Añadir datos sobre campañas de marketing:** identificando las campañas de marketing se podría analizar qué campañas están teniendo más impacto en la red social. Para esto habría que poder identificarlas en el tuit, mediante algún nombre único que no se confunda con el propio mensaje.
- **Mejora del análisis de sentimiento:** mejorar el análisis de sentimiento ya sea mediante la integración de una herramienta software empresarial o mediante un desarrollo a medida.
- **Realizar un tratamiento previo del texto no estructurado:** para mejorar los resultados del análisis de sentimiento, se podrían realizar procesos ETL de transformación para procesar primero el texto y luego enviarlo al componente de análisis de sentimiento.

8 Referencia

1. **Twitter.** Developer Twitter. *Twitter.* [En línea] [Citado el: 17 de 06 de 2014.] dev.twitter.com.

2. **Amazon.** Capa de uso gratuito AWS. *Amazon.* [En línea] Amazon. [Citado el: 17 de 06 de 2014.] http://aws.amazon.com/es/free/.

3. **Oracle.** Oracle Endeca Information Discovery. *Oracle.com.* [En línea] Oracle. [Citado el: 17 de 06 de 2014.] http://www.oracle.com/us/solutions/business-analytics/business-intelligence/endeca/overview/index.html.

4. **Twitter.** Rest API Twitter. *Twitter.* [En línea] Twitter. [Citado el: 18 de 06 de 2014.] https://dev.twitter.com/docs/api.

5. **McKinsey.** Big data: The next frontier for innovation, competition, and productivity. *McKinsey.* [En línea] McKinsey. [Citado el: 18 de 06 de 2014.] http://www.fujitsu.com/downloads/SVC/fla/03_Michael_Chui.pdf.

6. **Gartner.** Encuesta Gartner revela que el 64% han invertido o planean invertir en Big Data en 2013. *Gartner.* [En línea] Gartner. [Citado el: 18 de 06 de 2014.] http://www.gartner.com/newsroom/id/2593815.

7. Revealed, what happens in just ONE minute on the internet. *http://www.dailymail.co.uk/.* [En línea] [Citado el: 14 de 06 de 2014.] http://www.dailymail.co.uk/sciencetech/article-2381188/Revealed-happens-just-ONE-minute-internet-216-000-photos-posted-278-000-Tweets-1-8m-Facebook-likes.html.

8. The Internet of Things. *http://www.ibmbigdatahub.com/.* [En línea] [Citado el: 17 de 06 de 2014.] http://www.ibmbigdatahub.com/blog/internet-things.

9. Silicon Angle. *Building Big Data: Farming Big Data Goes To The Cows.* [En línea] [Citado el: 17 de 06 de 2014.] http://siliconangle.com/blog/2012/09/07/building-big-data-farming-big-data-goes-to-the-cows/.

10. Gartner Says the Internet of Things Installed Base Will Grow to 26 Billion Units By 2020. *Gartner.* [En línea] [Citado el: 17 de 06 de 2014.] http://www.gartner.com/newsroom/id/2636073.

11. Top 10 List – The V's of Big Data. *Datasciencecentral.* [En línea] Datasciencecentral. [Citado el: 18 de 06 de 2014.] http://www.datasciencecentral.com/profiles/blogs/top-10-list-the-v-s-of-big-data.

12. **BERTUCA, DAVID J.** *The World's Columbian Exposition.* 1996.

13. Nikola Tesla. *Moebious.* [En línea] Moebious. [Citado el: 14 de 06 de 2014.] http://www.moebius-bcn.com/?p=1802.

14. Changing Standards: What it Means to Be Green. *Sustainable Chicago.* [En línea] Sustainable Chicago. [Citado el: 18 de 06 de 2014.] http://www.sustainable-chicago.com/2010/12/16/changing-standards-what-it-means-to-be-green/.

15. **McKendrick, Joseph.** *BIG DATA, BIG CHALLENGES, BIG OPPORTUNITIES.* s.l. : Unisphere Research, 2012.

16. **MIT Sloan Management Review.** *Analytics: The New Path to Value* . s.l. : IBM, 2010.

