



UNIVERSIDAD DE JAÉN

Facultad de Ciencias Sociales y Jurídicas

Trabajo Fin de Grado

**MINERIA DE DATOS
EN REDES SOCIALES
PARA PYMES.**

Alumno: Ruiz Ramírez, Álvaro

Enero, 2018

Índice

1. Introducción	1
2. Objetivos	1
3. Antecedentes	2
3.1. ¿Qué es la Minería de Datos?	2
3.2. ¿Qué son las Redes Sociales?	4
3.3. Minería de datos en Redes Sociales.	6
3.4. Twitter como Red Social y su API.	6
3.5. Herramientas y otros servicios.	10
4. Estructura del software desarrollado	15
4.1. Entorno de desarrollo: Rstudio.	15
4.2. Registro y Creación de Twitter App.	16
4.3. Proceso de Autorización y Autenticación para Twitter API.	19
4.4. Extracción de la información y almacenamiento. TwitterR	20
4.5. Limpieza y preprocesado de la información.	21
4.6. Términos frecuentes y Bag Words	24
5. Ejemplo de uso: Binteak Music	25
6. Conclusiones y trabajo futuro	31
6.1. Conclusiones.	31
6.2. Líneas de trabajo futuro.	31
Bibliografía y Referencias	32

Índice de figuras

1.	3.1 Etapas KDD.	4
2.	3.2 Cuenta en Twitter de la Universidad de Jaén.	7
3.	3.3 Estructura Rest API.	9
4.	3.4 Estructura Streaming API.	10
5.	3.5 Interfaz gráfica Oracle Data Mining.	11
6.	3.6 Interfaz gráfica Tableau.	12
7.	3.7 Interfaz gráfica Octoparse.	12
8.	3.8 Interfaz gráfica Rapid Miner.	13
9.	3.9 Interfaz gráfica Knime.	14
10.	3.10 Interfaz gráfica Orange Data Mining.	15
11.	4.1 IDE de Rstudio.	16
12.	4.2 Formulario de Twitter App.	17
13.	4.3 Twitter App creada.	18
14.	4.4 Detalles de Twitter App.	18
15.	4.5 Claves de acceso a API de Twitter.	19
16.	4.6 Autorización Twitter App.	19
17.	4.7 Código PIN de acceso.	20
18.	4.8 Establecimiento de conexión en Rstudio.	20
19.	4.9 Estructura en lista de la información extraída.	21
20.	4.10 Tweets sin transformar.	22
21.	4.11 Transformaciones en tweets.	22
22.	4.12 Tweets transformados.	23
23.	4.13 Corpus.	23
24.	4.14 Imagen WordCloud.	24
25.	5.1 Visualización de tweet redactado.	25
26.	5.2 Conjunto de cuentas de Twitter.	26
27.	5.3 Gráfico de frecuencias para Hastags.	27
28.	5.4 Gráfico de frecuencias para menciones y cuentas.	28
29.	5.5 WordCloud con las palabras más frecuentes.	29
30.	5.6 Tweets redactados en base a la información relevante.	30

Agradecimientos

En primer lugar, quiero dar las gracias a mi familia y amigos por confiar en mis posibilidades y estar ahí en los buenos y malos momentos por los que he pasado a lo largo de mi etapa universitaria.

Dar las gracias a mi tutor Carlos Molina por guiarme y ayudarme durante la realización de mi trabajo en las dudas que se me pudieron ir generando sobre la marcha. Además del resto de profesores de los que durante mi tiempo de formación en mi carrera he ido tomando todos aquellos conocimientos sin los cuáles no podría haber llevado a cabo este trabajo.

Por último, dar gracias a esa gran comunidad de Internet que mediante su colaboración de manera altruista y desinteresada, con sus aplicaciones y métodos recogidos en diferentes foros de discusión han dado apoyo a resolver otros problemas de diversa magnitud que se han ido encontrando durante el transcurso de este trabajo.

Resumen

Las técnicas de minería de Datos permiten obtener información relevante de un conjunto de datos. Debido al aumento constante y al instante del volumen de la información que es generada por usuarios de cualquier parte del mundo y el uso de las redes sociales, se plantea el estudio de estas tecnologías para que pequeñas y medianas empresas puedan abrirse hueco en el mercado hallando patrones repetitivos, tendencias o reglas que permitan el estudio del comportamiento de un determinado tipo de datos en un contexto social.

Palabras clave: Twitter, Minería de Datos, Ciencia de Datos, Minería de Textos, Tendencias, Redes Sociales.

Abstract

Data mining techniques allow obtaining relevant information from a set of data. Due to the constant and instantaneous increase in the volume of social information, the study of these technologies for small and medium-sized companies opens the gap in the market, there are repetitive patterns, trends or rules that allow the study of the behavior of a type of data in a social context.

Keywords: Twitter, Data Mining, Data Science, Text Mining, Trends, Social Networks.

1. Introducción

Actualmente, parte de los propietarios de micro, pequeña y mediana empresa provienen de anteriores trabajos asalariados y otros muchos de una población económicamente inactiva. Es por ello que en una gran parte de los casos, sus fundadores crean este tipo de empresas por motivos meramente económicos, sin tener un perfil claramente empresarial, lo que genera que estos negocios no se encuentren administrados de la manera más adecuada (SDE,2008).

Considerando que las Pymes se encuentran limitadas ante grandes empresas por un simple factor de equidad frente a éstas, existen otras restricciones como obtener equipo, adaptar sus productos al cliente o poseer información de mercado relevante. Al no contar con los conocimientos administrativos necesarios o garantías para obtener créditos, tienden a limitar la utilización de nuevas tecnologías, desconocer el mercado que las margina, ver muy complicado conseguir posicionarse en ellos o entender a los posibles clientes que cada vez son más exigentes (Hernández, 2007).

Desde el nacimiento de Internet, y apoyado por el auge de las redes sociales y su gran uso, el crecimiento de la información existente ha ido aumentando de manera exponencial, continua y al instante. Así se plasma en el informe llevado a cabo de manera conjunta por la empresa de investigación IDC junto a la de almacenamiento EMC realizado en 2014, en la que los resultados arrojados calcularon que el universo digital alcanzaría un volumen aproximado de 44 zettabytes en 2020. Es por ello que cada vez resulta más interesante el poder llegar a analizar de una forma sistemática, estructurada y con el propósito de sacar valor para estas empresas la información relevante que poseen estos grandes volúmenes de datos.

2. Objetivos

Para que las Pymes puedan competir con grandes empresas a la hora de obtener conclusiones y hallar valor a los datos, se propone en este trabajo el desarrollo de ciertas técnicas en Minería de Datos que permitan solventar algunos de estos problemas de desventaja competitiva, sobre todo en Redes Sociales, debido al gran contenido en información generado al instante por el gran volumen de usuarios que las utilizan día a día, opinando sobre productos, marcas, tendencias y otros factores que pueden ser interesantes de analizar para estos tipos de negocio.

Es por ello que el objetivo de este trabajo se centrará en desarrollar, en fase inicial, una herramienta con la que las Pymes puedan hacer frente a otras empresas, por medio de técnicas de Minería de Datos como el Text Mining en redes sociales para llevar a cabo un análisis de tendencias y patrones para poder así crear una ventaja competitiva.

3. Antecedentes

En esta sección se desarrollará la evolución en el concepto de la Minería de datos desde sus inicios hasta nuestra época. También se explicará el concepto de Redes Sociales en Internet. Por otro lado se procederá a relacionar ambos conceptos y la aplicación de la Minería de Datos en Redes Sociales. Para finalizar este apartado se detallarán herramientas y otros servicios ya existentes.

3.1. ¿Qué es la Minería de Datos?

El concepto de Minería de Datos tiene sus inicios en los años sesenta, época en la cual los estadísticos de entonces ya utilizaban términos como Data Fishing, Data Mining incluso Data Archeology teniendo en mente la búsqueda de correlaciones sin una definición previa de las bases de datos.

A finales de los años ochenta, diferentes autores como Robert Blum, Rakesh Agrawal o Gregory Piatetsky-shapiro entre otros, empezaron a asentar las bases del término de Minería de Datos.

Estos autores fueron mejorando estas herramientas durante el tiempo por lo que podemos dividir esta evolución en cuatro etapas principales:

- Colección de Datos (1960)
- Acceso de Datos (1980)
- Almacén de Datos y Apoyo a las Decisiones (principios de 1990)
- Minería de Datos Inteligente (finales de 1990)

Por aquel entonces, eran escasas las empresas que se dedicaban a esta nueva tecnología.

En 2002 empieza a incrementarse su uso, contando con este tipo de herramientas más de 100 empresas alrededor del mundo ofreciendo más de 300 soluciones. Para entonces, investigadores de todo el mundo, en más de ochenta países creaban foros de discusión sobre la Minería de Datos.

- **Definición formal**

La Minería de datos es conocida actualmente como Data Mining¹, business intelligence incluso big data. El nombre depende bastante del tipo de área de actuación pero todos ellos hacen referencia a un mismo problema: Extraer el conocimiento existente a partir de bases de datos.

Su definición formal sería la extracción no trivial de información implícita, que se plantea de manera previa como desconocida y que será capaz de proporcionar otro tipo de información a partir de ella que es potencialmente útil a partir de los datos a estudio”. Por otro lado, también puede entenderse como “aquella exploración y análisis, por medios automáticos o semiautomáticos, de un gran volumen de datos con el propósito de hallar patrones con un significado.

¹ http://www.sinnexus.com/business_intelligence/datamining.aspx

A la hora de proceder a una investigación, las técnicas en minería de datos permiten a los científicos obtener la capacidad de clasificar, segmentar datos o formar diversas hipótesis. Es así que estas técnicas ayudan a encontrar aquella información escondida en los datos que no resulta fácil de ver a simple vista, principalmente debido al volumen gigantesco de datos existentes.

- **Protocolo de extracción del conocimiento (KDD)**

El proceso de KDD² (Knowledge Discovery in Databases) es comúnmente definido como “el proceso no trivial de identificar patrones válidos, novedosos y potencialmente útiles y en última instancia, comprensible a partir de los datos”. También puede encontrarse bajo el nombre de Data Archeology, Pattern Data Analysis o KnowledgeFishing.

KDD engloba de manera convergente a diversas disciplinas por la investigación en estas bases de datos. Entre muchas otras, algunas de ellas son el Machine Learning, la Estadística, Inteligencia Artificial (IA), Sistemas de apoyo para la toma de decisiones (DSS) o sistemas de gestión de DDBB.

Para llevar a cabo un buen proyecto en base a la Minería de datos se establece un protocolo básico KDD que cuenta con varias fases diferenciadas:

- **Selección de los datos:** Se buscará el objetivo y las herramientas necesarias para el proceso de minería, buscando e identificando los datos a extraer, buscando los atributos de entrada y la información que se arrojará de salida a la hora de presentarlos, es decir, qué se quiere obtener y qué datos nos facilitarán la información que se quiere conocer antes de comenzar el proceso en cuestión.
- **Preprocesamiento:** Para este momento del proceso se procede a la preparación y limpieza de los datos que son extraídos a partir de las diversas fuentes que hemos consultado para etapas posteriores. En este punto se tiende a manejar diferentes estrategias para el manejo de datos faltantes o en blanco, qué estén fuera de rango para que al final, la estructura de los datos sea la adecuada para su posterior transformación.
- **Transformación:** En esta etapa se trata de una manera preliminar los datos, se transforman y se generan nuevas variables a partir de las que ya se disponían, todo ello para contar con los datos ya estructurados a la hora de proceder al empleo de las técnicas de minería de datos adecuadas.
- **Data Mining:** La fase del modelamiento en la que se emplearán las diferentes herramientas de análisis en base a los datos que se tienen. El objetivo será extraer patrones que se consideraban previamente desconocidos o de nueva naturaleza, que puedan ser útiles, comprensibles y que se encontraban ocultos hasta el momento.

