UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

TESIS DE GRADO

MODELO DE EXTRACCIÓN Y ANÁLISIS DE INFORMACIÓN DE LA RED SOCIAL TWITTER

PARA OPTAR AL TITULO DE LICENCIATURA EN INFORMATICA

MENCION: INGENIERIA DE SISTEMAS INFORMATICOS

POSTULANTE: UNIV. JENNY SUSANA QUISPE AGUILAR
TUTOR METODOLOGICO: LIC. JAVIER HUGO REYES PACHECO
ASESOR: M.SC. ALDO RAMIRO VALDEZ ALVARADO

LA PAZ- BOLIVIA

2015

UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

LA CARRERA DE INFORMÁTICA DE LA FACULTAD DE CIENCIAS PURAS Y NATURALES PERTENECIENTE A LA UNIVERSIDAD MAYOR DE SAN ANDRÉS AUTORIZA EL USO DE LA INFORMACIÓN CONTENIDA EN ESTE DOCUMENTO SI LOS PROPÓSITOS SON ESTRICTAMENTE ACADÉMICOS.

LICENCIA DE USO

El usuario está autorizado a:

a) visualizar el documento mediante el uso de un ordenador o dispositivo móvil.
b) copiar, almacenar o imprimir si ha de ser de uso exclusivamente personal y privado.
c) copiar textualmente parte(s) de su contenido mencionando la fuente y/o haciendo la referencia correspondiente respetando normas de redacción e investigación.

El usuario no puede publicar, distribuir o realizar emisión o exhibición alguna de este material, sin la autorización correspondiente.

Dedicatoria

A mis padres Yola y Cirilo,

por ser mi motivación para superarme día a día.

Solo decirles reto cumplido.

Agradecimiento

A MI TUTOR Lic. Javier Hugo Pacheco Reyes, por haberme colaborado a realizar este trabajo, por el apoyo y sugerencias que ayudaron a concluir satisfactoriamente este difícil desafío.

A MI REVISOR Lic. Aldo Ramiro Valdez Alvarado, por su orientación y corrección en el desarrollo de este trabajo.

A TODA MI FAMILIA, por su compañía y apoyo en cada momento de mi formación profesional.

A MIS AMIGOS, por las experiencias vividas a lo largo de este reto, siempre los recordaré con cariño.

Muchas gracias a todos.

Jenny S. Quispe Aguilar

Índice General

Índice de Figuras y Tablas

Figuras

Tablas

Resumen

Hoy en día, cada vez tiene más importancia que el contenido de la web sea accesible en el mismo momento de su creación. Al mismo tiempo, Twitter es una red social ampliamente utilizada para acceder a información en tiempo real ya que la gran mayoría de su contenido es accesible de forma pública.

El objetivo de este proyecto es la extracción de información accesible a través de Twitter, así como la investigación de las posibilidades existentes para su procesamiento y análisis.

En este proyecto se hace una revisión tanto de artículos de investigación como de servicios relacionados con el uso de información que provee Twitter, seguida de la definición de un marco teórico que clasifique toda esa información. Se presenta el diseño de un modelo orientado en la extracción y análisis de información obtenida desde Twitter en español. Concretamente, se ha centrado en tweets publicados sobre diferentes temas de actualidad escritos por cuatro de los que más importantes medios de comunicación escrita y televisión.

Distintos aspectos de la aplicación implementada han sido evaluados. La evaluación del clasificador de texto obtiene una precisión del 97%. Los resultados obtenidos nos permiten concluir que el modelo es capaz de determinar las preferencias actuales locales, a partir de examinar los múltiples, dispersos y poco estructurados tweets, así como revelar información necesaria para que poder percatarse sobre qué temas se está hablando, cuáles han sido los mensajes más populares, cómo se relacionan esas tendencias entre sí y cómo ha sido la evolución en el tiempo de los trending topics más importantes.

Abstract

Today, every time is more important than the content of the web is accessible at the very moment of its creation. At the same time, Twitter is a social network widely used to access information in real time and that most of its content is publicly accessible.

The objective of this project is the extraction of information accessible through Twitter, as well as research of the possibilities for processing and analysis.

In this project, a review of research articles both as related to the use of information provided by Twitter, followed by the definition of a theoretical framework to classify all such information is made services. Oriented design in the extraction and analysis of information obtained from Twitter in Spanish model is presented. In particular, it has focused on tweets posted on various current issues written by four of the most important means of written communication and television.

Implemented various aspects of implementation have been evaluated. Evaluating the obtained text classifier accuracy 97%. The results allow us to conclude that the model is able to determine the current local preferences, from examining the many scattered and unstructured tweets as well as reveal information necessary for you to realize what is being talked about issues, what they were popular posts, how these trends relate to each other and how has been the evolution in time of the main topics trending.

1. MARCO REFERENCIAL

1.1. INTRODUCCIÓN

Durante estos últimos años, hemos sido conscientes del gran crecimiento que ha tenido la web, tanto es así que fue necesario dar paso a lo que hoy en día conocemos como Web 2.0. Este término está asociado a las aplicaciones web que permiten al propio usuario crear contenido y facilitan la compartición de estos. Es decir, es la representación de la evolución de las aplicaciones tradicionales a las aplicaciones enfocadas al usuario final. Algunos de los servicios que ofrece la nueva Web 2.0 son los sistemas de etiquetado (TAGS), Blogs, Wikis. Pero sin duda, los servicios que más crecimiento, en cuanto a usuarios, han tenido son los microblogging y las redes sociales (RS).

Las redes sociales han cambiado la función de la informática, del negocio de simplemente automatizar las transacciones a facilitarlas. Las redes sociales, como Facebook, Twitter y LinkedIn, Youtube, Google+ y otros, se han generalizado y la aceptación por parte de los usuarios se ha acelerado. Estas redes son un torrente de datos desatado que ofrece enormes oportunidades de negocio, tanto en términos de gestión de marca como de apertura de nuevos canales de mercado. Los usuarios se comunican entre esos miles de millones de mensajes y hablan de quiénes son y sobre lo que les gusta y lo que no.

Este incesante flujo de datos procedente de las redes sociales proporciona unos niveles de valoración sin precedentes y una oportunidad inigualable para que las empresas escuchen al cliente, obtengan información y entablen un diálogo interactivo para lograr una ventaja competitiva, además de nuevas oportunidades de negocio. Durante la selección del servicio de red social objeto del análisis, los principales factores tenidos en cuenta fueron: su popularidad en términos generales, la existencia de una API abierta que permitiera hacer uso de sus recursos de forma libre, y la utilidad para el análisis del tipo de información que albergaba.

Twitter, como otros servicios de red social del tipo de Facebook o MySpace, goza de una gran popularidad, es una de las diez páginas web más visitadas en el mundo y su número de usuarios está en constante crecimiento. En Twitter, los usuarios escriben mensajes cortos de 140 caracteres, y el servicio ha adquirido una enorme popularidad en buena parte del planeta, conjuntamente dispone de una API pública y abierta para todo tipo de desarrolladores, que

permite acceder a su información de forma práctica. Además dicha información es extremadamente densa en contenido útil dada su naturaleza “micro-blogging”. Todas estas características convierten a twitter en la opción más adecuada, de entre todas las presentes en el mercado, para el propósito final del proyecto: extracción de información útil por métodos, para un posterior análisis productivo.

1.2. ANTECEDENTES

La minería de datos se encuentra en pleno desarrollo y aplica a varias disciplinas como las bases de datos, estadística, bodegas de datos, visualización de datos y obtención de información. También se utilizan métodos en las áreas de reconocimiento de patrones, redes neuronales, análisis espacial de datos y procesamiento de señales.

Cada año, en los diferentes congresos, simposios y talleres que se realizan en el mundo se reúnen investigadores con aplicaciones muy diversas. Sobre todo en los Estados Unidos, la minería de datos se ha ido incorporando a la vida de empresas, gobiernos, universidades, hospitales y diversas organizaciones que están interesadas en explorar sus bases de datos. Podemos decir que "en minería de datos cada caso es un caso" (Molina, 2002).

La Minería Web es una técnica de análisis para webs que deriva de la minería de datos. Se usa para el estudio de varios aspectos esenciales de un sitio y ayuda a descubrir tendencias y relaciones en el comportamiento de los usuarios que sirven como pistas para, por ejemplo, mejorar la usabilidad de un sitio. Generalmente se analizan grandes volúmenes de información, utilizando algoritmos y luego se los representa en modelos para que puedan ser analizados.

Actualmente, existen trabajos de minería de datos aplicada al campo empresarial, la banca, la educación, al deporte y otros, pero investigaciones enfocadas a las redes sociales son pocas y particularmente en nuestro país por el momento ha sido mínima las investigaciones. Sin embargo, en la Carrera de Informática de la Facultad de Ciencias Puras y Naturales de la Universidad Mayor de San Andrés se encuentra tesis de grado relacionado con la minería de datos y la minería web:

- Título: *“Minería de datos en el sistema de información educativa”*, Caso: Educación Formal, Autor: Glizeth Rojas Fernández, **2008**, *Carrera de Informática,* Institución: *Universidad Mayor De San Andrés. Resumen:* proyecto de grado el cual describe la

aplicación de minería de datos sobre los datos de educación formal del Sistema de Información Educativa (SIE) del Ministerio de Educación y Culturas para descubrir patrones de comportamiento y características de la población estudiantil de las unidades educativas del área rural y urbana para conocer mejor el sector educativo y apoyar a la toma de decisiones. Aplicando el proceso KDD y árboles de decisión mediante reglas con la herramienta WEKA.