17. *Apuntes Data Mining 4º Ingeniería Informática UFV.* 2014.

18. **BBC.** Twitter co-founder Jack Dorsey rejoins company. *BBC.* [En línea] BBC. [Citado el: 08 de 06 de 2014.] http://www.bbc.co.uk/news/business-12889048.

19. **Kelly, Ryan.** Twitter Study Reveals Interesting Results About Usage. *Pearlanalytics.* [En línea] Pearlanalytics, 08 de 2009. [Citado el: 10 de 06 de 2014.] http://web.archive.org/web/20120503013715/http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf.

20. **Real Academia Española.** La tecnología al servicio de la palabra. *rae.es.* [En línea] RAE. [Citado el: 18 de 06 de 2014.] http://www.rae.es/sites/default/files/La_tecnologia_al_servicio_de_la_palabra.pdf.

21. *Sentiment analysis and opinion mining.* **Liu, Bing.** s.l. : Synthesis Lectures on Human Language Technologies, 2012, Vols. 5, no 1, p. 1-167.

22. **Turney, Peter D.** *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Review.* Otawwa, Canada : s.n., 2002.

23. **Twitter.** Twitter Analytics. *Twitter.* [En línea] Twitter. [Citado el: 18 de 06 de 2014.] https://analytics.twitter.com/.

24. **Twtrland.** Twtrland. *Twtrland.* [En línea] Twtrland. [Citado el: 19 de 06 de 2014.] http://twtrland.com/.

25. **Twitonomy.** Twitonomy. *Twitonomy.* [En línea] Twitonomy. [Citado el: 18 de 06 de 2014.] http://www.twitonomy.com/.

26. **Topsy.** Topsy. *Topsy.* [En línea] Topsy. [Citado el: 17 de 06 de 2014.] http://topsy.com/.

27. **Simply Measured.** Simply Measured. *Simply Measured.* [En línea] Simply Measured. [Citado el: 18 de 06 de 2014.] http://simplymeasured.com/.

28. **Tweetlevel.** Tweetlevel. *Tweetlevel.* [En línea] Tweetlevel. [Citado el: 18 de 06 de 2014.] http://tweetlevel.edelman.com/.

29. **Mashape.** List of 20+ Sentiment Analysis APIs. *mashape.* [En línea] Mashape. [Citado el: 18 de 06 de 2014.] http://blog.mashape.com/post/48757031167/list-of-20-sentiment-analysis-apis.

30. **Textalytics.** Textalytics. *Textalytics.* [En línea] Textalytics. [Citado el: 10 de 06 de 2014.] textalytics.com.

31. **Dadedalus.** Daedalus. *Daedalus.* [En línea] Daedalus. [Citado el: 18 de 06 de 2014.] http://www.daedalus.es/.

32. **Lexalytics.** Salience Engine. *Lexalytics.* [En línea] Lexalytics. [Citado el: 17 de 06 de 2014.] http://www.lexalytics.com/software.

33. —. Lexalytics. *Lexalytics.* [En línea] Lexalytics. [Citado el: 17 de 06 de 2014.] http://www.lexalytics.com/.

34. **Oracle.** Oracle WebLogic Server. *Oracle.com.* [En línea] Oracle. [Citado el: 17 de 06 de 2014.] http://www.oracle.com/technetwork/middleware/weblogic/overview/index.html.

35. —. Oracle Endeca Information Discovery Integrator Documentation. *Oracle.com.* [En línea] Oracle. [Citado el: 17 de 06 de 2014.] http://docs.oracle.com/cd/E40518_01/index_server_tab.htm.

36. —. Oracle Endeca Information Discovery Server Documentation. *Oracle.com.* [En línea] Oracle. [Citado el: 17 de 06 de 2014.] http://docs.oracle.com/cd/E40518_01/index_server_tab.htm.

37. —. Oracle Endeca Information Discovery Studio Documentation. *Oracle.com.* [En línea] Oracle. [Citado el: 17 de 06 de 2014.] http://docs.oracle.com/cd/E40518_01/index_server_tab.htm.