² <http://mnrva.io/kdd-platform.html>

- **Interpretación y Evaluación:** Tras la fase de modelamiento se toman en cuenta los resultados obtenidos gracias a esa etapa, analizando los patrones obtenidos y que resultan interesantes para su evaluación.

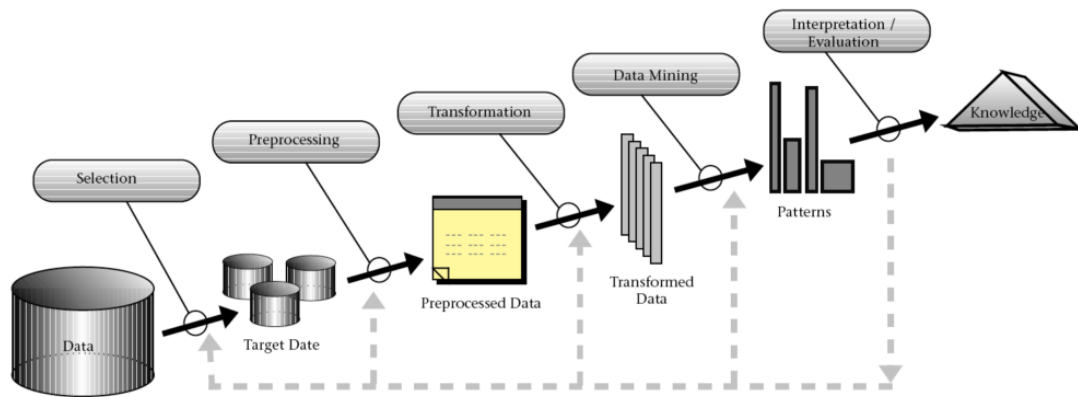


Figura 3.1: Etapas KDD.

- **Técnicas de minería de datos**

Son muchas y diversas las técnicas de minería de datos con las que se cuenta, dependiendo del problema a tratar por cada una de ellas, dónde se proceden a usarse o qué tipo de datos son los que se manejan.

Las técnicas más usadas para la minería de datos son las Redes Neuronales, Árboles de decisión, Algoritmos de agrupamiento Clustering o Reglas de asociación. Además, existen diferentes algoritmos supervisados y no supervisados. Por otro lado, se cuenta con otras técnicas más concretas como el Text Mining.

3.2. ¿Qué son las Redes Sociales?

Las redes sociales son estructuras que permiten comunicar entre sí a personas e instituciones, creando la posibilidad de establecer relaciones entre ellos para lograr construir comunidades online con intereses comunes.

Estar en estas redes sociales supone la posibilidad de tener un contacto de carácter ilimitado y en tiempo real. Esto es posible debido a la interactividad que ofrecen con sus seguidores. La actualización es instantánea, ya que la información proporcionada por sus usuarios se comparte en el preciso momento en el que ocurre, creciendo de manera exponencial.

- **Evolución de las Redes Sociales**

No hace mucho del origen del concepto de redes sociales. En 1995 surge la creación de la web classmates.com por parte del estadounidense Randy Conrads. Su intención fue la de crear un proyecto que pudiera reunir ex compañeros de universidad. Al comprobar el éxito, fueron apareciendo posteriormente diferentes redes para reunir amigos. A partir del

año 2003 diferentes redes como LinkedIn o MySpace con objetivos mucho más específicos fueron tomando gran popularidad.

Dos de las redes sociales que se encuentran en mayor uso a día de hoy son Facebook y Twitter. Facebook fue creado alrededor de 2004 por un grupo de estudiantes junto a Mark Zuckerberg. El propósito inicial de esta red social fue poner en contacto a aquellos estudiantes pertenecientes a la Universidad de Harvard (USA) pero al poco tiempo cualquier persona podía unirse. Facebook comenzó a ganar gran popularidad en el ámbito estudiantil y posteriormente se abrió aún más sumando diferentes tipos de perfiles más distintos. La posterior traducción de toda la red a diferentes idiomas produjo que se expandiera mundialmente, contando en la actualidad con más de 1000 millones de usuarios.

Twitter fue creado poco después de Facebook, en el año 2006. Con sus conocidos “tuits”, publicados en la página principal del usuario, permite comunicar información en tiempo real en breves cadenas de texto.

Antes del nacimiento de estas primeras redes, las empresas y los medios de información no tenían la posibilidad de comunicarse con su público en tiempo real y mucho menos conocer su opinión. En el momento actual, esto ha pasado a ser una mera anécdota donde Facebook y Twitter fueron adquiriendo un mayor protagonismo hasta alcanzar las primeras posiciones como redes sociales en uso.

- **Usos de las Redes Sociales**

- **Marketing:** A la hora de darse a conocer, las empresas utilizan las redes sociales como gran escaparate e imagen corporativa. Suelen usarse para comunicar de una manera mucho más económica frente a los elevados costes que supone publicitarse en otros medios como pueden ser radio o televisión. También permiten generar un mayor alcance y una elevada difusión. Permite mantener una relación directa y personalizada con el público objetivo.
- **Entretenimiento:** El uso de redes sociales por parte de los usuarios es notorio debido a que pueden ser usadas al instante. Este uso se está convirtiendo en uno de los pasatiempos más recurrentes.
- **Comunicación e información:** Con el avance de las nuevas tecnologías, se ha incrementado el uso de dispositivos electrónicos en general y de dispositivos móviles en particular. Las relaciones sociales han pasado a un nuevo plano en el que la gente puede relacionarse al instante por medio de estas redes, es por ello que se genere un aumento de las empresas que ofrecen productos y servicios dentro de ellas.

- **Tipos de Redes Sociales**

Según la función que tienen todas las Redes Sociales se puede clasificar en **genéricas**. Son las más numerosas y goza de mayor popularidad. Consideradas como nuevos medios de comunicación masivos. Las más comunes son Facebook y Twitter.

Las redes **profesionales** como LinkedIn, son las que aglutinan a individuos que comparten relación en el terreno laboral o que buscan conseguir un mayor número de contactos en el ámbito laboral. Este tipo de redes pueden ser abiertas o cerradas.

Por otro lado, en las redes temáticas posibilitan la conexión entre usuarios que poseen intereses en común más específicos, ya pueda ser aquellos como intereses musicales, hobbies, deportes o fotografía, siendo la más famosa Flickr.

3.3. Minería de Datos en Redes Sociales

La relación de las redes sociales con la minería de datos ha supuesto una verdadera revolución para las empresas. La posibilidad para estas empresas de analizar grandes volúmenes de datos en tiempo real y sacarle partido, es un valor añadido que cada vez más es tenido en consideración.

Según Nielsen Online, alrededor del 67 por ciento de los usuarios que usan redes sociales en Internet a nivel mundial lo hacen por medio de Facebook, Twitter o LinkedIn. Este elevado uso es de gran importancia debido al volumen de información que es generado sobre estos clientes o potenciales clientes que es posible llegar a analizar por las empresas, compartiendo sus datos personales, red de contactos, publicaciones e incluso conversaciones en red.

Actualmente se están creando empresas especializadas en vigilancia de medios de comunicación de carácter social en los que tienen como objetivo recopilar diferente información de los usuarios en redes sociales, los cuales generan gráficos sociales que buscan sintetizar la información y hacerla entendible.

Jesse Torres, presidente y CEO del banco Panamericano de Los Ángeles considera que estas empresas vienen a satisfacer una necesidad. En su opinión estas técnicas incrementan la satisfacción del cliente, que recibe ofertas personalizadas y de gran interés al haber sido diseñadas a partir de la información generada por ellos.

3.4. Twitter como Red Social y su API

Twitter³ consiste en una aplicación web gratuita de microblogging que combina a la perfección ventajas de tipo blog, redes sociales y mensajería instantánea entre otros servicios. Permite enviar mensajes de texto hasta un máximo de 140 caracteres por tweet, nombre que reciben las publicaciones que cada perfil de usuario y/o empresa va añadiendo a esta red social desde su perfil. El usuario de esta red social tiene la posibilidad de seguir a otros usuarios, ver y compartir tweets. A twitter se accede desde su web, mediante smartphones e incluso en algunos países a partir de SMS.

³ <https://twitter.com/>



Figura 3.2: Cuenta en Twitter de la Universidad de Jaén.

Twitter utiliza una serie de palabras que conforman su propio lenguaje dentro de su red social. Para poder entender los diferentes términos que aparecerán en posteriores apartados a continuación se detallan los más relevantes:

- **Tweet (Tuit):** Actualización o publicación en la red social.
- **Timeline (TL):** Compone la lista de tweets enviados desde las diferentes cuentas que un usuario sigue y que van apareciendo de manera cronológica y automática en el momento de su publicación.
- **Followers (Seguidores):** Son los distintos usuarios que siguen una cuenta y leen los tweets que son enviados.
- **Following (Seguidos):** Cuentas de Twitter que un usuario sigue. Sus tweets irán a la timeline del usuario que los siga.
- **Retweet (RT o retuit):** Permite volver a mostrar un tweet, ya sea propio o de otro usuario independientemente del momento en el que fue subido, mostrándose en nuestro timeline y muro de perfil, haciendo referencia a su autor.
- **Favoritos (Fav):** Corresponden a los tweets que un usuario de esta red considera de interés, ser afín a sus gustos o sentirse identificado con ellos. Aparecen en la pestaña con el mismo nombre dentro del perfil del usuario que los señaló.

- **Menciones (@usuario):** Como su mismo nombre indica tienen que ver con el llamamiento a la conversación por medio de enlaces de cadenas de tuits entre diferentes usuarios o cuentas para crear un hilo de conversación.
- **Replies:** Respuesta a la mención que se ha generado por otra cuenta o usuario, que hace que el hilo continúe creciendo en tweets.
- **Hashtags (#):** Son palabras clave o etiquetas que se muestran precedidas por el símbolo almohadilla #. Permite enlazar cadenas de tuits en las que aparezca esto mismo, lo que ayuda a recopilar y mostrar mensajes relacionados, permitiendo también una mayor comodidad en la búsqueda de un tema en concreto.
- **Trending Topics (TT):** Son aquellas palabras clave más utilizadas por un período de tiempo concreto en Twitter. Tratan de aquello que es tendencia y de lo que más se habla en ese momento dentro de la red social.

Por otro lado la información que se puede extraer de esta red social es A continuación se exponen los diferentes tipos de información que Twitter proporciona:

- **Quién.** El usuario que escribe, hace un retweet, junto con sus datos públicos: nombre, localización, lenguaje.
- **Qué.** Información relevante a modo de tweet, en el que se muestra el texto de cada mensaje, links, contenido multimedia, menciones a otras cuentas.
- **Cuándo.** Obtendremos la fecha y la hora en tiempo real a partir de las publicaciones de los usuarios.
- **Dónde.** Coordenadas geográficas de la ubicación desde donde fue publicado. En la mayoría de los tweets no sabremos desde donde se conecta cada usuario y en otros casos no sabremos si el lugar elegido por cada usuario se corresponde realmente con el lugar en el que se está escribiendo esos tweets.