- Título: *"Modelo Predictivo para la evaluación del rendimiento académico aplicando minería de datos"*, Caso, **2012**, *Carrera de Informática*, Institución: *Universidad Mayor De San Andrés. Resumen:* tesis de grado el cual realiza un estudio para evaluar el rendimiento académico estudiantil en relación a la deserción universitaria. La investigación toma para el análisis datos académicos de los estudiantes de la Carrera de Informática. Proponiendo un modelo predictivo aplicando minería de datos utilizando arboles de decisión.
- Título: "Minería Web utilizando lógica difusa*"*, Autor: Juan Pablo Poma Chuquimia, 2009, *Carrera de Informática*, Institución: *Universidad Mayor de San Andrés*. Resumen: tesis de grado que intenta resolver el problema de la falta de información con respecto a las preferencias de los usuarios que visitan un sitio web, mediante el desarrollo de un prototipo capaz de obtener conocimiento agrupando a los usuarios de un sitio web según las páginas que haya visitado utilizando como fuente los datos que guardan los servidores web. Entre los métodos usados se encuentran el time out para identificar sesiones; para realizar las agrupaciones se utiliza dos algoritmos: fuzzy c mean y fuzzy c medioid. Para la validación de los agrupamientos se utiliza dos métodos los cuales son el índice de realización difusa y entropía de la clasificación normalizada.

Por otra parte se encontró trabajos publicados en internet que involucran la aplicación de MW en el entorno del presente trabajo:

- Título: "*Metodología para extraer intereses de usuarios de twitter para generación de recomendaciones"*, Autor: Mario Castro Squella, 2010, *Facultad de Ciencias Físicas y Matemáticas Departamento de Ciencias de la Computación*, Institución: *Universidad de Chile*. Resumen: Este trabajo tuvo como objetivo aplicar una metodología para identificar y clasificar términos indicativos de intereses de los usuarios. El conocer estos intereses resulta útil para tareas de filtrado de información como lo es la generación de recomendaciones. Para identificar y clasificar los términos se hizo uso de técnicas estadísticas y resultados de búsquedas en el motor de búsqueda google.

- Título: "*Sistema de extracción de entidades y análisis de opiniones en contenidos Web generados por usuarios"*, Autor: Álvaro José Casado Valverde, 2013, *Escuela Politécnica Superior*, Institución: *Universidad Autónoma de Madrid*. Resumen: En este trabajo de fin de grado se ha diseñado y desarrollado una herramienta capaz de identificar entidades nombradas en la plataforma twitter, y analizar las valoraciones o sentimientos vertidos sobre las mismas, concluyendo con un resumen del análisis realizado a través de una interfaz Web. Se ha tratado una temática novedosa y en auge, que es el análisis de sentimiento, donde no hay una solución definitiva. En este proyecto se da una propuesta de análisis de sentimiento, añadiendo como elemento diferenciador respecto a otras herramientas de sentimiento en twitter el reconocimiento de sentimiento sobre entidades y usuarios.

La investigación de Banerjee (2007) fue la que resultó ser de mayor influencia para este trabajo, en ella se investigó twitter como una potencial fuente de información contextual acerca de sus usuarios, y se encontró que existe una cantidad suficiente de palabras clave indicativas que manifiestan intereses de los usuarios y que en estas se pueden encontrar asociaciones y aplicarles técnicas de clasificación. Twitter es una red social donde la gente comparte públicamente información sobre sus opiniones personales. Un usuario puede también hacer re-tweet de la información de otro usuario y así compartirla.

1.3. PLANTEAMIENTO DEL PROBLEMA

La aceptación y el progresivo uso de Internet por la sociedad, han cambiado radicalmente los hábitos de comunicación de las personas, haciendo posible que millones de individuos puedan estar interconectados permanentemente desde cualquier lugar, a través de la web 2.0 y las redes sociales.

Las redes sociales son un torrente de datos desatado que ofrece enormes oportunidades de negocio. Los usuarios se comunican entre esos miles de millones de mensajes y hablan de quiénes son y sobre lo que les gusta y lo que no. Este incesante flujo de datos procedente de las redes sociales ha despertado el interés de las empresas, que en muchos casos establecen contratos con las propias redes sociales para acceder a todos los mensajes públicos y realizar así un proceso de minería de datos, de gran utilidad para realizar estudios sociológicos sobre el comportamiento de los usuarios en las redes sociales o simplemente para conocer la

satisfacción de los consumidores respecto a un producto o servicio, pues se sospecha que la mayor parte de los mensajes contienen alta dosis de opinión.

Medios de comunicación, realizar campañas publicitarias con las que ampliar su cartera de clientes, difundir información de manera pública a sus clientes.

La gran cantidad de información presente en la redes sociales hace de su recopilación y tratamiento una tarea compleja donde, a su vez, la correcta realización de dichas tareas permite generar datos útiles en diversos ámbitos, desde capturar las preferencias actuales (tecnología, música, deportes, etc.) de distintos perfiles sociales, tendencias o temas del momento, por ejemplo.

Los medios de comunicación, empresas y firmas comerciales han encontrado el medio ideal para interactuar con los usuarios, se han tenido que adaptar para formar parte de este cambio, lo que ha producido la presencia de estos medios en la red mediante sus propias páginas web tratando de captar la atención de su audiencia, ofreciéndole contenido masivo y disperso.

Hoy por hoy ese flujo de contenido e información no se está aprovechando en todo su potencial porque existen las siguientes dificultades:

- Se genera cantidades masivas de información dispersa entre los mensajes que se publican provocando que los usuarios fácilmente se saturaren de información.
- Existe una mezcla de relevancia con irrelevancia de los mensajes publicados induciendo a no determinar patrones de comportamiento.
- Los mensajes publicados son poco estructurados y por tanto se dificulta el procesamiento esta información.
- Solo grandes empresas se benefician con este flujo de información, ya que establecen contratos con las redes sociales, esto provoca una desventaja para las pequeñas y medianas empresas.

Dentro de lo que son las redes sociales, una de los más relevantes hoy en día es Twitter, que contiene una gran cantidad de contenido público escrito por los usuarios con respecto a determinados temas. Por lo mencionado y para la mejor compresión de aquí en adelante citaremos como ***información*** el contenido público de esta red social.

Ante la existencia de los problemas ya mencionados se construye la pregunta de investigación:

¿Cómo se puede conocer las preferencias actuales locales, examinando los múltiples, dispersos y poco estructurados tweets?

1.4. OBJETIVOS

1.4.1. Objetivo General

Plantear un modelo de minería web que permita extraer y analizar la información de los tweets publicados por los usuarios de Twitter para conocer sus preferencias actuales locales, a partir de la masiva, dispersa y desestructurada información generada en esta red social.

1.4.2. Objetivos Específicos

Los objetivos específicos de este trabajo son los siguientes:

- Realizar un estudio sobre el análisis y procesado de mensajes de la red social Twitter utilizando la API pública disponible.
- Obtener tweets de cuentas oficiales de medios de comunicación locales y usuarios particulares para el análisis correspondiente.
- Clasificar la información relevante contenida en tweets recolectados en twitter.
- Realizar un estudio de la información poco estructurada existente en los twwets.
- Realizar el modelo de extracción y análisis de información de la red social twitter aplicando el proceso de minería web.
- Explicar y concluir sobre los resultados obtenidos.

1.5. HIPOTESIS

“El modelo de minería web logrará la extracción y análisis de información para conocer los intereses actuales de los usuarios de twitter, a partir de la masiva, dispersa y poca estructurada información generada en esta red social”.

1.6. JUSTIFICACIÓN

1.6.1. Justificación Científica

El presente trabajo no pretende ser un nuevo paradigma computacional para el desarrollo de sistemas de minería web, pues solo se propone un estudio de extracción y el análisis de información para determinar intereses actuales y locales de usuarios de twitter que es de gran importancia para la ciencia de informática ya que la mayor parte de las aproximaciones existentes a este tema se han realizado para el inglés, y hay pocos trabajos dedicados al español. Debido a que no hay una solución definitiva ante este problema, en este trabajo intentará dar una aproximación al mismo, desarrollando un modelo para la extracción y análisis de información a partir de tweets publicados por medios de comunicación locales. Se puede concluir argumentando la particularidad de esta investigación, ya que no es explorada esta línea de investigación en nuestro país.