38. **Lexalytics .** Named Entity Extraction. *Lexalytics.com.* [En línea] Lexalytics . [Citado el: 17 de 06 de 2014.] http://www.lexalytics.com/technical-info/named-entity-extraction.

39. **Twitter.** Rate Limit API Twitter. *Twitter.* [En línea] Twitter. [Citado el: 17 de 06 de 2014.] https://dev.twitter.com/docs/rate-limiting/1.1.

40. **Sommerville, Ian.** *Ingeniería del Software.* s.l. : Pearson, 2004. 84-7829-074-5.

41. **Barry W. Boehm, TRW Defense Systems Group.** *A Spiral Model of Software Development and.*

42. *Asignatura Ingeniería del Software de 3º de Ingeniería Informática de la UFV.* Madrid : s.n., 2012.

43. Topsy. [En línea] http://topsy.com/.

44. **Ministerio de Eduación, Cultura y Deporte.** Fundación del Gobierno de España para la proyección internacional de las universidades españolas. *Ministerio de Eduación, Cultura y Deporte.* [En línea] Ministerio de Eduación, Cultura y Deporte. [Citado el: 18 de 06 de 2014.] http://universidad.es/es/universidades/provincias/madrid.

45. **Amazon.** Amazon AMI Linux. *Amazon.* [En línea] Amazon. [Citado el: 17 de 06 de 2014.] http://aws.amazon.com/es/amazon-linux-ami/2013.09-release-notes/.

46. Putty. *putty.org.* [En línea] [Citado el: 17 de 06 de 2014.] http://www.putty.org/.

47. Credentials Amazon AWS. *Amazon.* [En línea] Amazon. [Citado el: 17 de 06 de 2014.] http://docs.aws.amazon.com/AWSSecurityCredentials/1.0/AboutAWSCredentials.html.

48. HTTP Apache Server. *Apache.* [En línea] [Citado el: 17 de 06 de 2014.] http://httpd.apache.org/.

49. PhpMyAdmin. *phpmyadmin.net.* [En línea] [Citado el: 17 de 06 de 2014.] http://www.phpmyadmin.net/home_page/index.php.

50. **Green, Adam.** 140dev. *140dev.com.* [En línea] 140dev. [Citado el: 17 de 06 de 2014.] http://140dev.com/free-twitter-api-source-code-library/.

51. —. Arquitectura Twitter Database Server. *140dev.com.* [En línea] 140dev. [Citado el: 17 de 06 de 2014.] http://140dev.com/free-twitter-api-source-code-library/twitter-database-server/code-architecture/.

52. —. Esquema MySQL 140.dev. *140dev.* [En línea] 140dev.com. [Citado el: 17 de 06 de 2014.] http://140dev.com/free-twitter-api-source-code-library/twitter-database-server/mysql-database-schema/.

53. Heartbleed. [En línea] http://heartbleed.com/.

54. **Amazon.** Amazon AWS Openssl. *Amazon.* [En línea] Amazon. [Citado el: 14 de 06 de 2014.] http://aws.amazon.com/es/security/security-bulletins/aws-services-updated-to-address-openssl-vulnerability/.

55. —. Amazon Vulnerabilities. *Amazon.* [En línea] Amazon. [Citado el: 18 de 06 de 2014.] http://www.rapid7.com/db/vulnerabilities/amazon-linux-ami-alas-2013-171.

56. Snowball. [En línea] http://snowball.tartarus.org/.

57. **TASS.** Taller de Análisis de Sentimientos en la SEPLN. *daedalus.es.* [En línea] Daedalus. [Citado el: 17 de 06 de 2014.] http://www.daedalus.es/TASS.

58. **Lexalytics.** Lexalytics Salience Dev Wiki. *Lexalytics.* [En línea] Lexalytics. [Citado el: 17 de 06 de 2014.] http://dev.lexalytics.com/wiki/pmwiki.php?n=SalienceWorkbench.Projects.

59. **El Mundo.** El sótano de los horrores de la Universidad Complutense. *El Mundo.* [En línea] El Mundo. [Citado el: 18 de 06 de 2014.] http://www.elmundo.es/madrid/2014/05/18/5378f7d8268e3e14768b4573.html.