Particularmente en este trabajo, no se tendrá en cuenta la geolocalización de los diferentes tweets por lo anteriormente citado.

• **Twitter API**

Twitter ofrece a sus desarrolladores su propia API (Application Programming Interface) oficial. Esta API nos proporciona la comunicación entre los distintos softwares que lo conforman por medio de una capa que los engloba. Esta API permite controlar nuestra cuenta, además de la posibilidad de recuperar la información desde el código. Actualmente a 2017 la API de Twitter se encuentra en la versión 1.1 y se encuentra limitado entre 150 y 350 solicitudes por hora dependiendo de si hemos registrado o no nuestra aplicación en el apartado de desarrolladores de esta red.

Este es uno de los motivos por los que Twitter se ha expandido en los últimos años. La posible creación de APIs de carácter gratuito que proporciona, ayuda a la creación de aplicaciones de terceros para poder permitir la conexión y utilizar datos de la aplicación.

Hay tres maneras en las que se puede extraer información de Twitter: Search API, Rest API y Streaming API.

- **Search API**⁴: Suministra los tweets buscados con un límite de 7 días y un total de 1.500 tweets como máximo. En este tipo de API es posible la filtración de tweets por cliente utilizado, así como por lenguaje y su localización.
- **Rest API**⁵: Esta API web funciona por medio de HTTP a la que se accede a partir de URLs que devuelven contenido en distintos formatos (JSON, XML, HTML) A diferencia.

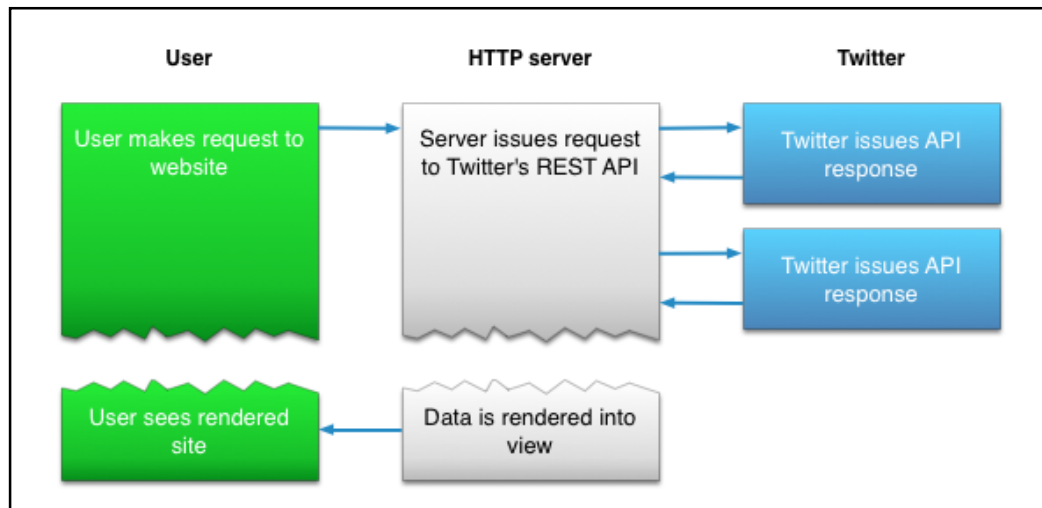


Figura 3.3: Estructura Rest API.

- **Streaming API**⁶: Proporciona información en tiempo real. Los contenidos que devuelve tienen formato JSON. Se obtienen muestras aleatorias o por filtrado de palabras clave, usuarios. Además, también existen otros métodos más interesantes como el poder obtener un conjunto de tweets o filtrar únicamente por tweets retuiteados o con enlaces.

⁴ <https://dev.twitter.com/rest/public/search>

⁵ <https://dev.twitter.com/rest/public>

⁶ <https://dev.twitter.com/streaming/overview>

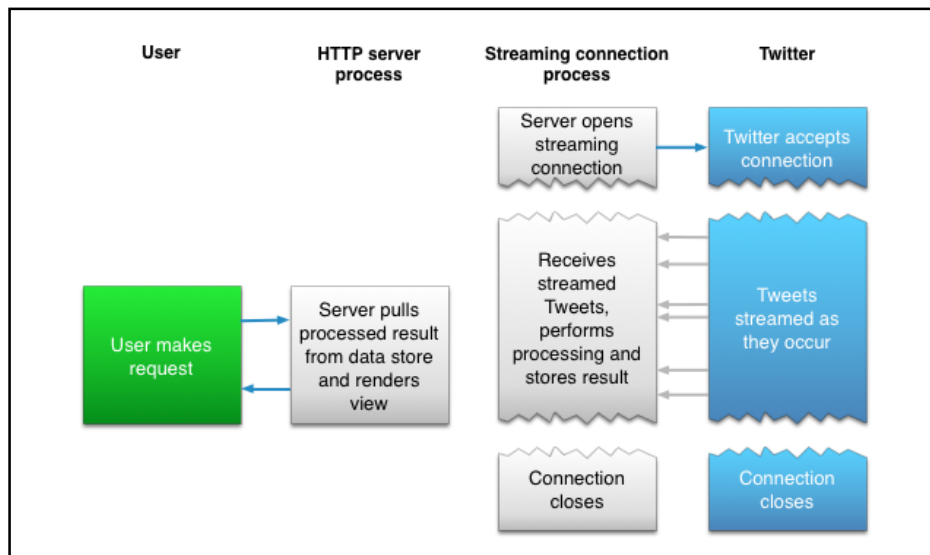


Figura 3.4: Estructura Streaming API.

3.5. Herramientas y otros servicios

En la actualidad existen una gran cantidad de empresas que deciden elegir diferentes herramientas y software que les permitan llevar a cabo análisis en redes sociales. La mayor parte de ellas son de pago y se encuentran enfocadas a satisfacer necesidades por parte de las empresas que los demandan. Algunas de las más importantes son:

- **Oracle Data Mining**



Esta herramienta dispone de una serie de algoritmos de minería de datos alojados por medio de una interfaz gráfica de usuario IBM y su IDE (Entorno de desarrollo), los cuáles se encuentran dispuestos al flujo de trabajo consiguiendo así que a largo plazo el cliente pueda ir desarrollando esos algoritmos de minería de datos con el tiempo.

La herramienta dispone de diferentes técnicas de Data Mining como son Análisis Clúster, Algoritmos bayesianos, Árboles de decisión o modelos lineales generalizados.

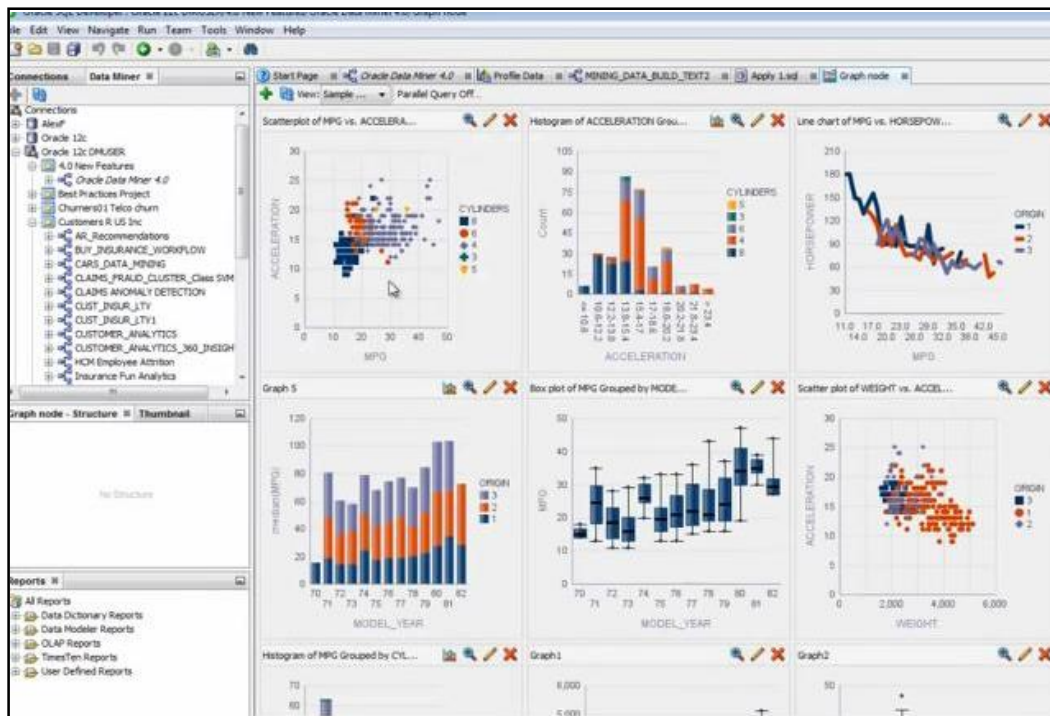


Figura 3.5: Interfaz gráfica Oracle Data Mining.

- **Tableau**



Consta de servicios de escritorio así como online y móvil. Dispone de diferentes servicios. Desde Análisis de encuestas y visualización intuitiva de datos, pasando por análisis de sitios webs, series temporales hasta herramientas Big Data o Dashboards de negocios.

En el caso de Redes Sociales ayuda por medio de análisis de patrones a aumentar la popularidad de marca.

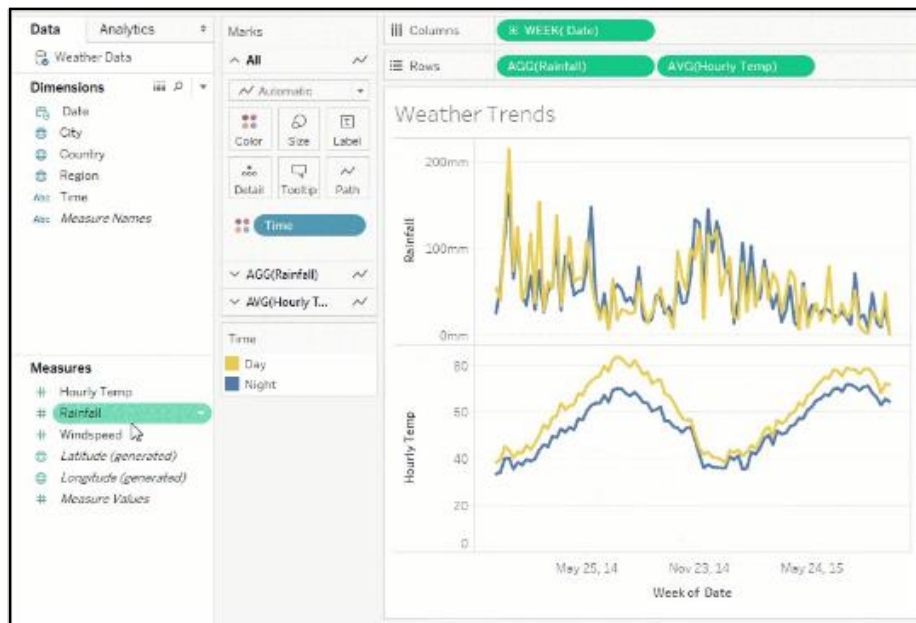


Figura 3.6: Interfaz gráfica Tableau.