1.6.2. Justificación Social

La investigación se justifica socialmente ya que este modelo será capaz de dar indicios de las tendencias e intereses de los usuarios de twitter, a través de la extracción y análisis de los mensajes publicados en esta red social, si bien el modelo no determinará de manera concreta cuales son los productos y servicios que marcarán tendencia a futuro, este sí será capaz de mostrar los que sí están en boga, cuáles van aumentando su presencia e impacto en twitter. Observando la dinámica de la información, viendo cómo se forman grupos de opinión, a veces independientes. Estas fuentes de información que llegan a ser del interés de grandes empresas, logran establecer contratos con las propias redes sociales para acceder a dicha información, la cual utilizan para realizar estudios sociológicos sobre el comportamiento de los usuarios en las redes sociales o simplemente para conocer la satisfacción de los consumidores respecto a un producto o servicio, pues se sospecha que la mayor parte de los mensajes contienen alta dosis de opinión. Por ello la importancia de este trabajo para conocer los intereses de los usuarios y orientar de mejor manera un producto o servicio.

1.6.3. Justificación Económica

La extracción y el análisis de información son tareas difíciles y costosas, que están siendo abordadas por investigadores y empresas, provocando que haya una falta de herramientas comerciales eficaces. La construcción del modelo propuesto constituye una disminución

económica considerable referente a lo ya mencionado y a la vez incrementar las posibilidades para la mejor toma de decisiones en medios de comunicación y empresas.

1.7. ALCANCES

Si bien en un principio se consideró utilizar el mayor número de redes sociales para relacionar cada una de ellas con la solución más adecuada, fue necesario hacer una selección de éstas ya que no sólo necesitan un interfaz de programación, sino que cada una de ellas necesita un análisis propio. Por esta razón se optó por la red social twitter: ya que un usuario de esta red social crea una serie de contenidos que se engloban en ciertos intereses. En una cuenta de twitter el usuario deja mucha información propia de todo tipo y que, en su mayor parte, está accesible a todos los públicos. Uno de estos perfiles no sólo muestra los temas que le interesan a este usuario, sino su círculo de amistades, personas que considera expertos o más fiables en cualquier materia e incluso enlaces con otras páginas que dan aún más información sobre este usuario. Es bastante información, por lo que en este trabajo se tiene como alcances los siguientes puntos:

- Se monitoreará los tweets de texto que realizan los usuarios de determinados medios de comunicación de esta red social respecto a temas o servicios que ofrecen.
- Se analizaran y seleccionaran tweets de cuentas en Twitter de determinados medios de comunicación y de usuarios particulares.
- Se analizara la verdadera información relevante e excluirá la irrelevante, entre los múltiples tweets que los usuarios publican.

1.8. LIMITACIONES

El presente trabajo de tesis se limita a realizar la extracción y análisis de información de la red social Twitter, enfatizando al contenido de tweets, excluyendo aquellos que contengan una imagen o video, además de los que contengan símbolos y/o emoticonos que no aportan al análisis del presente estudio.

Por otro lado mencionar que esta propuesta de modelo solo se aplica al estudio de tweets en idioma español.

2. MARCO TEÓRICO

En este capítulo se dará a conocer la revisión bibliográfica realizada para definir el marco teórico necesario para entender, tanto el problema presentado en este trabajo de investigación, como la solución que se describirá en detalle en los siguientes capítulos. Para lograr este objetivo de la mejor manera posible, se partirá describiendo de lo general a lo particular en cada una de las secciones que se mencionan a continuación.

Debido a la naturaleza del problema descrito en el capítulo anterior, y por razones que se darán a conocer en el capítulo 3, se considera que las ramas de investigación relevantes para este trabajo son las de extracción y análisis de información de tweets publicados en la red social twitter.

2.1. MINERÍA DE DATOS

La Minería de Datos (**MD**) es una disciplina cuyo objetivo principal es la extracción de conocimiento a partir de un conjunto de datos, este conocimiento es desconocido a primera vista pero es potencialmente útil. La extracción de este conocimiento llega a ser útil, si se ha obtenido a partir de un conjunto de datos válidos y representativos (Vallejos, 2006).

El conocimiento que nos ayuda a descubrir la MD se encuentra implícito en los datos, por lo que sería difícil extraerlo con técnicas comunes de reportes, es por eso que las técnicas de MD surgieron por la necesidad de extraer el conocimiento en bases de datos. Hoy en día es frecuente ver que el tamaño de las bases de datos, utilizados por todo tipo de entidades públicas y privadas, se ha incrementado considerablemente gracias a los avances en el manejo, capacidad y precio de los medios de almacenamiento físico. Por tanto el reto es contar con herramientas que puedan explotar eficientemente la información contenida en las bases de datos (Vallejos, 2006).

Para Vallejos (2006) *"La MD se convierte importante cuando existe la acumulación de información y esta crece aceleradamente, además la necesidad de extraer el conocimiento implícito en esas colecciones de datos. Actualmente existen una gran cantidad de herramientas que ayudan aplicar técnicas de minería de datos; las podemos encontrar en soluciones empresariales y de pago, así también como muchas propuestas en el ámbito del software libre y de código abierto".*

Si bien las herramientas existentes en el mercado son muy buenas, y se ha logrado un gran trabajo en la optimización y modificación de las técnicas heredadas de otras disciplinas para facilitar su uso en minería de datos, hay temas que actualmente resultan de suma importancia plantear.

Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación del confronto entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento. En la Figura 2.1 se ilustra la jerarquía que existe en una base de datos entre dato, información y conocimiento. Se observa igualmente el volumen que presenta en cada nivel y el valor que los responsables de las decisiones le dan en esa jerarquía. El área interna dentro del triángulo representa los objetivos que se han propuesto. La separación del triángulo representa la estrecha unión entre dato e información, no así entre la información y el conocimiento (Vallejos, 2006).

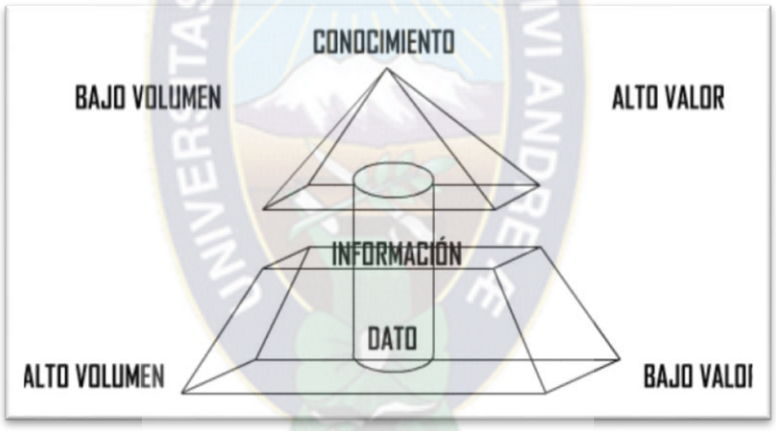


Figura 2.1. Pirámide conocimiento

Fuente: Vallejos (2006)

2.1.1. Definición

La definición de MD de la cual la literatura entrega definiciones del área que las engloban, como trabaja y en qué consiste la minería de datos. A continuación unas definiciones validas:

Fayyad (1996, Pág. 6) lo define como "*el paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre-procesados*".

Witten (2000, Pág. 5), lo define como "*el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos*".

Según Molina (2009, Pág. 23) lo define como "*el proceso de extraer conocimiento de bases de datos. Su objetivo es descubrir situaciones anómalas y/o interesantes, tendencias, patrones y secuencias en los datos. Su input son los datos pre-procesados en las fases anteriores de la metodología, el objetivo es construir un modelo a partir de ellos, el cual pueda producir nuevo conocimiento que sea útil para el usuario*".

Para este trabajo la MD lo definimos como el proceso de la extracción no trivial de información que reside de manera implícita en los datos, información que previamente es desconocida y podrá resultar útil para algún proceso.

2.1.2. Tipos De Modelos

La MD tiene como objetivo extraer conocimiento a partir del análisis de datos. Este conocimiento puede ser en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa, estas formas de representación constituyen el modelo de los datos analizados. Estos modelos pueden ser de dos tipos: descriptivos o predictivos, como se muestra en la Figura 2.2.

Figura 2.2. Tipos de modelos de minería de datos

Fuente: Elaboración propia

2.1.2.1. Modelo Descriptivo

Los modelos descriptivos logran identificar patrones que explican o resumen los datos, es decir, sirve para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Este modelo presenta las características fundamentales de los datos a ser estudiados, sin que para ello se creen trabas a la cantidad de los datos. Como resultado se logra un entendimiento más preciso de las relaciones de los datos y sus correspondientes estructuras (Hernández Orallo, 2004).

Un modelo descriptivo se preocupa de proporcionar información sobre las relaciones entre los datos. Por Ejemplo:

- Los clientes que suelen comprar pañales, compran cerveza.
- El tabaco y el alcohol son los factores más importantes en la enfermedad "X".
- Los clientes sin televisión y con bicicleta tienen características muy distintas del resto.