60. **Ministerio de Educación, Cultura y Deporte.** Avance de la Estadística de estudiantes universitarios. Curso 2012-2013. *mecd.gob.es.* [En línea] Ministerio de Educación, Cultura y Deporte. [Citado el: 18 de 06 de 2014.] http://www.mecd.gob.es/educacion-mecd/areas-educacion/universidades/estadisticas-informes/estadisticas/alumnado/2012-2013.html.

61. **Oracle.** Oracle Enterprise Linux Intallation Guide. *Oracle.* [En línea] Oracle. [Citado el: 18 de 06 de 2014.] http://docs.oracle.com/cd/E37670_01/E41137/E41137.pdf.

62. —. Oracle Database Intallation Guide. *Oracle.* [En línea] Oracle. [Citado el: 18 de 06 de 2014.] http://docs.oracle.com/cd/E11882_01/install.112/e47689.pdf.

63. —. Oracle Weblogic Installation Guide. *Oracle.* [En línea] Oracle. [Citado el: 18 de 06 de 2014.] http://www.oracle.com/technetwork/middleware/weblogic/documentation/index.html.

64. —. Oracle Enedca Information Discovery Documentation. *Oracle.com.* [En línea] Oracle. [Citado el: 17 de 06 de 2014.] http://docs.oracle.com/cd/E40518_01/index.htm.

65. —. Oracle Endeca Text Enrichment. *Oracle.* [En línea] Oracle. [Citado el: 23 de 06 de 2014.] http://docs.oracle.com/cd/E40520_01/integrator.311/IntegratorETLTextEnrichmentQuickStartGuide.pdf.

66. Limite Rest API Twitter. [En línea] https://dev.twitter.com/docs/api/1.1/get/search/tweets.

67. **Universidad Politécnica de Valencia. .** *Proceso de desarrollo de software* . Valencia : s.n., 2003.

Anexo I: Términos de búsqueda en la primera fase y estudio del número de tuits.

En este anexo se muestran los términos escogidos en una primera fase y el estudio que se realizó para estimar el número de tuits que se debían obtener.

Los datos de la tabla 72 son los tuits generados en el último mes desde la fecha de consulta, el día 14 de febrero de 2014, la fuente de información es Topsy (42).

Tabla 72. Términos de búsqueda y número de tuits. Primera fase. Fuente: Topsy (42)

Término de búsqueda	Número de twits
ceu	78000
ucm	27000
UPM	22000
uam	18000
uah	12000
uned	5700
U_tad	5200
URJC	3400
LA_UPM	3200
ufv	2200
Universidad Rey Juan Carlos	2000
esic	1900
esne	1600
uc3m	1500
ieb	1300
UIMP	1200
NEBRIJA	1000
Universidad Carlos III de Madrid	860
UAX	777
Universidad Internacional Menéndez Pelayo	696
U-TAD	470
ucjc	414
Universidad Alfonso X el Sabio	301
Universidad Complutense de Madrid	261
Universidad Politécnica de Madrid	217
Universidad Camilo José Cela	181
Universidad Europea de Madrid	181
Universidad Francisco de Vitoria	147
UDIMA	124
Universidad CEU San Pablo	123
Universidad Nacional de Educación a Distancia	117
idcsalud	107
eserp	103
Universidad Pontificia Comillas	79
universidad_uax	67
universidadcjc	62
CES Felipe II	57
cunef	43
escuni	39
UEuropea	30
Universidad a Distancia de Madrid	24
Universidad Antonio de Nebrija	24
CUGC	23
Centro de Estudios Garrigues	22
ucomplutense	16
UCOMILLAS	14
uspceu	12
UAM_Madrid	11
UAHes	10
ufvmadrid	10
Centro Universitario de la Defensa	8
lasallemad	6
esne_es	5
RCU María Cristina	5
Escuela Universitaria de Artes y Espectáculos TAI	4

upcomillas	4
villanuevacu	3
escuelatai	1
cucc_educacion	1
Centro Superior de Estudios Universitarios “La Salle”	1
universidadeuropea	1
cud.uah	0
gredossandiego	0
cardenalcisneros	0
euemadrid	0
euf.once	0
lasallecentrouniversitario	0
cesfelipesegundo	0
cesdonbosco	0
CisnerosCU	0
rcumariacristina	0
sanrafaelnebrija	0
centrogarrigues	0
RCU_2013	0
IEB_Spain	0