- Octoparse



Se caracteriza por su fácil manejo extrayendo datos a un simple click sin necesidad de crear código y a una gran velocidad. Extrae datos de diferentes webs y los exporta en formatos como CSV, Excel o salvándolos en bases de datos.

Por otro lado ofrece servicios en la nube y webscrapping. Disponen de un servicio dedicado y profesional que se adapta a las características de cada empresa.

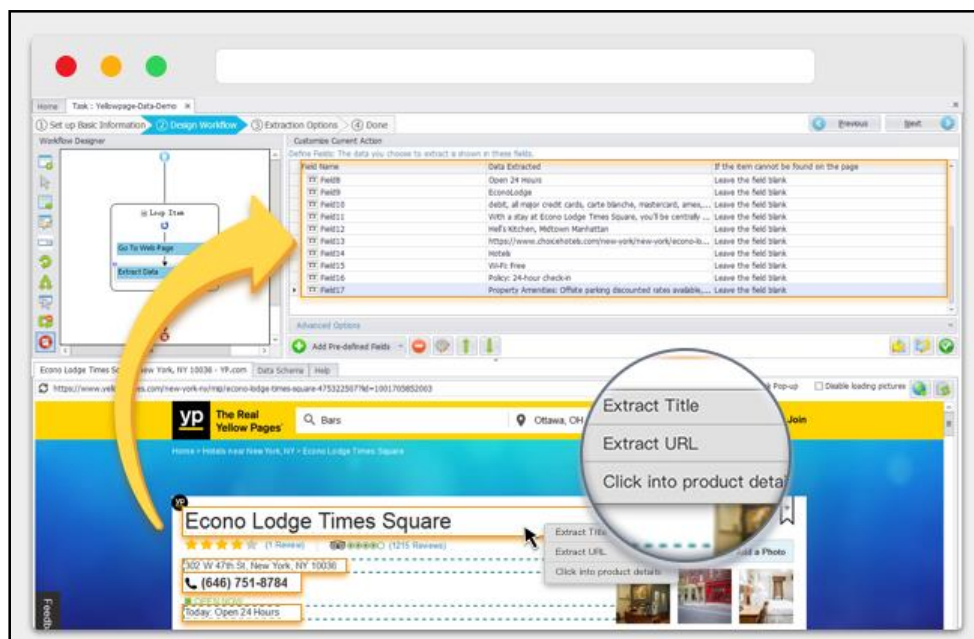


Figura 3.7: Interfaz gráfica Octoparse.

- **RapidMiner**



Esta herramienta hace a los equipos de ciencia de datos más productivos a través de una plataforma de código abierto para preparación de datos, machine learning y modelos de negocio.

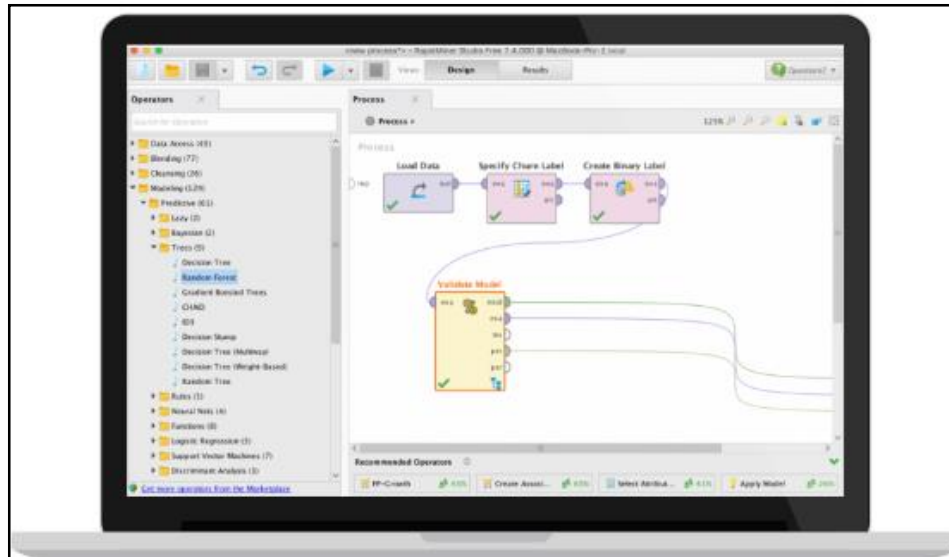


Figura 3.8: Interfaz gráfica Rapid Miner.

- **Knime**



Sus productos son muy variados extendiéndose desde plataformas de innovación a partir de los datos que ayudan a mostrar el potencial oculto en los datos y poder así posteriormente predecir nuevos en el futuro hasta la búsqueda de la mejora en la productividad de tu negocio. Todo ello es llevado a cabo a través de un software especializado en servicios en la nube, mejora de la productividad, herramientas Big Data, servicios clúster y herramientas de Social Media dedicadas al Análisis de Sentimientos entre otros.

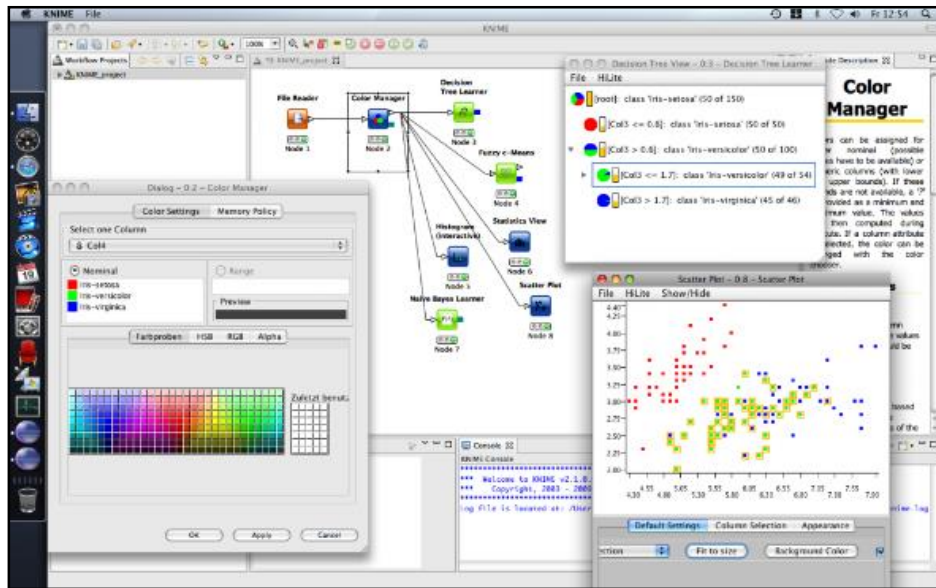


Figura 3.9: Interfaz gráfica Knime.

- Orange Data Mining



Este software libre provee de aprendizaje automático y herramientas de Data Mining. Entre sus características principales podemos encontrar el uso de herramientas de programación de tipo visual *front-end* para la exploración de datos y visualización de sus resultados.

Con Orange podemos crear nuestros propios flujos de trabajo interactivos para el análisis y visualización de datos con mayor amplitud para después poder adaptar esta herramienta a las necesidades que exija cada empresa. La información es visualizable en diferentes tipos de formatos como por ejemplo diagramas de dispersión, árboles de decisión o redes neuronales proporcionando hasta más de 100 widgets en esta herramienta personalizable.

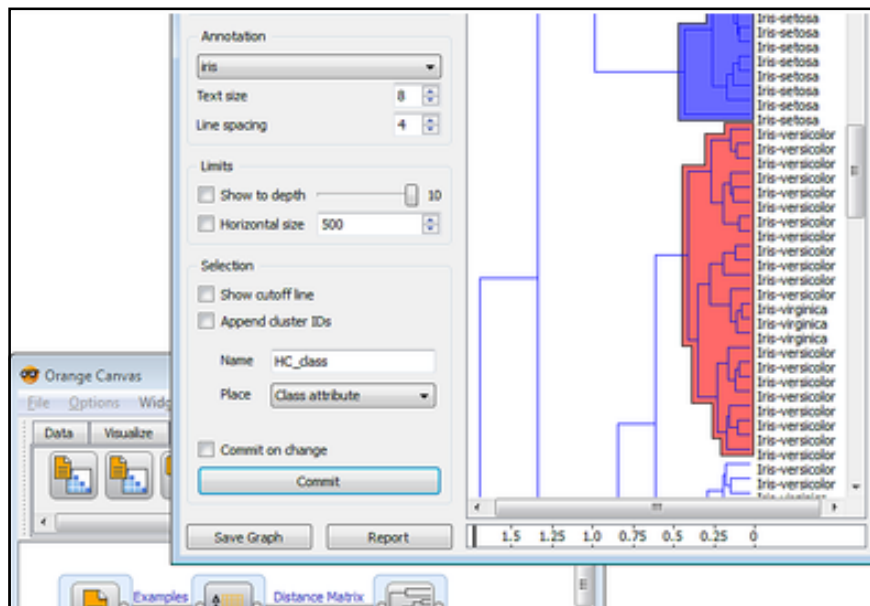


Figura 3.10: Interfaz gráfica Orange Data Mining.

4. Estructura del software desarrollado

En este apartado se describirá paso a paso las diferentes etapas con las que contará la herramienta desarrollada, desde la creación de nuestra aplicación para poder recolectar los datos necesarios, hasta su posterior análisis por medio de las técnicas de minería de datos dentro de la IDE de Rstudio.

4.1. Entorno de desarrollo: Rstudio

Rstudio⁷ es un entorno de desarrollo integrado (IDE) para el lenguaje de programación estadístico R, en el que poder visualizar de una mejor manera el código que crearemos para nuestra herramienta. Rstudio puede correr en diferentes sistemas operativos ya sea Windows, Mac OS X o Linux como aplicación de escritorio, la cual es libre de descargar desde la red.

Cuando Rstudio es cargado por primera vez podemos ver algo parecido a la Figura x. Existen cuatro paneles en los que encontraremos (De izqda. a dcha., de arriba abajo) el script desarrollado, la consola con los resultados de ejecutar el código y un último apartado en el que podremos obtener información acerca de librerías, cargar paquetes y obtener acceso a información sobre funciones así como la posibilidad de visualizar las diferentes opciones gráficas que vayan apareciendo.