2.1.2.2. Modelo Predictivo

Según Hernández (2004) "*los modelos predictivos pretenden estimar valores futuros o desconocidos asociados a variables de interés"*. Para estudiarlos se pueden agrupar en dos tipos importantes, las dos tareas más usadas en los modelos predictivos son:

- *Clasificación:* Tarea que genera una función cuya salida es del tipo nominal, y permite segmentar elementos de similares características para una determinada instancia, cabe decir que para cada conjunto de entradas existe un único valor de salida.
- *Regresión:* Es una tarea que, al igual que la clasificación, genera una función cuyas variables corresponden a los de la entrada de la instancia y el resultado es el valor de la salida, la diferencia está en que la salida es una variable numérica.

En muchos casos, un modelo desarrollado con el propósito de predicción de igual forma puede convertirse eficaz para la interpretación y es el que se encarga primordialmente de responder preguntas sobre datos futuros (Hernández Orallo, 2004). Por ejemplo:

- ¿Cuáles serán las ventas el año próximo?
- ¿Es esta transacción fraudulenta?
- ¿Qué tipo de seguro es más probable que contrate el cliente X?

Muchas veces, los modelos no aspiran a ser modelos perfectos, sino modelos aproximados. En cualquier caso, al estar trabajando con hipótesis, es necesario realizar una evaluación de los patrones obtenidos, con el objetivo de estimar su validez y poder compararlos con otros. Por tanto, la minería de datos, más que verificar patrones hipotéticos, usa los datos para encontrar estos patrones, llegando a ser un proceso inductivo.

2.1.3. Métodos o Técnicas de la Minería de Datos

En la MD existen muchos métodos o técnicas que resuelven las tareas explicadas anteriormente, hay métodos que tratan distintas tareas para distintos modelos.

La MD se apoya en la aplicación de métodos matemáticos de análisis, utilizando diferentes algoritmos y técnicas de clasificación, tales como agrupamiento, regresión, inteligencia artificial, redes neuronales, reglas de asociación, arboles de decisión, algoritmos genéticos, entre otras, que son de gran utilidad para realizar el análisis inteligente de grandes volúmenes de información digital (Vallejos, 2006).

Las técnicas de MD pueden clasificarse de acuerdo con los dos grandes grupos de minería de datos, como se observa en la Tabla 2.1 (Moreno, 2001).

Supervisados	No supervisados
Arboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (clustering)
Series temporales	Reglas de asociación
	Patrones secuenciales

Tabla 2. 1. Métodos o técnicas de minería de datos

Fuente: Moreno (2001).

2.1.3.1. Arboles De Decisión

Los árboles de decisión son clasificados como método o técnica de aprendizaje supervisado, pues deben ser entrenados con información que contiene un histórico de los propios datos y los resultados que han sido consecuencia de dichos datos para poder utilizarse con el fin de crear predicciones.

Un árbol de decisión tiene unas entradas las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta, esta herramienta utiliza valores para tomar las entradas y salidas correspondientes, los cuales pueden ser valores discretos o continuos. Regularmente se manejan los valores discretos debido principalmente a su simplicidad, además se debe destacar que en una función al utilizar un valor discreto, la aplicación se denomina “clasificación”. En cambio al utilizar los valores continuos, nos encontramos hablando de una “regresión”.

Durante el proceso se lleva a cabo un test a medida que este árbol de decisión se recorre hasta las hojas para alcanzar una determinación. El árbol además contiene nodos internos, nodos de probabilidad, nodos hojas y arcos, los cuales se encargan de diversas propiedades. Un nodo interno contiene un test sobre algún valor de una de las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada.

En resumen, los árboles de decisión son diagramas de decisiones secuenciales que nos muestran sus posibles resultados. Las empresas son unas de las entidades que más utiliza este tipo de técnica, ya que les ayuda a determinar cuáles son sus opciones al mostrarles las distintas decisiones y sus resultados.

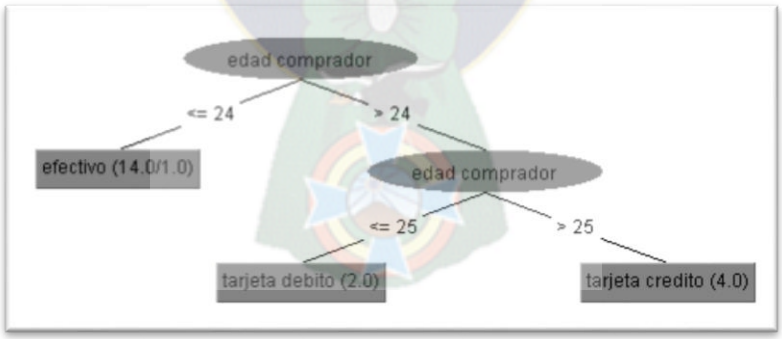


Figura 2.3. Ejemplo de Árbol de decisión

Fuente: Elaboración propia

Podemos ver en la Figura 2.3 un *ejemplo* del cual el objetivo es predecir si una persona dada que es entrevistada para un puesto de trabajo, según sus condiciones, lo obtendrá o no.

2.1.3.2. Agrupación

Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores que utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

A los vectores de un mismo grupo se les denomina “clústeres”, de aquí el nombre del proceso, los cuales comparten propiedades comunes. El conocimiento de los grupos te permite hacer una descripción sintética de un conjunto de datos multidimensional complejo. Esta se consigue sustituyendo la descripción de todos los elementos del clúster, por una descripción característica de un representante del grupo.

Desde un punto de vista práctico, el clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos espaciales, aplicaciones web, marketing, diagnóstico médico, análisis de ADN en biología computacional, y otras.

2.1.3.3. Redes Neuronales

Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Como indica Warehousing (2009) se trata de técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas, la topología de la red y los pesos de las conexiones determinan el patrón aprendido.

Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Las redes neuronales se utilizan para la clasificación y para el reconocimiento de patrones. Dado un conjunto fijo de ejemplos de entrenamiento, existen muchos modelos que podrían representar esos datos, y cada algoritmo de aprendizaje determina alguno de esos modelos (Craven et al., 1998).

Se dice que las redes neuronales son, probablemente, el algoritmo más complicado de clasificación, necesita una gran cantidad de variables y bastante tiempo para la etapa de entrenamiento, pero una vez entrenada la red, puede realizar predicciones incluso en tiempo

real, otra característica es que sólo operan con números, por lo cual, todas las variables no numéricas deben ser convertidas a número.

2.2. MINERÍA WEB

La Minería Web **(MW)** hoy por hoy es un área de investigación amplia dentro de varios grupos de trabajo, primordialmente interesados debido al alto crecimiento de la información que existe en la web y por el tendencia económica que ha generado el e-commerce y sobre todo para intentar resolver los problemas que se han mencionado anteriormente, ya sea de manera directa o indirecta (Baeza R. , 2004). Actualmente, el objetivo principal es aprender de comportamientos de los usuarios en su andar por la web y así proporcionarles información realmente relevante, útil y personalizada en muchos casos.

En la última década se ha producido un crecimiento explosivo de la información disponible en la World Wide Web (www). Hoy en día los buscadores de internet proporcionan enormes cantidades de fuentes de datos de texto y multimedia. Esta profusión de recursos ha creado la necesidad de desarrollar técnicas automáticas de MD en la www, a las cuales se les ha dado el nombre genérico de minería web (Baeza R. P., 2005).

Para Baeza (2004) *"la necesidad de crear sistemas inteligentes tanto por parte del servidor como del cliente para extraer eficientemente el conocimiento tanto de internet como de sitios web particulares ha atraído la atención de investigadores procedentes de los dominios de Recuperación de Información (Information Retrieval), Descubrimiento de Conocimiento (Knowledge Discovery), Aprendizaje Automático (Machine Learning) e Inteligencia Artificial (Artificial Intelligence) entre otros".*

2.2.1. Definición de minería web

En el ámbito del acceso, recuperación y organización de información, la MW es un campo importante de aplicación en Internet. Se utiliza para el estudio del comportamiento de ciertos aspectos esenciales para mejorar la arquitectura de un sitio ayuda a descubrir conocimientos potencialmente útiles a las organizaciones.

Pighin (2001) la define el proceso que agrupa a todas las técnicas, métodos y algoritmos utilizados para extraer información y conocimiento desde los datos originados en la web. Parte

de estas técnicas apuntan a analizar el comportamiento de los usuarios, con miras a mejorar continuamente la estructura y contenido de los sitios que son visitados.

Etzioni (1996) la define como el empleo de las técnicas de la MD para descubrir y extraer información automáticamente del web. Entre sus campos de aplicación principales se encuentran:

- Los motores de búsqueda.
- El comercio electrónico.
- El diseño Web.
- El posicionamiento Web.
- La seguridad.

Liu (2007) lo define términos generales como la aplicación de métodos de MD adaptados a la web para descubrir y extraer información de documentos y servicios web, analizando el contenido de documentos web, las páginas web que están vinculadas a través de hipervínculos y esta-dísticas de uso para ayudar a los usuarios a satisfacer necesidades de información.

Teniendo en cuenta lo anteriormente, la MW es un proceso de recuperación de información, de descubrimiento y análisis de información relevante que involucra el uso de técnicas y acercamientos basados en la MD orientados al descubrimiento y extracción automática de información de documentos y servicios de la web, tomando en consideración el comportamiento y preferencias del usuario.