ESICEducation	0
CEG_Garrigues	0
Centro Universitario "Gredos San Diego"	0
EU de Magisterio "Cardenal Cisneros"	0
EU de Enfermería de la Cruz Roja	0
EU de Enfermería Fundación Jiménez Díaz	0
EU de Fisioterapia de la ONCE	0
Centro Universitario de la Guardia Civil (CUGC)	0
CES CUNEF	0
CES Educación Don Bosco	0
CES Villanueva	0
CES Cardenal Cisneros	0
EU de Profesorado Escuni	0
Instituto de Estudios Bursátiles (I.E.B)	0
CES de Gestión y Marketing (ESIC)	0
Centro Universitario San Rafael	0
ESNE-Escuela Universitaria de Diseño	0
Universidad Autónoma de Madrid	0
Universidad de Alcalá	0

Anexo II: Términos de búsqueda segunda fase, whitelist y texttagged

Para la extracción de los términos de búsqueda se ha seguido el criterio de escoger el nombre de la universidad, su acrónimo y la cuenta de Twitter asociada en la página principal de la universidad.

Se han tomado las 16 universidades de la Comunidad de Madrid, según indica la Fundación del Gobierno de España para la proyección internacional de las universidades españolas del Ministerio de Educación, Cultura y Deporte (43).

El fichero whitelist relaciona el término de búsqueda con la universidad (columna 1 con la columna 2 de la tabla 73). Después del proceso de etiquetado, figura 76 en la fase 3 del desarrollo, se guarda en el metadato texttaged el campo que ha encontrado. Por ejemplo, si encuentra UDIMA en el tuit pondrá en el metadato texttagged "Universidad a Distancia de Madrid".

Tabla 73. Relación de términos de búsqueda con centros asociados y universidad. Fuente: elaboración propia.

Etiqueta (término de búsqueda)	Centro asociado	Universidad
Universidad a Distancia de Madrid	Universidad a Distancia de Madrid	Universidad a Distancia de Madrid
UDIMA	Universidad a Distancia de Madrid	Universidad a Distancia de Madrid
Universidad Alfonso X el Sabio	Universidad Alfonso X el Sabio	Universidad Alfonso X el Sabio
UAX	Universidad Alfonso X el Sabio	Universidad Alfonso X el Sabio
universidad_uax	Universidad Alfonso X el Sabio	Universidad Alfonso X el Sabio
sanrafaelnebrija	Centro Universitario de Ciencias de la Salud de San Rafael	Universidad Antonio de Nebrija
centrogarrigues	Centro de Estudios Garrigues	Universidad Antonio de Nebrija
CEG_Garrigues	Centro de Estudios Garrigues	Universidad Antonio de Nebrija
Centro Universitario San Rafael	Centro Universitario San Rafael	Universidad Antonio de Nebrija
Centro de Estudios Garrigues	Centro de Estudios Garrigues	Universidad Antonio de Nebrija
Universidad Antonio de Nebrija	Universidad Antonio de Nebrija	Universidad Antonio de Nebrija
NEBRIJA	Universidad Antonio de Nebrija	Universidad Antonio de Nebrija
euemadrid	EU de Enfermería de la Cruz Roja	Universidad Autónoma de Madrid
idcsalud	EU de Enfermería Fundación Jiménez Díaz	Universidad Autónoma de Madrid
euf.once	EU de Fisioterapia de la ONCE	Universidad Autónoma de Madrid
lasallecentrouniversitario	Centro Superior de Estudios Universitarios "La Salle"	Universidad Autónoma de Madrid