⁷ <http://www.rstudio.com/>

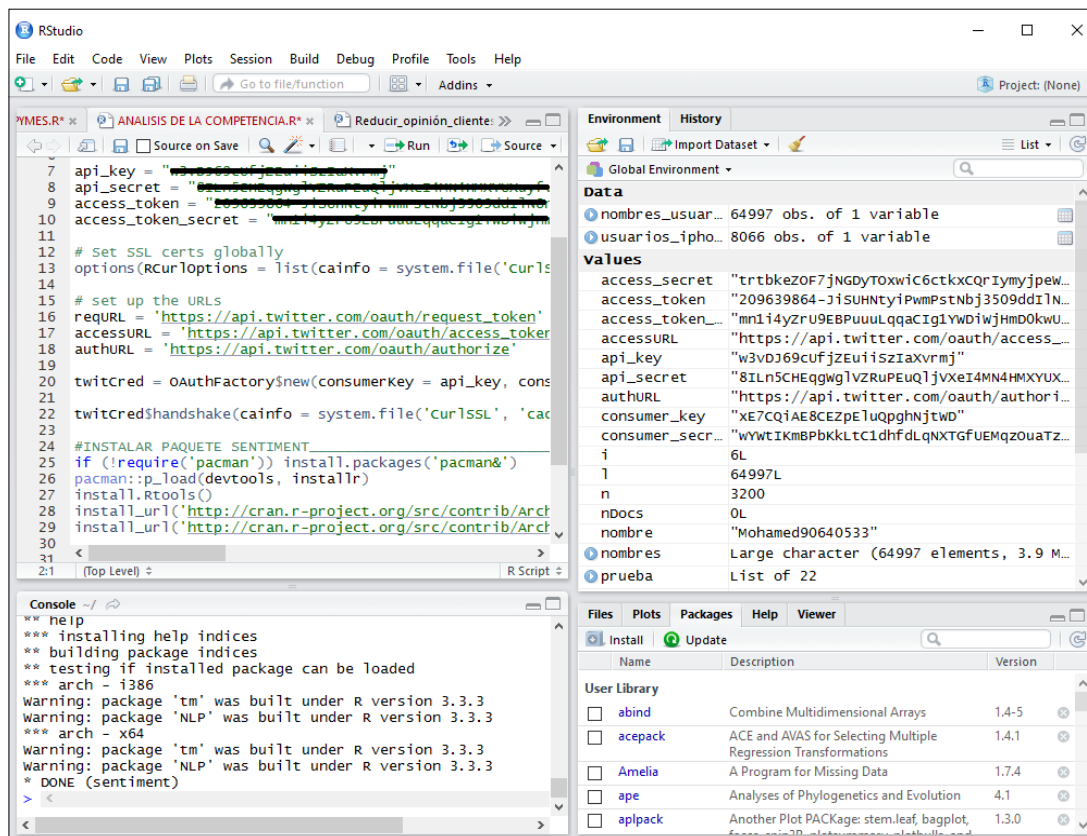


Figura 4.1: IDE de Rstudio.

4.2. Registro y Creación de Twitter App

Para comenzar, hay que dirigirse a la página de desarrolladores de Twitter⁸, en la que accederemos con el usuario y contraseña de nuestra cuenta de Twitter.

Una vez hayamos entrado, nos dirigiremos al menú situado en la parte superior donde aparecerá el icono de nuestro perfil de usuario para que al pasar por encima se nos despliegue otro menú con la opción “*My aplicaciones*”. Al clicar en esta opción nos dirigirá a la página de apps para desarrolladores⁹, desde la cual podremos crear nuestra aplicación pulsando en el botón llamado “*Create a new application*”

Opciones de configuración al crear la aplicación en Twitter. A continuación, se nos mostrará una serie de campos a rellenar para poder crear la aplicación. Esto es también importante para poder conectar posteriormente

⁸ <https://dev.twitter.com/>

⁹ <https://dev.twitter.com/apps>

ASP_BinteakMusic Test OAuth

[Details](#) [Settings](#) [Keys and Access Tokens](#) [Permissions](#)

Application Details

Name *

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Callback URL

Where should we return after successfully authenticating? [OAuth 1.0a](#) applications should explicitly specify their `oauth_callback` URL on the request token step, regardless of the value given here. To restrict your application from using callbacks, leave this field blank.

Privacy Policy URL

The URL for your application or service's privacy policy. The URL will be shared with users authorizing this application.

Terms of Service URL

Figura 4.2: Formulario de Twitter App.

Los campos que encontramos por orden son:

- **Name:** Introduciremos el nombre que queramos que tenga nuestra aplicación. Cabe aclarar que puede contener el nombre que queramos mientras que no contenga la palabra Twitter.
- **Description:** Nos da la opción de escribir una breve descripción acerca de la funcionalidad de nuestra aplicación.
- **Website:** Aquí pondremos el sitio web de la aplicación. En el caso de no necesitarla, podemos escribir una y en el futuro podremos cambiarla.
- **Callback URL:** Para desarrollar una aplicación web deberemos tener en cuenta este campo ya que es una URL a la que trae de vuelta cuando se autentifica de manera correcta en la aplicación. En el caso nuestro no es necesario y se puede dejar en blanco.

Después de rellenar correctamente los apartados anteriormente descritos procederemos a marcar la casilla de condiciones de uso y proceder a verificar que somos humanos a través de un código captcha. Posteriormente pulsaremos el botón que sigue bajo el nombre de “*Create your Twitter Application*” con la que habremos terminado el proceso de creación de nuestra aplicación Twitter.

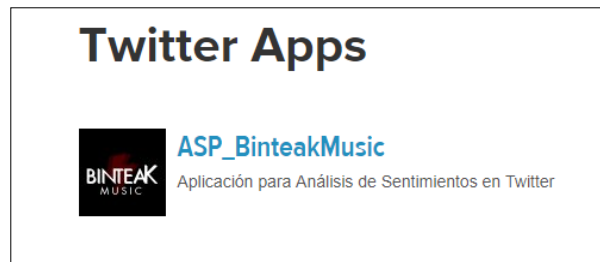


Figura 4.3: Twitter App creada.

Al finalizar estos pasos ya habremos obtenido una clave de usuario y la contraseña correspondiente. Estos valores son denominados por Twitter bajo el nombre de “Consumer Key” y “Consumer Secret” respectivamente, los cuales serán de vital importancia en el proceso posterior necesario para conectar la API de Twitter con RStudio, ya que nos proporcionarán acceso a nuestra cuenta en Twitter a través de la aplicación, es por ello por lo que nunca se deberá compartir estos datos con terceros.

En la pestaña principal podemos observar el nivel de acceso del que disponemos, pudiendo modificar nuestros permisos en cualquier momento. Además, podremos añadir información sobre nuestra organización, como se detalla en la siguiente figura:

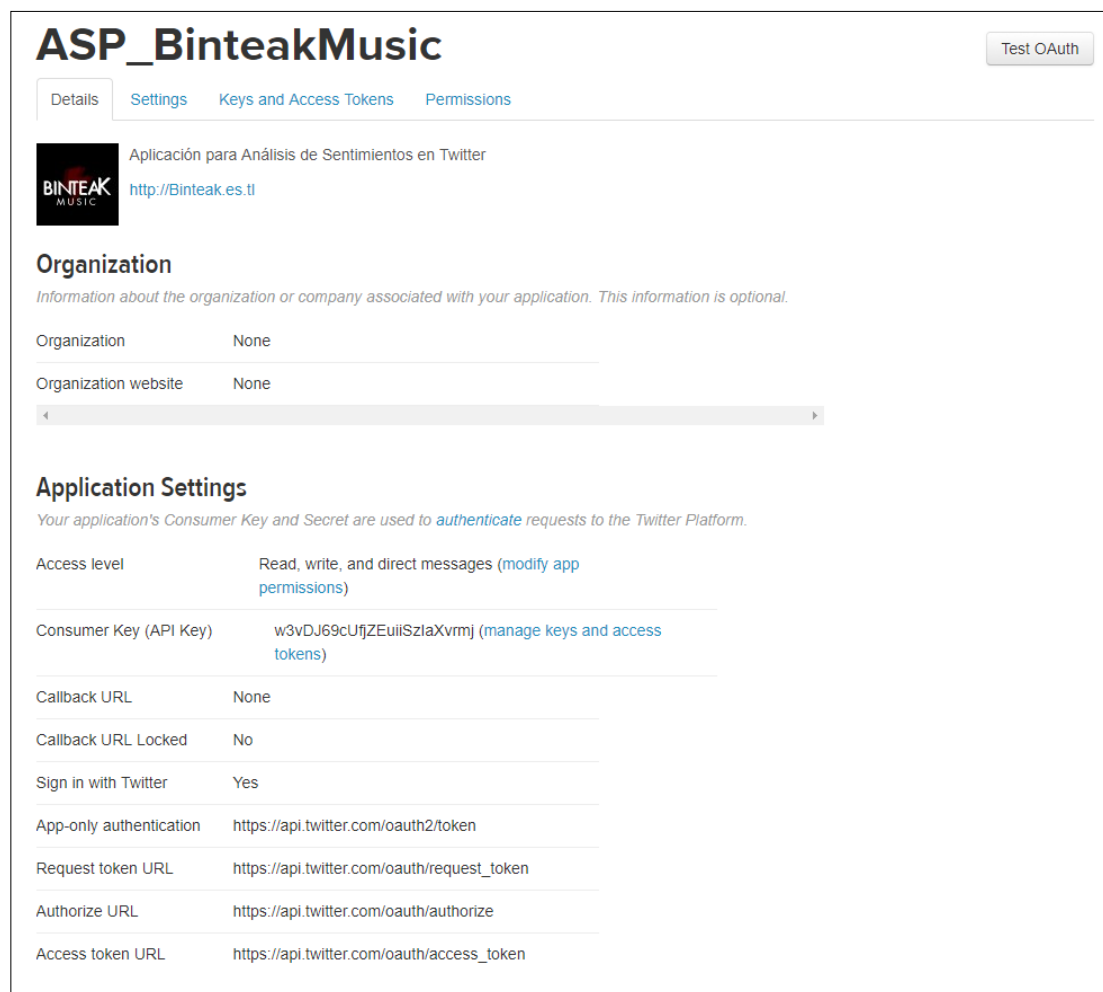
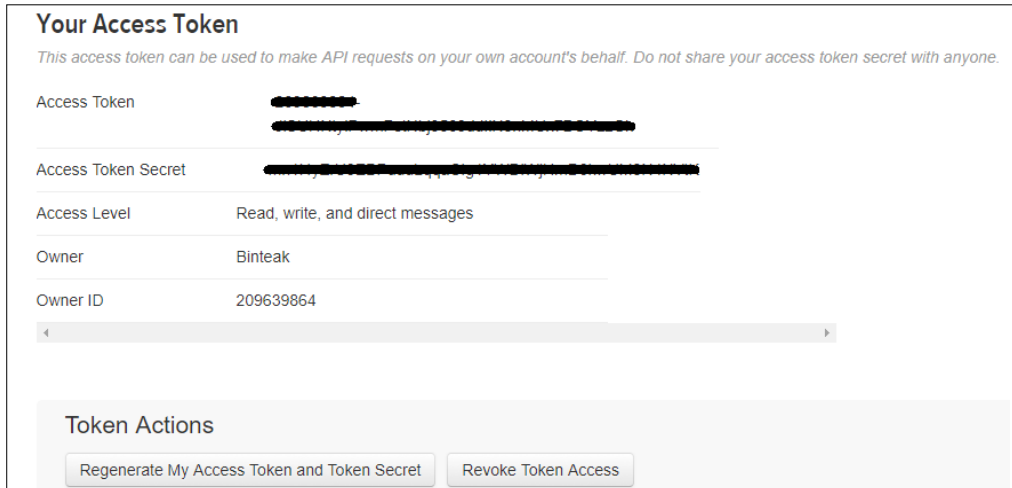


Figura 4.4: Detalles de Twitter App.