2.2.2. Datos en minería web

Según Baeza (2004) "*la web no es más que una colección enorme de datos diversos y dinámicos, lo que provoca problemas de escalabilidad, heterogeneidad y dinamismo. La MD intenta dar respuesta a la nada trivial tarea de identificar los datos válidos, novedosos o potencialmente útiles que encierra la web"*.

En minería web, los datos pueden ser recogidos por parte del servidor, el cliente u obtenidos de bases de datos de una organización. Dependiendo de la localización de la fuente, el tipo de datos puede diferir, existiendo una gran variación en los contenidos (textos, imágenes, audio, símbolos). Esto hace que las técnicas utilizadas en MW cambien según la tarea

particular que haya que llevar a cabo (Baeza R. , 2004). A continuación algunas características comunes de los datos web:

- No etiquetados
- Distribuidos
- Heterogéneos
- Semiestructurados
- Variables en el tiempo

Por lo tanto, la MW trata básicamente con información de gran tamaño, con hiperenlaces y con las características antes mencionadas. Además, al ser un medio interactivo, la interfaz humana es un componente clave en la mayoría de los usos de la web (Baeza R. , 2004).

Así, la MW, en vez de ser considerada un caso particular de MD, se garantiza un campo de investigación independiente, principalmente debido a las características anteriores de los datos y a las cuestiones relacionadas con el razonamiento humano.

2.2.3. Áreas o categorías de la minería web

Se suele usar la denominación MW para catalogar tres tipos de áreas consideradas diferentes. Todas estas actividades se enmarcan dentro de la MD y, además, están relacionadas con la web, pero los datos que son objeto de la minería son diferentes (Linoff, 2001).

En la Figura 2.4 se explica que en el caso de la MW los datos pueden ser obtenidos desde el lado del servidor, del cliente, de los servidores proxy o de la base de datos corporativa de la entidad a la cual pertenece el sitio, consiguiendo ser clasificados en tres tipos predominantes.

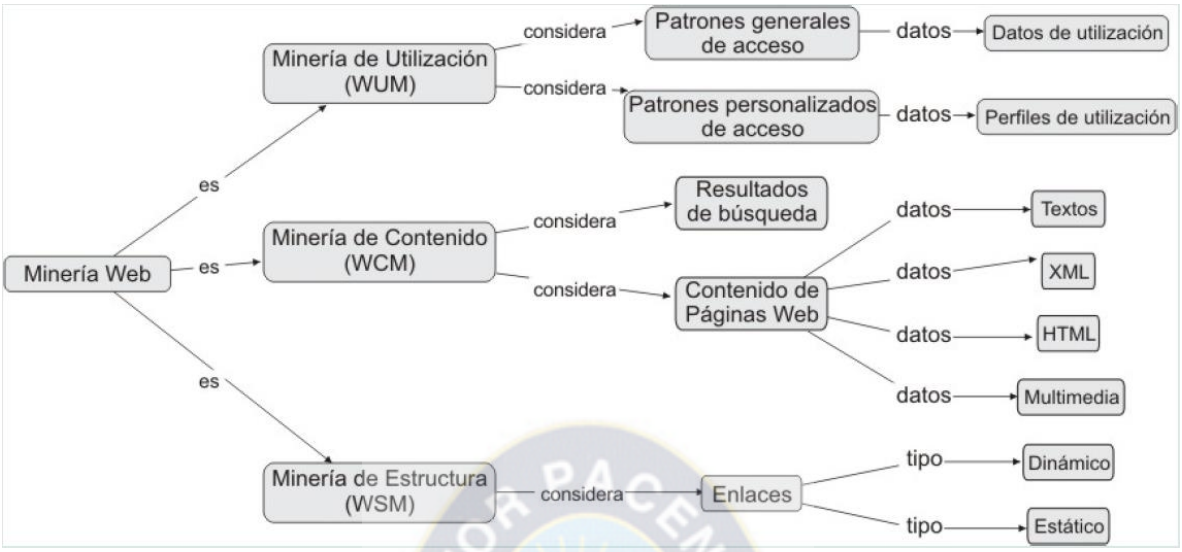


Figura 2.4. Mapa conceptual de las categorías de la minería web

Fuente: Dürsteler (1995)

2.2.3.1. Minería de contenido web

Minería de contenido web (Web Content Mining) se centra en el contenido, por lo que se pueden obtener datos acerca de la forma de escribir que sea más atractiva para el usuario, de si la catalogación que se usa sirve para mejorar la relevancia del sitio, si los temas que se tratan interesan o no. Esta área de la MW tiene dos vertientes: recuperación de la información y base de datos (Blockeel, 2000).

Se enmarcan aquí los procesos cuyo objeto es extraer información sobre la topología de la web, es decir, de los enlaces entre las páginas. Se trataría por tanto, de responder a preguntas del tipo ¿qué páginas son usualmente accedidas desde otras páginas?, o ¿cuáles son las páginas origen que llevan a otras determinadas páginas? (Cooley, 1997).

Como se conoce, los sitios de web están compuestos de colecciones de documentos de hipertexto. La recuperación de la información se realiza a través de la exploración semántica de los documentos, mediante dos enfoques: la minería de textos y el análisis semántico de los textos (Blockeel, 2000).

Considerando que los sitios de web también son colecciones de documentos semiestructurados, se pueden descubrir y extraer esquemas para formularios que capturen información semántica relevante de fuentes de datos heterogéneas. Los enfoques están

basados en lenguajes de consultas para web, base de datos múltiples y descubrimiento de jerarquías.

2.2.3.2. Minería de estructura web

Minería de estructura web (Web Structure Mining) se refiere al grado de dificultad que tienen los usuarios para encontrar la información, si la estructura del sitio es simple o muy profunda, si los elementos están colocados en los lugares adecuados dentro de la página, si la navegación es comprensible, cuáles son las secciones menos visitadas y su relación con el lugar que ocupan en la página central (Blockeel, 2000).

La minería de estructura web revela más información que simplemente la información contenida en los documentos. Por ejemplo, enlaces o eslabones que apuntan a un documento indican su nivel de popularidad, mientras que los enlaces o eslabones que salen de un documento indican la riqueza o quizás la variedad de temas que se abarcan en el documento.

2.2.3.3. Minería de uso de la web

Minería de uso web (Web Usage Mining) tiene como objetivo la extracción de patrones de navegación que se pueden descubrir en los usuarios que visitan un sitio y que pueden ser útiles para mejorar la navegación (Blockeel, 2000).

Según Linoff (2011) "*para llevar a cabo un proyecto de minería de uso de la web, como en todo proyecto de minería de datos, es necesario seguir un proceso perfectamente definido. En la fase inicial, se establecen los objetivos desde el punto de vista del negocio, así como las estrategias de validación de estos objetivos. En la fase siguiente se reúne los datos que formarán parte del análisis, pudiendo ser ficheros históricos de logs del servidor o servidores del sitio web a analizar, datos de los clientes/usuarios, datos demográficos, datos de facturación y marketing, entre otros. Una vez recopilados los datos, se llevarán a cabo tareas de limpieza y selección de los mismos, donde se identificarán las sesiones y transacciones de usuario".*

2.2.4. Etapas de la minería web

La MW puede dividirse en cuatro etapas (Etzioni, 1996) ver la Figura 2.5:

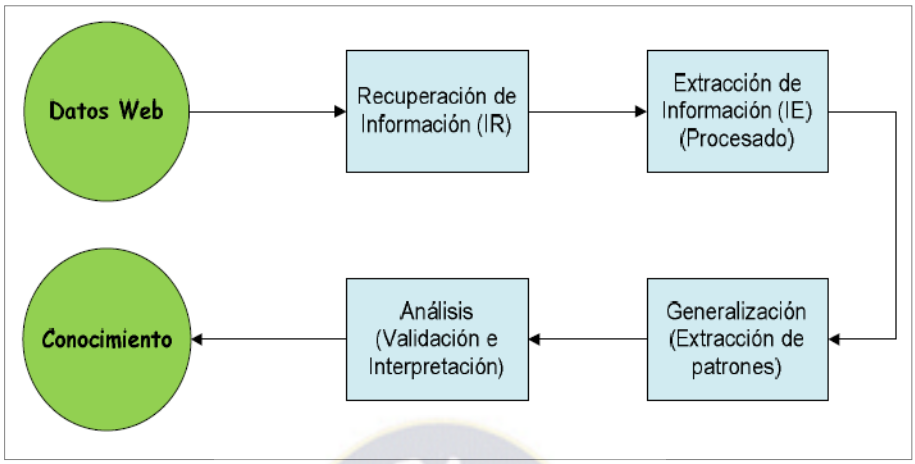


Figura 2.5. Etapas de minería web

Fuente: Etzioni (1996)

2.2.4.1. Recuperación de la Información

La Recuperación de Información (**RI**) trata acerca de la recuperación automática de todos los documentos relevantes de un conjunto de conocimiento, asegurando al mismo tiempo que los documentos no relevantes recuperados sean los menos posibles. El proceso de IR incluye principalmente la representación de documentos, el indexado, y la búsqueda de documentos (Etzioni, 1996).