lasallemad	Centro Superior de Estudios Universitarios “La Salle”	Universidad Autónoma de Madrid
EU de Enfermería de la Cruz Roja	EU de Enfermería de la Cruz Roja	Universidad Autónoma de Madrid
EU de Enfermería Fundación Jiménez Díaz	EU de Enfermería Fundación Jiménez Díaz	Universidad Autónoma de Madrid
EU de Fisioterapia de la ONCE	EU de Fisioterapia de la ONCE	Universidad Autónoma de Madrid
Centro Superior de Estudios Universitarios “La Salle”	Centro Superior de Estudios Universitarios “La Salle”	Universidad Autónoma de Madrid
Universidad Autónoma de Madrid	Universidad Autónoma de Madrid	Universidad Autónoma de Madrid
uam	Universidad Autónoma de Madrid	Universidad Autónoma de Madrid
UAM_Madrid	Universidad Autónoma de Madrid	Universidad Autónoma de Madrid
esne	ESNE-Escuela Universitaria de Diseño	Universidad Camilo José Cela
U_tad	U-TAD	Universidad Camilo José Cela
esne_es	ESNE-Escuela Universitaria de Diseño	Universidad Camilo José Cela
ESNE-Escuela Universitaria de Diseño	ESNE-Escuela Universitaria de Diseño	Universidad Camilo José Cela
U-TAD	U-TAD	Universidad Camilo José Cela
Universidad Camilo José Cela	Universidad Camilo José Cela	Universidad Camilo José Cela
ucjc	Universidad Camilo José Cela	Universidad Camilo José Cela
universidadcjc	Universidad Camilo José Cela	Universidad Camilo José Cela
CUGC	Centro Universitario de la Guardia Civil (CUGC)	Universidad Carlos III de Madrid
Centro Universitario de la Guardia Civil (CUGC)	Centro Universitario de la Guardia Civil (CUGC)	Universidad Carlos III de Madrid
Universidad Carlos III de Madrid	Universidad Carlos III de Madrid	Universidad Carlos III de Madrid
uc3m	Universidad Carlos III de Madrid	Universidad Carlos III de Madrid
Universidad CEU San Pablo	Universidad CEU San Pablo	Universidad CEU San Pablo
uspceu	Universidad CEU San Pablo	Universidad CEU San Pablo
ceu	Universidad CEU San Pablo	Universidad CEU San Pablo
cesfelipesegundo	CES Felipe II	Universidad Complutense de Madrid
cunef	CES CUNEF	Universidad Complutense de Madrid
cesdonbosco	CES en Humanidades y Ciencias de la Educación Don Bosco	Universidad Complutense de Madrid

villanuevacu	CES Villanueva	Universidad Complutense de Madrid
CisnerosCU	CES Cardenal Cisneros	Universidad Complutense de Madrid
rcumariacristina	RCU María Cristina	Universidad Complutense de Madrid
escuni	EU de Profesorado Escuni	Universidad Complutense de Madrid
ieb	Instituto de Estudios Bursátiles (I.E.B)	Universidad Complutense de Madrid
RCU_2013	RCU María Cristina	Universidad Complutense de Madrid
IEB_Spain	Instituto de Estudios Bursátiles (I.E.B)	Universidad Complutense de Madrid
CES Felipe II	CES Felipe II	Universidad Complutense de Madrid
CES CUNEF	CES CUNEF	Universidad Complutense de Madrid
CES Educación Don Bosco	CES Educación Don Bosco	Universidad Complutense de Madrid
CES Villanueva	CES Villanueva	Universidad Complutense de Madrid
CES Cardenal Cisneros	CES Cardenal Cisneros	Universidad Complutense de Madrid
RCU María Cristina	RCU María Cristina	Universidad Complutense de Madrid
EU de Profesorado Escuni	EU de Profesorado Escuni	Universidad Complutense de Madrid
Instituto de Estudios Bursátiles (I.E.B)	Instituto de Estudios Bursátiles (I.E.B)	Universidad Complutense de Madrid
Universidad Complutense de Madrid	Universidad Complutense de Madrid	Universidad Complutense de Madrid
ucm	Universidad Complutense de Madrid	Universidad Complutense de Madrid
ucomplutense	Universidad Complutense de Madrid	Universidad Complutense de Madrid
cud.uah	Centro Universitario de la Defensa	Universidad de Alcalá
gredossandiego	Centro Universitario "Gredos San Diego"	Universidad de Alcalá