A continuación, nos iremos a la pestaña de “*Keys and Access Tokens*” para conseguir nuestras credenciales de acceso, necesarias para el proceso de autorización y autenticación a la API de Twitter desde Rstudio. Al final de este apartado encontraremos la opción de Token Actions, donde pulsaremos en el botón “*Create my Access Tokens*”. Posteriormente este apartado pasará a obtener la característica de poder regenerar nuestras claves anteriormente citadas en el caso de ser necesario.



Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	[REDACTED]
Access Token Secret	[REDACTED]
Access Level	Read, write, and direct messages
Owner	Binteak
Owner ID	209639864

Token Actions

[Regenerate My Access Token and Token Secret](#) [Revoke Token Access](#)

Figura 4.5: Claves de acceso a API de Twitter.

Con los pasos anteriormente descritos ya tenemos completamente configurada nuestra aplicación que nos permitirá obtener la información de Twitter para su posterior análisis dentro de nuestro entorno de desarrollo RStudio.

4.3. Proceso de Autorización y Autenticación para API Twitter

El siguiente paso es obtener permiso para utilizar la API de Twitter desde Rstudio utilizando nuestros tokens generados como cliente. De esta manera se nos desplegará una ventana en nuestro navegador para proceder a la autorización de nuestra aplicación desde Rstudio.



¿Autorizas a ASP_BinteakMusic a utilizar tu cuenta?

[Autorizar la aplicación](#) [Cancelar](#)

Esta aplicación podrá:

- Leer Tweets de tu cronología.
- Ver a quién sigues y seguir a nuevas personas.
- Actualizar tu perfil.
- Publicar Tweets por ti.
- Acceder a tus mensajes directos.

No podrá:

- Ver tu dirección de correo electrónico.
- Ver tu contraseña de Twitter.

ASP_BinteakMusic
Binteak.es.tl
Aplicación para Análisis de Sentimientos en Twitter

Figura 4.6: Autorización Twitter App.

Al aceptar el proceso de autorización se generará un código PIN por el cual se concede el acceso a la app creada en Twitter, el cual tendremos que alojar en la última línea de código generada en nuestra consola dentro de Rstudio:

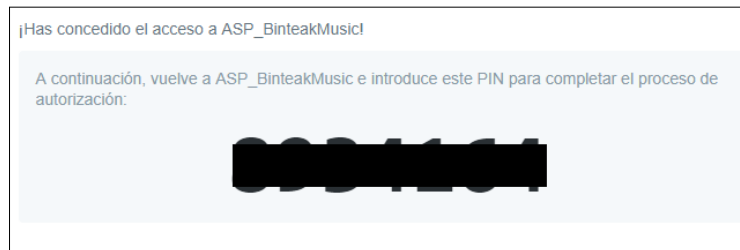


Figura 4.7: Código PIN de acceso.

```
> twitcred$handshake(cainfo = system.file('CurlSSL', 'cacert.pem', package = 'RCurl'))
To enable the connection, please direct your web browser to:
https://api.twitter.com/oauth/authorize?oauth_token=BaS2_QAAAAAaWS0NAAABX2lupQo
when complete, record the PIN given to you and provide it here: [redacted]
```

Figura 4.8: Establecimiento de conexión en Rstudio.

A partir de este momento Rstudio está autorizado a llevar a cabo búsquedas en Twitter.

4.4. Extracción de la información y almacenamiento. TwitterR

A la hora de extraer la información de Twitter tendremos que cargar la librería TwitterR. TwitterR¹⁰ es un paquete de R que nos permite acceder a la API de Twitter. La mayor parte de las funcionalidades de la API son compatibles aunque se pueden encontrar diferentes sesgos en base a la información a extraer.

El proceso de instalación en equipo de la librería twitterR es el siguiente:

```
> install.packages("twitter")
```

Posteriormente accederemos a ella llamando a librería:

```
> library(twitter)
```

Existen diferentes funciones dentro de este paquete del cual utilizaremos UserTimeline. UserTimeline nos proporciona un número de Tweets del TimeLine de un usuario específico en Twitter. El número de tweets puede variar en base al valor de parámetro que le asignemos. Este número puede moverse desde 1 hasta un máximo de 3200 por cuenta de usuario en Twitter.

Para nuestro estudio se eligen 3200 tweets por cuenta aunque este número será algo menor en cantidad para algunas de ellas.

¹⁰ <https://www.rdocumentation.org/packages/twitteR/versions/1.1.9>

11	Entra en... https://t.co/9fxzTVIYfp ,FALSE,14,NA,2017-12-22 18:59:05,TRUE,NA,"944281154206814208",NA,"
12	5,"Si vienes a #ViñaRock desde Bilbao, Burgos, Madrid, San Sebastián, Valencia o Vitoria hazte ya con una de las últim... https://t.co/nAEMcJgkWW ",FALSE,4,NA
13	6,"Empezamos el viernes escuchando algunas de las canciones que sonarán en #ViñaRock2018 e imaginándonos ya allí... https://t.co/7pdEYroDbd ",FALSE,16,NA
14	7,"En #ViñaRock no suele hacer mucho sol... pero cuando sale lo disfrutamos al máximo.
15	Imagen de @davidalgas93
16	
17	Abonos... https://t.co/MHg8Ozj0Uf ",FALSE,26,NA,2017-12-21 19:00:37,TRUE,NA,"943919151000838144",NA,"
18	8,"Somos los @SegisOficial y vamos al puto lio" - Os esperamos en Viña Rock
19	Si todavía no tienes tu abono no esperes... https://t.co/LXx6TuBhGm ",FALSE,10,NA,2017-12-21 13:04:10,TRUE,NA,"943829447077191681",NA,"<a href=""https://
20	9,"Buenos días #Viñarockeros!
21	https://t.co/5GLrQHAzb https://t.co/JSnZNxFhhm ",FALSE,20,NA,2017-12-21 09:59:31,FALSE,NA,"943782978987216896",NA,"<a href=""https://postcron.com"" r
22	10,"Así de felices estamos sabiendo que por solo 3€ + gastos nos podemos duchar con agua caliente de forma ilimitada d... https://t.co/YbGBwnEbLR ",FALSE,5,
23	11,"@LosNarco Los ganadores del sorteo de 1 CD firmado por @LosNarco son @Mercaque y @ElenaFuckYou (escribidnos un mens... https://t.co/emdh6aTish ",
24	12,"Queremos que la música empiece a sonar y estar en #ViñaRock dándolo todo
25	
26	Abonos a la venta en... https://t.co/mPHXxR3dJn ",FALSE,17,NA,2017-12-20 09:13:48,TRUE,NA,"943409087299096576",NA,"<a href=""http://twitter.com"" rel=""r
27	13,"Así fue #ViñaRock 2017para @juannn.gc. ¡Nos volveremos a ver este año!
28	
29	No te quedes sin tu entrada en... https://t.co/6oKHPEjhSU ",FALSE,13,NA,2017-12-19 16:39:15,TRUE,NA,"943158800718196736",NA,"<a href=""http://twitter.com
30	14,"@GoitiAizpea Ya queda menos!! <ed><U+00A0><U+00BE><ed><U+00B4><U+009F>",FALSE,3,"GoitiAizpea",2017-12-19 12:27:34,FALSE,"943091814810177536"
31	15,"@saramarug @CandelaGleek _La verdad es que mucho sol no hizo... pero nos lo pasamos muy bien <ed><U+00A0><U+00BD><ed><U+00B8><U+
32	16,"@VicHacheR ""Hipotekt"" de @ofertaespecial.
33	Muy recomendable este grupo <ed><U+00A0><U+00BE><ed><U+00B4><U+009F>",FALSE,5,"VicHacheR",2017-12-19 12:01:16,FALSE,"943077632207458304", "94308
34	17,"Llegamos a la recta final de 2017 con los recuerdos de la pasada edición de #ViñaRock muy presentes y con el conven... https://t.co/khMAJf61ze ",FALSE,44,
35	18,"Hoy nos hemos levantado con las pilas cargadas recordando el concierto de @SonsofAguirre la pasada edición del fest... https://t.co/DWDQWJ4N6g ",FALSE
36	19,"@SegisOficial Allí os esperamos <ed><U+00A0><U+00BE><ed><U+00B4><U+009F><ed><U+00A0><U+00BE><ed><U+00B4><U+009F>",FALSE,4,"SegisOficial",2
37	20,"@irelose)",FALSE,11,"irelosa30",2017-12-18 16:15:15,FALSE,"942789001194467333", "942790370739609600", "129277967", "<a href=""http://twitter.com"" rel=
38	21,"El año pasado ya pudimos ver de qué son capaces @Russkaja . Este año será un concierto que seguro que no nos vamos... https://t.co/MS0N2LkSN ",FALSE,
39	22,"@2ds_ni)",FALSE,1,"2ds_nintendo",2017-12-18 16:00:21,FALSE,"942768500292161536", "942786620364861440", "2221924396", "<a href=""http://twitter.com""
40	23,"Estas son las vistas que queremos tener por la mañana.
41	¡Buenos días #Viñarockeros! - Imagen de @latasand
42	
43	Abonos e... https://t.co/vGHWxRpoO ",FALSE,36,NA,2017-12-18 09:29:00,TRUE,NA,"942688136299245568",NA,"<a href=""https://postcron.com"" rel=""nofollow
44	24,"@Jose17)",FALSE,1,"Jose17PM",2017-12-18 08:33:29,FALSE,"942628571721461761", "942674162614788096", "586424008", "<a href=""http://twitter.com"" rel=
45	25,"#ViñaRock nos espera el 28, 29 y 30 de abril.
46	Hazte hoy con tu abono al precio actual en https://t.co/upBDC2q2NG https://t.co/Bkn7AHBgxU ",FALSE,21,NA,2017-12-17 20:29:01,FALSE,NA,"9424918451117957

Figura 4.10: Tweets sin transformar.

Las estructuras de código utilizadas para su tratamiento son descritas a continuación:

Estructura del código utilizado	Descripción de la transformación en la información
<code>gsub, "(RT via)((?:\\b\\W* @\\w+)+)"</code>	Remover RT's
<code>gsub, "http[^:blank:]]+"</code>	Remover links anexos al tweet
<code>gsub, "[[:punct:]]"</code>	Remover caracteres especiales
<code>library(stringi), stri_trans_general(, "Latin-ASCII")</code>	Remover tildes
<code>gsub, "[[:digit:]]"</code>	Remover dígitos numéricos
<code>as.list(tolower())</code>	Transformar todo el texto a minúsculas
<code>gsub, "^ ", " \$"</code>	Remover espacios en blanco en texto al inicio y al final de los strings
<code>gsub, "[\\t]{2,}"</code>	Remover tabulación
<code>duplicated(), unique()</code>	Encontrar y eliminar filas duplicadas
<code>iconv(, "latin1", "ASCII", sub="")</code>	Remover elementos extraños
<code>gsub, "\\n\\n", "\\n", ":", "/"</code>	Otros

Figura 4.11: Transformaciones en tweets.

Para obtener una visión clara y concisa de lo anteriormente descrito y una vez aplicadas todas las transformaciones se muestra a continuación como quedaría el texto de todos los tweets después de su tratamiento.