Existen diferentes técnicas de minería de contenidos web utilizadas por distintos autores para la Recuperación de Información en documentos semiestructurados. Estas técnicas se basan generalmente en la utilización de índices. Un índice es, básicamente, una colección de términos con indicadores a los lugares en los que puede encontrarse la información sobre los documentos. Sin embargo, indexar páginas web para facilitar la recuperación es un proceso bastante complejo, y un reto si se compara con el problema correspondiente asociado a bases de datos clásicas, donde las técnicas directas son suficientes (Etzioni, 1996).

El enorme número de páginas web, su dinamismo, y su frecuente puesta al día hace que las técnicas de indexado parezcan aparentemente imposibles de aplicar. Actualmente, existen cuatro aproximaciones para el indexado de documentos en la web: indexado humano o manual; indexado automático; indexado inteligente o basado en agentes; e indexado basado en metadatos.

2.2.4.2. Extracción y Preprocesamiento

Una vez los documentos se han recuperado, el desafío es extraer conocimiento automáticamente y otras informaciones requeridas sin la interacción humana. La Extracción de Información (**EI**) consiste en la transformación de una colección de documentos, habitualmente con la ayuda de sistemas de RI, en información más fácil de asimilar y analizar. EI intenta extraer hechos relevantes de los documentos, mientras que RI selecciona los documentos relevantes. Por tanto, podríamos decir que EI trabaja con una granularidad más fina que RI. En todo caso, los conceptos EI e RI pueden llegar a confundirse en la práctica (Blockeel, 2000).

Para Blockeel (2000) *"la EI tiene como objetivo extraer el nuevo conocimiento de los documentos recuperados en la estructura y representación del documento mediante la conversión en mayúsculas, teniendo en cuenta que los expertos de RI consideran que el texto del documento es una bolsa de palabras y no prestan atención a la estructura del documento. La escalabilidad es el mayor desafío más para los expertos de EI; no es factible construir sistemas de EI que sean escalables al tamaño y dinamismo de la web"*. Por tanto, la mayoría de los sistemas de EI extraen información de sitios específicos y se enfocan en áreas definidas.

Una de las técnicas de preprocesamiento usadas para EI es el índice semántico latente, del inglés latent semantic index, que busca transformar los vectores del documento original a un espacio dimensional más bajo mediante el análisis de la estructura correlacional en esa colección del documento de modo que documentos similares que no comparten los mismos términos se colocan en la misma categoría (tema) (Etzioni, 1996).

2.2.4.3. Generalización

Para Etzione (1996) una vez que se ha automatizado el descubrimiento y la extracción de la información procedente de los sitios web, el siguiente paso es tratar de generalizar a partir de la experiencia acumulada. Para ello, la MW ha adaptado técnicas de MD (reglas de asociación, clustering, entre otras), de la RI (algunas técnicas para la categorización y la clasificación de textos).

En esta etapa, se utilizan el reconocimiento de patrones y las técnicas de aprendizaje automático sobre la información extraída. La mayoría de los sistemas de aprendizaje en máquinas utilizados en la web aprenden más sobre los intereses de los usuarios que sobre la

propia web. Un gran obstáculo en el aprendizaje web es el problema del etiquetado: los datos son abundantes en la web, pero no están etiquetados. Muchas de las técnicas de MD necesitan entradas etiquetadas como positivas o negativas respecto a algún concepto (Etzioni, 1996).

Desafortunadamente, las páginas web no están etiquetadas. Técnicas como el muestreo de incertidumbre reducen la cantidad necesaria de datos etiquetados, pero no eliminan completamente el problema. Una aproximación a la solución está basada en el hecho de que la web es mucho más que una colección enlazada de documentos, es un medio interactivo (Baeza R. , 2004).

Los documentos procesables han llevado al desarrollo del concepto web semántica, que está inspirado en el hecho de que la mayoría de la información en la web se diseña para el consumo humano e incluso si deriva de una base de datos, la estructura de los datos no es evidente para búsquedas automáticas. A diferencia de la AI, en la que se procura emular el comportamiento del cerebro humano, la web semántica, en su lugar, desarrolla lenguajes para expresar información de forma procesable para las máquinas.

2.2.4.4. Análisis

El análisis es un problema de manipulación de datos que requiere que existan datos suficientes disponibles para que la información potencialmente útil se pueda extraer y analizar. Los seres humanos juegan un papel importante en el proceso de descubrimiento del conocimiento en la web, considerando que la web es un medio interactivo. Esto es especialmente importante para la validación y la interpretación de los patrones de minería que tienen lugar en esta etapa (Etzioni, 1996).

Una vez que los patrones han sido descubiertos, los analistas necesitan herramientas apropiadas para comprender, visualizar e interpretar estos patrones. Algunos usan el Procesamiento Analítico en Línea, del inglés Online Analytical Processing (OLAP) con el propósito de simplificar el análisis de las estadísticas a partir de los logs de los servidores, y otros mecanismos SQL, como el sistema WEBMINER que propone un lenguaje de consultas, similar a SQL, que posee un mecanismo de consultas para preguntar acerca del conocimiento descubierto (en forma de reglas de la asociación y modelos secuenciales) (Hernández, 2004):

En definitiva, la MW puede verse como el uso de las técnicas de MD aplicadas a la búsqueda, extracción y evaluación automática de información para el descubrimiento del conocimiento de los documentos y servicios web, como se mencionó anteriormente.

2.3. WEB 2.0

El término Web 2.0 está acuñado a Tim O'Reilly, fundador de O'Reilly Media durante una sesión de brainstorming para desarrollar ideas para una conferencia.

2.3.1. Definición

El termino Web 2.0 sugiere una nueva versión de la Web, aunque no se refiere a un cambio en las especificaciones técnicas, sino más bien a los cambios en cuanto a las especificaciones de los usuarios finales y los cambios en el desarrollo de páginas web (Alag, 2008).

Según Alag (2008) *"la Web 2.0 propone una nueva visión de la World Wide Web, basada en la compartición de la información, el diseño centrado en el usuario y la colaboración entre los usuarios. Un sitio Web 2.0 es aquel que permite a los usuarios interactuar entre sí y crear contenido dentro de una comunidad virtual, a diferencia de la Web tradicional en la que los usuarios se limitan a observar los contenidos".*

Todo lo mencionado también trajo consigo un cambio en el modelo de negocio utilizado en la World Wide Web. La Figura 2.6, conocida como mapa de Markus Angermeier, muestra de forma visual este concepto.

Figura 2.6. Mapa de Markus Angermeier

Fuente: Internet (sf)

2.3.2 Servicios de Web 2.0

La Web 2.0 ofrece una serie de servicios o herramientas diseñadas para la compartición o creación de contenido o la interacción de los usuarios. En líneas generales, estos servicios son los siguientes:

- **Etiquetado de recursos**: Consiste en categorizar o etiquetar los recursos de la web, de tal forma que la etiqueta representa una alusión semántica al recurso referenciado. Normalmente, esta organización o clasificación la introducen los propios usuarios. Una vez indizados los recursos, los sitios web proporcionan nubes de tags en las que el usuario puede ver los contenidos con etiquetas comunes.
- **Blog**: Un blog es un sitio web que recopila de forma cronológica textos, artículos o determinada información de uno o varios artículos. La información se muestra en el orden en el que se crea apareciendo primero el más reciente. El término Blog viene de Web + Log → Blog y se refiere a el registro de comentarios o de conversación entre usuarios en la web. Existen herramientas que permiten una fácil creación y alojamiento. Las más populares son Blogger y WordPress.
- **MicroBlogging:** También conocido como NanoBlogging, es un servicio que permite a sususuarios publicar mensajes breves (generalmente, solo texto) con un tamaño alrededor de 140 caracteres. Las actualizaciones se muestran vía web, sms,

aplicaciones ad hoc etc. Estas actualizaciones se muestran en la página de perfil del usuario y también son enviadas a otros usuarios que tengan la opción de recibirlas. El mejor ejemplo de este tipo de servicios es el que usaremos a lo largo del proyecto por su gran popularidad: Twitter.

- **Wikis:** Término proveniente del hawaiano wiki – rápido, de forma sencilla – Una wiki es una web cuyo contenido puede ser editado por múltiples usuarios a través del navegador web de forma sencilla y rápida. Los usuarios pueden crear, modificar o eliminar texto de la web de forma colaborativa. A menudo, las wikis funcionan como enciclopedias colectivas. La wiki más famosa es Wikipedia.
- **Compartición de recursos**: La web 2.0 llegó para adecuarse a la demanda de los usuarios en el uso de la web. Uno de los factores que más influyeron en esto fue la necesidad de compartir recursos a través de la web. Gracias a los servicios de compartición de recursos un usuario puede subir a la nube de internet cierto contenido al que cualquier otro usuario tendrá acceso. Existen infinidad de servicios de compartición de recursos, que podemos clasificar según el tipo de recurso compartido:
    - Documentos: Google Docs, Issuu, Calameo.
    - Fotos: Flickr, Pinterest, MetroFlog.
    - Presentaciones: Slideshare, Prezzi.
    - Vídeos: Youtube, Vimeo.
    - Generalistas: MediaFire, Rapidshare.
- **Redes Sociales:** Las redes sociales son un medio de comunicación utilizado para relacionarse on-line por personas que comparten algún tipo de relación. En los últimos años han cobrado mucha popularidad y cada día el número de usuarios de estas aumenta considerablemente. Existen multitud de redes sociales, que podemos clasificar entre generalistas y específicas.
- **Sistemas de recomendación**: Se encargan de recolectar la información sobre la que un determinado usuario está interesado para poder mostrarle después información similar (música, libros, páginas webs y otros). Como ejemplos contamos con Genius o Amazon.