cardenalcisneros	EU de Magisterio "Cardenal Cisneros"	Universidad de Alcalá
cucc_educacion	EU de Magisterio "Cardenal Cisneros"	Universidad de Alcalá
Centro Universitario de la Defensa	Centro Universitario de la Defensa	Universidad de Alcalá
Centro Universitario "Gredos San Diego"	Centro Universitario "Gredos San Diego"	Universidad de Alcalá
EU de Magisterio "Cardenal Cisneros"	EU de Magisterio "Cardenal Cisneros"	Universidad de Alcalá
Universidad de Alcalá	Universidad de Alcalá	Universidad de Alcalá
uah	Universidad de Alcalá	Universidad de Alcalá
UAHes	Universidad de Alcalá	Universidad de Alcalá
Universidad Europea de Madrid	Universidad Europea de Madrid	Universidad Europea de Madrid
UEuropea	Universidad Europea de Madrid	Universidad Europea de Madrid
universidadeuropea	Universidad Europea de Madrid	Universidad Europea de Madrid
Universidad Francisco de Vitoria	Universidad Francisco de Vitoria	Universidad Francisco de Vitoria
ufv	Universidad Francisco de Vitoria	Universidad Francisco de Vitoria
ufvmadrid	Universidad Francisco de Vitoria	Universidad Francisco de Vitoria
Universidad Internacional Menéndez Pelayo	Universidad Internacional Menéndez Pelayo	Universidad Internacional Menéndez Pelayo
UIMP	Universidad Internacional Menéndez Pelayo	Universidad Internacional Menéndez Pelayo
Universidad Nacional de Educación a Distancia	Universidad Nacional de Educación a Distancia	Universidad Nacional de Educación a Distancia
uned	Universidad Nacional de Educación a Distancia	Universidad Nacional de Educación a Distancia
Universidad Politécnica de Madrid	Universidad Politécnica de Madrid	Universidad Politécnica de Madrid
UPM	Universidad Politécnica de Madrid	Universidad Politécnica de Madrid
LA_UPM	Universidad Politécnica de Madrid	Universidad Politécnica de Madrid
Universidad Pontificia Comillas	Universidad Pontificia Comillas	Universidad Pontificia Comillas
UCOMILLAS	Universidad Pontificia Comillas	Universidad Pontificia Comillas
upcomillas	Universidad Pontificia Comillas	Universidad Pontificia Comillas
eserp	CES Escuela Superior Empresarial de Relaciones Públicas ESERP	Universidad Rey Juan Carlos
esic	CES de Gestión y Marketing (ESIC)	Universidad Rey Juan Carlos

escuelatai	Escuela Universitaria de Artes y Espectáculos TAI	Universidad Rey Juan Carlos
ESICEducation	CES de Gestión y Marketing (ESIC)	Universidad Rey Juan Carlos
CES de Gestión y Marketing (ESIC)	CES de Gestión y Marketing (ESIC)	Universidad Rey Juan Carlos
Escuela Universitaria de Artes y Espectáculos TAI	Escuela Universitaria de Artes y Espectáculos TAI	Universidad Rey Juan Carlos
Universidad Rey Juan Carlos	Universidad Rey Juan Carlos	Universidad Rey Juan Carlos
URJC	Universidad Rey Juan Carlos	Universidad Rey Juan Carlos

Anexo III: Instalación de componentes de Oracle Endeca Information Discovery

Se adjunta al proyecto una máquina virtual con todos los componentes instalados y preparada para su despliegue.

La máquina virtual contiene 3 discos duros dinámicos:

- OracleLinux65: sistema operativo
- Database: base de datos
- Endeca: componentes Endeca

El procedimiento de arranque viene adjunto con los scripts de arranque (no será necesario copiar nada ya que están dentro de la máquina virtual) en la documentación electrónica, en el fichero "Manual.pdf". En ese documento se especifican las contraseñas necesarias para su acceso.