12	11,"losnarco los ganadores del sorteo decd firmado por losnarco son mercaque y elenafuckyou escribidnos un mens..."
13	12,"queremos que la musica empiece a sonar y estar en vinarock dandolo todoabonos a la venta en..."
14	13,"asi fue vinarock para juannngc nos volveremos a ver este anono te quedas sin tu entrada en..."
15	14,"goitiaizpea ya queda menos"
16	15,"samarug candelagleek ourwreck la verdad es que mucho sol no hizo pero nos lo pasamos muy bien"
17	16,"vichacher hipotekt de ofertaespecial muy recomendable este grupo"
18	17,"llegamos a la recta final decon los recuerdos de la pasada edicion de vinarock muy presentes y con el conven..."
19	18,"hoy nos hemos levantado con las pilas cargadas recordando el concierto de sonsofaguirre la pasada edicion del fest..."
20	19,"segisoficial alli os esperamos"
21	20,"irelosa en el vina se hacen amigos rapido tranquila"
22	21,"el ano pasado ya pudimos ver de que son capaces russkajaeste ano sera un concierto que seguro que no nos vamos..."
23	22,"dsnintendo claro puedes escribimos a infovinarockcom con tu duda"
24	23,"estas son las vistas que queremos tener por la manana buenos dias vinarockerosimagen de latasandabonos e..."
25	24,"josepm buenas jose puedes escribir a infodefestivalescom con tu solicitud"
26	25,"vinarock nos espera elyde abrilhazte hoy con tu abono al precio actual en"
27	26,"un segundo puede cambiarlo todo somos los duenos de la eternidad desequilibrados por no hallar el modo de no..."
28	27,"no te pierdas la xxiii edicion de vinarock consigue hoy uno de los ultimos abonos al precio actual en..."
29	28,"dsnintendo vinarock se podran reservar pero el precio puede que sea superior saludos"
30	29,"dias nos separan de volver a vinarockimagen compartida por joacojfp en instagramabonos a la venta en..."
31	30,"ahorra pasta comprando tus tuments moneda para comprar bebida dentro del festival de forma anticipada por cada ..."
32	31,"al mas puro estilo vinarockeroimagen de josepr compartida en instagram hazte hoy con tu abono en..."
33	32,"este ano podremos escuchar colaboraciones como este temazo de sfdkoficial con fyahbwoy y gordomaster blondapet..."
34	33,"el ano que viene esos seremos nosotros abonos a la venta en"
35	34,"todo lo malo del tiempo lo tuvo de bueno la companiarecuerdo de vinarock compartido por..."
36	35,"vamos vinarockeros que ya es viernesultimos abonos al precio actual en"
37	36,"hoy sale a la venta espichufrenia el ultimo trabajo de losnarco y vamos a sortearcds firmados para que nos va..."
38	37,"preparaos para escuchar lo nuevo de reincidentesofi en vinarockabonos a la venta en"
39	38,"skalejandro la foto sera de antes de que empezara la lluvia"
40	39,"amongil losmareaooficial seguro que seria una puta pasada pero todavia no estan dando conciertos esperamos que..."
41	40,"esta foto que comparte lauraspringsteen en instagram representa lo que es vinarock"
42	41,"contando los dias para volver a vinarockabonos en"
43	42,"billygrmg windows"
44	43,"adala amp the same song band es uno de esos descubrimientos que puedes hacer este ano en vinarockabonos en..."
45	44,"deja tu coche en casa y vente a vina rock en los autobuses de defestivales sin riesgo de cancelacion ultimas plaz..."
46	45,"sfdkoficial vuelve a vinarock lo que nos espera este ano en villarrobledoabonos en"
47	46,"losnarco nos espera este ano en vinarockabonos en"

Figura 4.12: Tweets transformados.

- **Corpus**

Después de obtener el texto de una manera limpia y adecuada, se procede al tratamiento del Corpus. En el ámbito de la lingüística computacional se llama Corpus¹¹ a aquel conjunto o colección que contenga más de un texto.

```
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 5618
```

Figura 4.13: Corpus.

Esta estructura nos permitirá realizar posteriores procesos tanto en la limpieza de la información recogida como posteriormente el uso para su análisis.

- **Stopwords**

Las Stopwords¹² son palabras recurrentes dentro de un idioma, conectores, determinantes, preposiciones y otras palabras que no aportan información interesante. Estas palabras dependerán del idioma en el que se encuentre la información extraída. En el caso a estudio se tomaron en cuenta las que corresponden a tres idiomas, el inglés, el español y el catalán ya que algunas de estas cuentas se dirigen a sus seguidores en uno, dos o tres de estos idiomas.

¹¹ <http://elies.rediris.es/elies18/23.html>

¹² <https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html>

Como las Stopwords venían con tildes, tuvieron que eliminarse, así como la letra ñ. Además, posteriormente se crearon archivos en formato Excel que fueron cargados y aplicados sobre el texto en R, incluyendo también palabras que se consideraron sin necesario interés.

4.6. Términos frecuentes y WordCloud

A la hora de mostrar las palabras y términos con mayor aparición en el Corpus debemos reunificar la información. Contabilizando uno por uno los términos con mayor aparición se creará una tabla en la que poder comprobar qué Menciones, Hastags y Palabras en general aparecen con mayor frecuencia.

La partición se llevó a cabo para Menciones, Hastags y Palabras debido a que si se analizaban juntas podría perderse la información que proporcionan al mostrarlas con las demás palabras.

Podremos comprobar si parte del texto analizado presenta algún tipo de patrón o existe alguna tendencia en los diferentes dataframes.

- **WordCloud**

Un WordCloud¹³ o nube de palabras es una representación de manera gráfica de todas aquellas palabras presentes en un conjunto de texto en los que visualmente se especifica en base al tamaño de las palabras mostradas la frecuencia de aparición que poseen frente a las demás.

De esta manera, podremos comprobar rápidamente si existen patrones en nuestros datos y ver si la posibilidad de asociaciones en el texto.



Figura 4.14: Imagen WordCloud.

¹³ <https://datascienceplus.com/building-wordclouds-in-r/>

5. Ejemplo de uso: Binteak Music

Binteak Music, una empresa que se dedica al mundo de la música redactando noticias en base a información relevante a modo de revista online sobre reviews, festivales y artistas en géneros como rock, música alternativa, indie nacional e internacional.



Figura 5.1: Visualización de tweet redactado.

El objetivo principal que busca esta pyme es posicionarse en el mercado abriéndose un hueco entre consolidadas revistas online viendo necesario saber a tiempo real qué noticias o tendencias son relevantes en el ámbito musical. Es por ello que busca un modelo que le permita obtener conclusiones relevantes de la red que conforman diferentes cuentas creadoras de información y que a su vez por sí sola pueda generar información relevante para los usuarios que tengan esa demanda y atraerlos a sus propios servicios.

Esta información creada al instante y de volumen masivo será de gran interés ya que permitirá saber en todo momento artistas y tendencias a tiempo real en estas redes sociales.

La manera en la que se busca conseguir el objetivo planteado es conseguir analizar la información que tuitean tanto empresas competidoras como sus usuarios activos, reunirla, preprocesarla, sacar conclusiones y crear información en blog sobre temas o artistas tendencia que será transmitida a las redes en forma de tweets para que los usuarios puedan acceder desde esta red social.

La pyme conseguirá crear un volumen de tráfico de visitas de usuarios en esta red social que demanden información similar o parecida hacia su blog, para a largo plazo tener la posibilidad en base a su contenido de atraerlos para futuras consultas y fidelizarlos como lectores activos.

En un primer momento se estudió qué empresas del mismo sector eran relevantes en cuanto a géneros musicales y de igual manera se hizo para las cuentas de los festivales de referencia que incorporaban en sus carteles a artistas de estos estilos.

El conjunto de cuentas en Twitter que se extrajeron fueron 31 en dos apartados, festivales y revistas online:

Revistas online	Festivales
Rock Sin Subtítulos	Viñarock
Cultura Rock Música	Primavera Sound
Dod Magazine	Sonar Festival
Mondo Sonoro	Bilbao BBK Live
El Quinto Beatle	Cruilla Fest
Rocking News	Al Rumbo Festival
El Mundo de Tulsa	Festival Internacional de Benicassim
Muzikalia	Low Festival
Indienauta	Arenal Sound
	Dream Beach Festival
	Mad Cool Festival
	Azkena Rock Festival
	Contempopránea
	Santander Music Festival
	DCODE Festival
	Download Festival Madrid
	Tomavistas Festival
	San San Festival
	Festival de Les Arts
	Vida Festival
	Sonorama Ribera

Figura 5.2: Conjunto de cuentas de Twitter.

En total el conjunto de la información recogida abarcó un total de 5899 tweets pertenecientes al total de las 31 cuentas de Twitter descritas anteriormente.

Posteriormente se procede a la limpieza y preprocesamiento de la información y partición en tres tablas de frecuencias, una para RT y menciones, otra para hastags y la última para el total de la información relevante disponible.

Esta partición se hace debido a que los hastags y las menciones obtendrán diferentes frecuencias por lo tanto al unirlas con el total de palabras se perderán así como la información relevante que puedan contener.

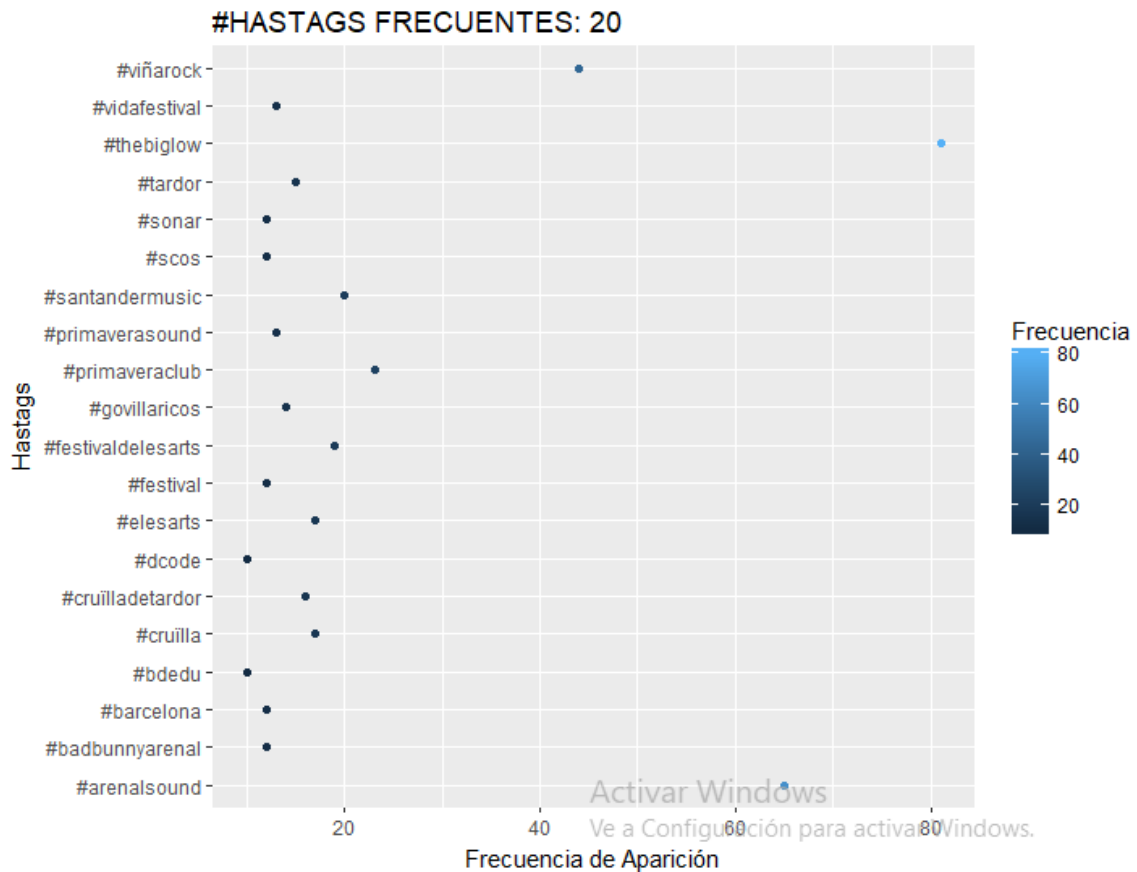


Figura 5.3: Gráfico de frecuencias para Hastags.