2.3.3. Información en Redes Sociales

Al igual que sucede con la búsqueda de información en Internet, no es conveniente salir a recoger datos sin haber diseñado una estrategia previa, o cuando menos, sin haber

reflexionado con detenimiento qué queremos, para qué lo queremos y dónde lo vamos a buscar (Moya, 2012).

Para Moya (2012) *"estos pilares son indispensables incluso para el caso de la monitorización o vigilancia de la red, donde a pesar de que no buscamos nada en concreto, sí debemos diseñar planes que incluyan sistemas de alertas tempranas que puedan ser útiles a nuestros intereses".*

Si bien es cierto que muchas veces no tenemos claros todos los objetivos, podemos realizar un acercamiento inicial que permita una "tormenta de ideas creativa" para diseñar, a continuación, una lista de objetivos finales junto a una estrategia que nos ayude a alcanzarlos.

Podemos mencionar algunos de los objetivos de búsqueda de información en redes sociales pueden ser los siguientes:

a. Contactar o conocer expertos en una materia.
b. Conocer e interactuar con perfiles de influencia.
c. Descubrir qué se está haciendo en otros lugares.
d. Determinar posibles relaciones entre perfiles y grupos.

Respecto a los objetos de la búsqueda, pueden ser nombres de organizaciones y personas, número de seguidores y a quién siguen, quién está hablando de qué, noticias en medios de comunicación, etc. El éxito de que el objeto de búsqueda conduzca hacia el objetivo dependerá, además, de la pericia del investigador a la hora de pensar en cuáles son las claves que le facilitarán extraer la información (Moya, 2012).

2.4. REDES SOCIALES

El ser humano, a lo largo de su historia, ha buscado el apoyo de las personas que le rodean ya que estas le pueden ayudar a satisfacer sus necesidades básicas como de seguridad, protección y afecto. Debido a que el ser humano es social y vive sumergido en un tramado de vínculos interpersonales que afectan su vida. Así, el comportamiento de cada individuo afecta y a su vez se ve afectado por las interacciones sociales en las que participa (Enterría, 2012).

Según Castañeda (2010) "*las redes sociales en internet son sitios web donde las personas se dan de alta, creando así un perfil (una página web personal) para posteriormente agregar un*

perfil de sus amigos. De esta manera permite la interacción entre personas que no necesariamente se conocen pero que comparten intereses, preocupaciones o necesidades, porque las redes sociales no solo sirven para mostrar fotografías o documentos".

Las redes sociales se diseñan para ser ampliamente accesibles, hay varios tipos de redes sociales para cada interés o necesidad, dividiendo así a las redes sociales en tres tipos: generales, profesionales y temáticas. Estas permiten al usuario dejarse ver en internet, encontrar personas, mantener relaciones distantes, conocer gente nueva, compartir conocimiento, aprender, compartir contenidos, participar en grupos de interés comunes, debatir, divertirse, realizar negocios, encontrar y ofrecer trabajo (Anguita, 2014).

Una de las principales redes sociales es Twitter[1], la segunda mayor, por detrás de Facebook[2]. Twitter fue creada en 2006 y cuenta con más de 500 millones de usuarios registrados, los cuales generan más de 65 millones de tweets y 800.000 peticiones de búsqueda al día. Por otro lado tenemos a Google+[3] cuya estrategia de ir generando servicios distintos para los usuarios con una integración entre ellos que no fuera más allá que el compartir nombre de usuario y contraseña, ha tenido una historia más problemática en su introducción en las redes sociales.

2.5. TWITTER

Se trata de un servicio de red social estadounidense creado inicialmente en 2006 con la intención de facilitar el intercambio de mensajes de texto breves entre sus miembros. Considerado por muchos como un servicio de microblogging, su principal virtud junto a la sencillez de uso y simplicidad, es la limitación de los mensajes o actualizaciones lo que ha supuesto que numerosos medios y autores describan esta herramienta como los SMS de Internet (L. D'Monte, 2009).

Twitter se comenzó a fraguar en 2006 en el seno de una compañía dedicada a los podcast (archivos multimedia por suscripción) como idea para comunicar a pequeños grupos de personas partiendo del concepto de los SMS[4] en telefonía móvil. A esta idea la llamaron twttr,

[1] www.twitter.com
[2] www.facebook.com
[3] www.google.com
[4] Short Message Service

inspirado en lo que ya era Flickr[5], y tras un periodo de pruebas interno, fue lanzada al público el 15 Julio de 2006 (L. D'Monte, 2009).

Twitter tiene su sede en San Francisco, California, pero es utilizado por personas de cualquier país del mundo.

2.5.1 ¿Qué es Twitter?

Según se explica en la página de soporte del propio sitio (Support, sf), *Twitter es un mecanismo que ayuda a la gente a comunicarse y a mantenerse en contacto a través de un intercambio rápido y frecuente de mensajes de hasta 140 caracteres*. Esta limitación en la longitud de los mensajes lo hace ideal para intercambiar contenido de forma rápida, el cual se almacena de manera persistente en los servidores de Twitter, y tiene su explicación en el origen de esta herramienta.

Twitter es una red de información de tiempo real, que permite a sus usuarios conectarse a lo que consideran interesante: basta con encontrar cuentas de personas que publican contenido que les gusta y seguir sus conversaciones. Los usuarios pueden cambiar su preferencia de idioma en su configuración de usuario si así lo desean (Twitter).

Twitter se basa en pequeños trozos de información llamados tweets. Conectado a cada tweet hay un panel de detalles que permite añadir información adicional, mayor contexto y contenido multimedia embebido. En la Figura 2.7 puede visualizarse una página de perfil de un usuario arbitrario.

[5] Servicio de fotografías: www.flickr.com

Figura 2.7. Perfil de usuario de Twitter

Fuente: Cuenta oficial en twitter de @UMSABolivia

El funcionamiento de Twitter es muy simple: el usuario escribe en su página de perfil mensajes (tweets) con un máximo de 140 caracteres, que pueden ser leídos por cualquiera que tenga acceso a su página. Cada usuario tiene, además, una lista de seguidores (followers) y seguidos (following). Una vez que un usuario twittea un mensaje, este aparecerá en la página de perfil de todos sus seguidores (L. D'Monte, 2009).

Un usuario puede mencionar a otro. En este caso, se envía el tweet referido a un usuario específico, sin embargo, se diferencia de los mensajes directos en que al mencionar el contenido del tweet es visible por otros usuarios mientras que en los mensajes directos el contenido es solamente visible por el usuario destino.

Además de todo esto, twitter permite a los usuarios retwittear mensajes de otro usuarios, en ese caso, el tweet saldrá en el timeline del usuario con el usuario original del tweet, e indicando el usuario que retwitteó el mensaje original. El siguiente diagrama de caso de uso (Figura 2.8.) resume la principal funcionalidad con la que cuenta un usuario de Twitter:

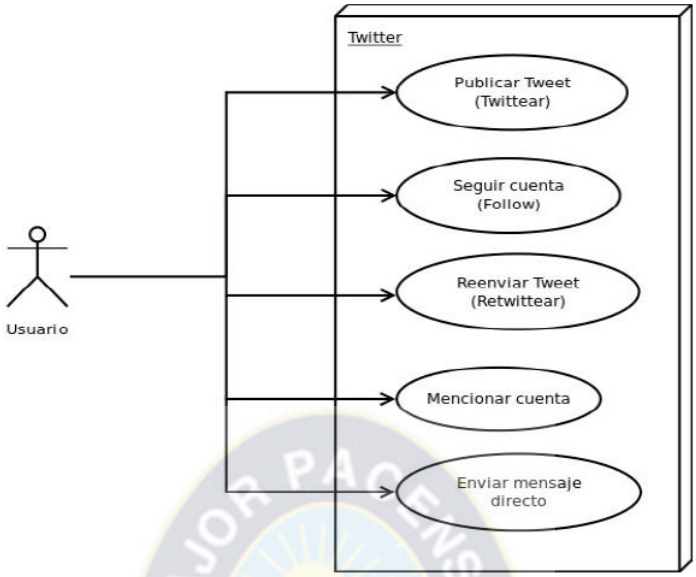


Figura 2.8. Diagrama de casos de uso Twitter

Fuente: Elaboración propia

2.5.2. Definición de Elementos Importantes

Twitter (sf) cuenta con un cierto lenguaje para algunos términos, en este apartado, daremos el significado a los términos más utilizados dentro de lo que corresponde a esta res social.