Los siguientes componentes están instalados en la máquina virtual.

- **Oracle Enterprise Linux 6.5**
  - Documentación (60)
- **Oracle Database 11gR2**
  - Documentación (61)
- **Oracle WebLogic Application Server 10.3.6**
  - Documentación (62)
- **Oracle Endeca Server 7.6.1**
  - Documentación (35)
- **Oracle Endeca Studio 3.1**
  - Documentación (36)
- **Oracle Endeca Provisioning Service 3.1**
  - Documentación (63)

Los siguientes componentes están instalados en local:

- **Oracle Integrator ETL**
  - Documentación (34)
- **Lexalytics Salience Engine**
  - Documentación (64)

Anexo IV: Vistas y metadatos.

Las 16 vistas utilizadas vienen en formato electrónico en la carpeta “Vistas”.

Los metadatos son los datos que se envían de un componente a otro. Visto de otra manera, equivalen a las columnas en una tabla. Desde que se adquiere el dato hasta que se carga en el servidor, estos van modificándose hasta obtener los siguientes metadatos. Algunos de ellos no se utilizarán pero se cargarán para por si se necesitan.

Los metadatos más importantes, con los que se trabajará desde Oracle Endeca Studio son los siguientes metadatos (tabla 74).

Tabla 74. Metadatos utilizados. Fuente: elaboración propia.

Categoría	Nombre del metadato	Descripción
Date	tweeted_time	Hora
Date	tweeted_date	Día
Date	tweeted_month	Mes
Measures	followers_count	Número de seguidores
Measures	friends_count	Numero de amigos
Measures	retweet_count	Número de retuits
TwitterAPI	created_at	Fecha de creación
TwitterAPI	text	Tuit (solo el texto)
TwitterAPI	source	Fuente
TwitterAPI	from_user_id	Id de usuario
TwitterAPI	from_user_name	Nombre de usuario
TwitterAPI	location	Localización
TwitterAPI	User Description	Descripción del usuario
TwitterAPI	profile_image_url	URL de la imagen
TwitterAPI	profile_image_url_https	
TwitterAPI	time_zone	Zona horaria
TwitterAPI	geo	Coordenadas de localización
Text Enrichment	DocumentSentiment	Puntuación de sentimiento
Text Enrichment	EntitiesPerson	Entidades persona identificadas
Text Enrichment	EntitiesProduct	Entidades producto identificadas

Text Enrichment	EntitiesCompany	Entidades empresa identificadas
Text Enrichment	EntitiesPositive	Entidades positivas identificadas
Text Enrichment	EntitiesNeutral	Entidades neutrales identificadas
Text Enrichment	EntitiesNegative	Entidades negativas identificadas
Text Enrichment	EntitiesList	Entidades configuradas por el usuario (en este caso relación centros asociados con universidades)
Text Enrichment	Themes	Temas extraídos
Text Enrichment	ThemesPositive	Temas positivos
Text Enrichment	ThemesNeutral	Temas neutrales
Text Enrichment	ThemesNegative	Temas negativos
Text Enrichment	Summary	Resumen
Text Enrichment	texttagged	Universidad a la que pertenece
Text Enrichment	Sentiment	Sentimiento positivo, negativo o neutral
Other	TweetKey	Id del registro (puesta por un proceso ETL)

Anexo V: Contenido del medio electrónico adjunto

El CD adjunto viene con las siguientes carpetas y documentos:

- Carpeta Vistas: contiene las 16 vistas utilizadas.
- Carpeta Proyecto Integrator: contiene el proyecto en Integrator ETL con todos los procesos ETL que han sido necesarios.
- Carpeta Análisis de sentimiento: contiene las mejoras realizadas en el diccionario así como los ficheros de pruebas.
- PFG_Alvaro_Ullate.docx: memoria final.
- PFG_Alvaro_Ullate.pdf: memoria final.
- Manual.pdf: contiene el manual de arranque y primeros pasos de la máquina virtual.
- Leame.txt: explicación del contenido.

Además, se adjunta una memoria USB con la máquina virtual de 50GB aproximadamente (carpeta Maquina_virtual).