Para los Hastags, obtenemos hasta 20 hastags más frecuentes. Vemos que los que obtienen una mayor aparición en frecuencia son #thebiglow, #arenalsound y #viñarock. Esto es debido a que en esta época es cuando mayor actividad tienen aquellas cuentas dedicadas a festivales y se incrementa la capacidad de interacción con sus usuarios sobre estos temas. Suelen crear hastags en los que van indicando información relevante en base al propio festival o sobre los artistas que irán en sus carteles además de los usuarios que comentan en base al hashtag.

Entre los principales hastags aparece un artista, #badbunnyarenal, por lo tanto podría ser una publicación por parte de Binteak Music interesante a tratar ya que además está relacionado con el segundo festival en orden de aparición por frecuencia.

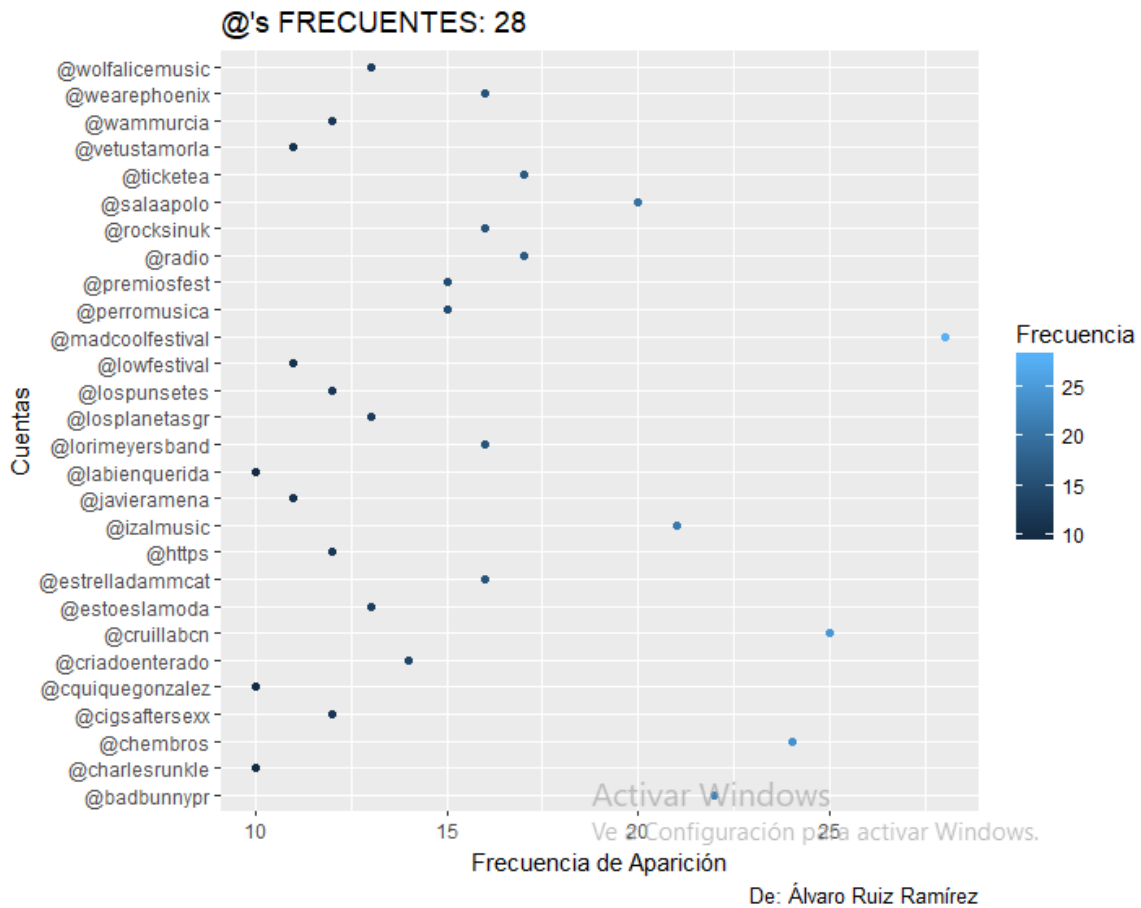


Figura 5.4: Gráfico de frecuencias para menciones y cuentas.

En el apartado de Menciones podemos encontrar hasta 28 con mayor frecuencia. Las menciones de cuentas que generan una mayor frecuencia de aparición son @madcoolfestival, @cruillabcn, @chembros, @badbunnypr, @izalmusic y @salaapolo.

Encontramos de nuevo entre los primeros puestos menciones a cuentas dedicadas a festivales, a 3 artistas musicales (@chembros, @badbunnypr e @izalmusic), y a una sala de conciertos (@salaapolo) en este orden.



Figura 5.5: WordCloud con las palabras más frecuentes.

De una forma más visual podemos comprobar que las palabras que aparecen con mayor frecuencia en el total de los 5899 tweets son nuevo, disco, concierto, anuncia, Madrid, videoclip y estrena, seguidos en menor medida por single, abonos, festival, Barcelona etc.

Posteriormente se hace una diferenciación por temas a tratar en base a las palabras, temas como festivales y encontramos dos artistas entre las primeras menciones.

Al final, después de hacer una búsqueda rápida en Twitter a partir de los artistas, comprobamos que palabras, hastags y menciones tienen relación de la siguiente manera:

[nuevo,disco] corresponde con @izalmusic
[concierto,anuncia,Madrid] está relacionado con @badbunnypr
#thebiglow, #arenalsound, #viñarock con @madcoolfestival
#arenalsound #badbunnyarenal con @badbunnypr
@salaapolo con barcelona
@chembros con #lowfestival

Por lo tanto se crearán 5 tweets que llevarán un enlace al blog de Binteak Music donde se desarrollarán estas noticias musicales.

Los tweets quedarán estructurados de la siguiente manera:



Figura 5.6: Tweets redactados en base a la información relevante.

6. Conclusiones y Trabajo futuro

En este apartado se desarrollan aquellas conclusiones que tras la realización de este trabajo se han ido generado, además de las líneas de un posible trabajo futuro para la extensión del mismo.

6.1. Conclusiones

La importancia de métodos de análisis de la información por parte de las pymes aporta actualmente una ventaja competitiva frente a grandes empresas. Ante un mundo globalizado en el que se genera información constante de manera instantánea y la dificultad que se presenta en base a factores como el coste por servicios para conseguir posicionarse en el mercado o la poca visibilidad con la que se parte desde el inicio de creación, cada vez más el uso de nuevas alternativas va tomando protagonismo en este sentido y estas pequeñas y medianas empresas pueden conseguir grandes resultados sin un gran desembolso económico.

Es por ello que en este trabajo se ha optado por crear una solución para empresas que tenían esta necesidad, aplicando técnicas de Minería de Datos y en especial la Minería de Textos para extraer conclusiones relevantes que desde un principio aportarán valor y las podrá hacer competir.

Ha sido interesante poder crear código para crear un hipotético caso real en el que conseguir afianzar diferentes conocimientos, sobre todo aquellos relacionado con el desarrollo de conocimientos adquiridos durante el transcurso del grado en lo relacionado al lenguaje de programación R. Por otro lado, la limpieza de la información, reestructurización de los datos, carga de datos, creación de gráficos, el uso de diferentes funciones, la interfaz de Rstudio y su conexión con entornos externos no vistos antes han sido de gran interés.

Se han estudiado todos estos conocimientos en aplicación ante un problema real en el que se ha buscado la consecución de los objetivos planteados de los que se partía inicialmente, con datos generados al momento y en una mayor cantidad.

El código generado permite que las empresas puedan investigar los datos generados en Redes Sociales para su análisis y que a partir de ellos puedan lograr el objetivo de conseguir mayor visibilidad y beneficios económicos, basándose en la extracción de conocimiento que no se tenía en cuenta hasta el momento.

6.2. Líneas de trabajo futuro

Si bien la lo desarrollado sirve a modo de consecución de los objetivos de este trabajo, creo que aún se puede extender con el tiempo y desarrollarse de una manera más amplia, ya sea añadiendo nuevas funcionalidades o ámbitos de actuación y desarrollo.

Para mejorar en un futuro cercano este proyecto, se procede a la creación de un listado que recoge en breves líneas aquellos aspectos que podrían aportar nuevas funcionalidades a lo ya descrito y expuesto y que podrán ser interesantes y útiles en posteriores trabajos:

- Partición en subsets de la información para diferentes cuentas
- Cruce de información entre cuentas por pares
- Agrupamiento por clusterización
- Aumento de la información a extraer
- Seguimiento por días de tendencias
- Mejora del código planteado
- Conexión y extracción con otras fuentes de información como Facebook

Bibliografía y Referencias

- (SDE). (2008). *Incremento a la productividad*. Obtenido de http://www.sde.guanajuato.gob.mx/index.php?option=com_content&task=228&Itemid=282
- Hernández, F. (2007). *Propuesta de un Modelo Estratégico de Globalización económica*. Obtenido de <http://www.caei.com.ar/ebooks/ebook27.pdf>
- Orallo, J. H. (2004). *Introducción a la minería de datos*. Editorial Alhambra S. A. (SP).
- Zhao, Y. (2015). *R and Data Mining. Examples and Case Studies*. Elsevier. Obtenido de <http://www2.rdatamining.com/uploads/5/7/1/3/57136767/rdatamining-book.pdf>
- Evolución Historia de la Minería de Datos
<http://www.dataprix.com/171-evoluci-n-historia-miner-datos>
- [1] Datamining (Minería de datos)
http://www.sinnexus.com/business_intelligence/datamining.aspx
- [2] KDD: ¿Qué es el Knowledge Discovery in Databases o KDD?
<http://mnrvai.io/kdd-platform.html>
- [3] Twitter web
<https://twitter.com/?lang=es>
- [4] Search API
<https://dev.twitter.com/rest/public/search>
- [5] Rest API
<https://dev.twitter.com/rest/public>
- [6] Streaming API
<https://dev.twitter.com/streaming/overview>
- [7] Rstudio programming language
<http://www.rstudio.com/>

[8] Twitter Developer

<https://dev.twitter.com/>

[9] Twitter Apps

<https://dev.twitter.com/apps>

[10] R Documentation, R Based Twitter Client

<https://www.rdocumentation.org/packages/twitteR/versions/1.1.9>

[11] El concepto de corpus y su definición

<http://elies.rediris.es/elies18/23.html>

[12] Dropping common terms: stop words

<https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html>

[13] Building Wordclouds in R

<https://datascienceplus.com/building-wordclouds-in-r/>