- ***Entrada ("tweet")****: mensaje de texto de 140 caracteres como máximo que es dado a conocer públicamente o notificado en exclusiva a los amigos.*
- ***Amigo ("following")****: usuario de twitter que notifica sus nuevas entradas a otro en tiempo real. En función de la configuración de privacidad, puede requerirse el consentimiento previo del primero para establecer la relación.*
- ***Seguidor ("follower")****: usuario de twitter que es notificado de las nuevas entradas de otro en tiempo real.*
- ***Re-entrada ("retweet")****: consiste en reproducir íntegramente el mensaje de otro usuario, incluyendo su nombre y la palabra "retweet" o RT al inicio.*
- ***Respuesta ("reply")****: respuesta al "tweet" de otro usuario de forma pública.*
- ***Mención ("mention")****: el signo @ seguido por un nombre de usuario permite a los usuarios mencionar a otros en sus "tweets".*
- ***Mensaje dirigido ("direct message")****: forma privada de comunicación entre usuarios de twitter.*

- ***Tema ("topic" o "hashtag")****: palabras o frases con un prefijo #. Normalmente los temas con mayores menciones diarias* ***("trending topics")****, suelen estar relacionadas con noticias de actualidad o acontecimientos recientes.*
- ***Listas ("lists")****: permiten clasificar a los amigos twitter en grupos a voluntad del usuario, para poder seguirlos de forma más fácil.*
- ***Usuario popular ("twitterati")****: usuario muy leído o influyente que posee un gran número de seguidores. En general suele tratarse de personajes públicos o celebridades, por ejemplo Barack Obama (http://twitter.com/BarackObama).*

2.5.3. La API de Desarrollo de Twitter

API (Application Programming Interface) es un conjunto de reglas (código) y especificaciones que las aplicaciones pueden seguir para comunicarse entre ellas: sirviendo de interfaz entre programas diferentes de la misma manera en que la interfaz de usuario.

El API de Twitter es abierta, se usa para cualquier acceso a Twitter que no sea la web de Twitter, la cual tiene actualmente una restricción de 100 usos por hora para usuarios normales, y las aplicaciones que la usan masivamente desde direcciones IP únicas o limitadas, tienen otro tipo de limitaciones. Las mejores están incluidas en una lista blanca que no tiene límite de peticiones por hora, aunque esto tiene un límite de días (esTwitter,sf).

Twitter ofrece una completa API para desarrollar aplicaciones integradas con twitter, está formada por tres grandes bloques: REST API, Search API y Streaming API. Cada una de ellas tiene unas funcionalidades específicas (Twitter, API).

2.5.4. Herramientas para el proceso de consultas

La ***API STREAMING*** no establece una conexión en respuesta a la consulta de un usuario. Por un lado un proceso streaming toma tweets y realiza parseo, filtrado y/o agregación antes de almacenar el resultado. Por otro lado el proceso manejador HTTP consulta la base de datos donde se almacenan los tweets en respuesta a la solicitud del usuario (Twitter Developers, sf). La opción API Streaming:

- ✓ Proporciona un flujo constante de una gran variedad de tweets en tiempo real.
- × Los elementos de consulta sobre la propia API son menores, solo con unos pocos parámetros en común si así se desea.

- ✗ Por cada aplicación solo puede haber una conexión.
- ✗ El flujo de datos proporcionado tiene un límite máximo de tweets por cantidad de tiempo por lo que cuando se incrementen drásticamente los tweets publicados sobre alguna tendencia todo aquel que no entre en el pipe de información se perderá ya que es imposible realizar consultas que capturen tweets anteriores en el tiempo.

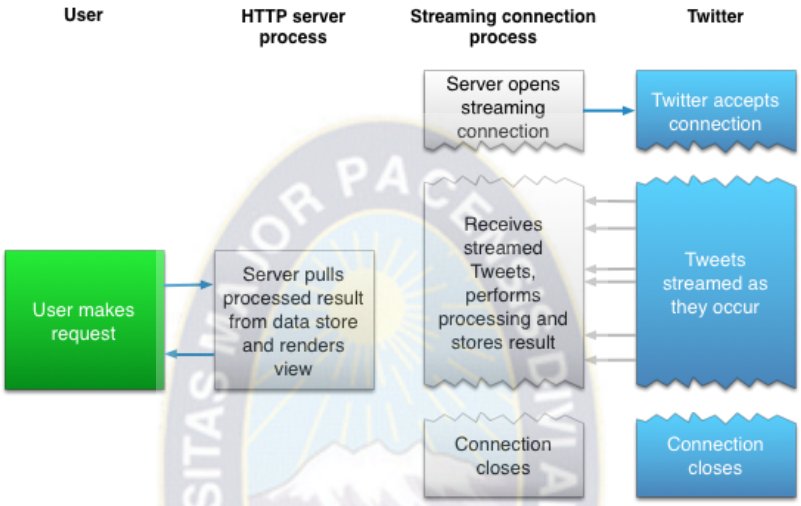


Figura 2.9.Funcionamiento de STREAMING API

Fuente: Streaming (sf)

La ***API REST*** no necesita tener abierta una conexión HTTP persistente. Una aplicación web con esta API podría realizar una o más consultas (Twitter, API). La opción API REST:

- ✓ Los elementos de consulta son mayores y más específicos que en Streaming API, pudiendo realizar consultas con más detalle.
- ✓ Por cada aplicación creada puede haber varias conexiones utilizándola.
- ✗ Tiene un límite de consultas y tiempo por usuario que el desarrollador deberá controlar, esto significa que habrá tiempos muertos en la recolección de tweets.

La API de REST de Twitter permite acceder a algunas de las primitivas básicas de Twitter, incluyendo líneas cronológicas, actualizaciones de estado y la información del usuario. Se enfoca a aplicaciones que aprovechan los principales objetos de Twitter.

El funcionamiento de la API de REST de una aplicación que acepta solucitudes de un usuario se puede observar en la Figura 2.10 se pueden realizar, una o varias peticiones a la API de Twitter, la cual formatea y devuelve el resultado al usuario en respuesta a la solicitud inicial.

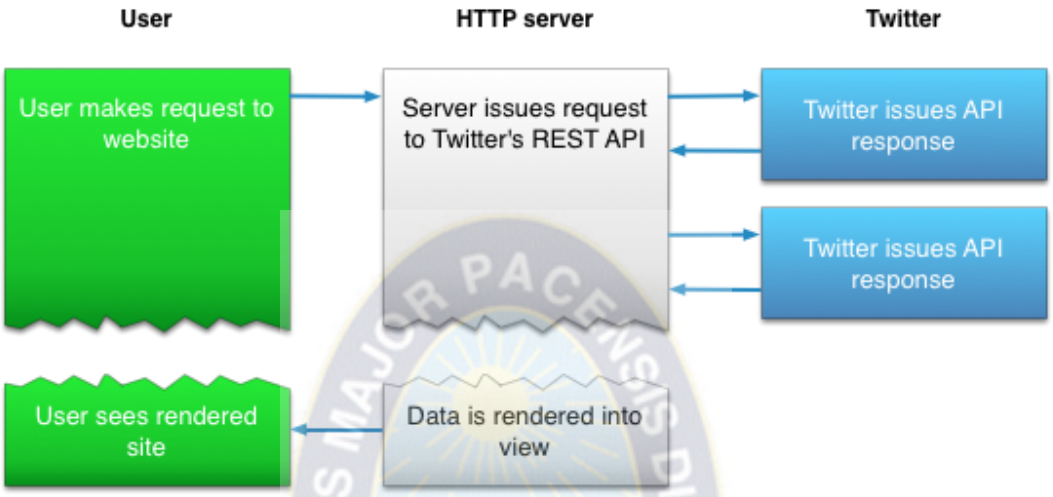


Figura 2.10. Funcionamiento de REST API

Fuente: (Rest Api, sf)

Existe un proceso de manipulación HTTP que hace la consulta a Twitter según las peticiones de los usuarios. Por el contrario, en la Figura 2.9 se percibe como la conexión en Streaming se ejecuta en un proceso adicional separado del proceso que maneja las solicitudes HTTP.

Con esto se percibe que la API de REST es más ligera y menos compleja. No es necesario el almacenamiento de grandes cantidades de datos, y a su vez permite al usuario realizar peticiones concretas en tiempo real y pasado interactuando directamente con Twitter.

2.5.5. ¿Qué devuelve el API?

Cuando el usuario realiza una solicitud de datos a Twitter, éste envia una respuesta en formato JSON. El resultado es un único objeto que se divide en dos campos principales: *search_metadata* y *statuses.*

search_metadata: Este campo es un objeto que tiene información sobre la solicitud que se ha realizado a Twitter. Se divide en la siguiente información:

completed_in: Tiempo que tarda en realizarse la consulta a Twitter.

count: Número de resultados que se han solicitado en la consulta.

max_id: Máximo identificador de que se ha obtenido, es decir, el identificador del Tweet más antiguo de la consulta. Sirve de referencia a next_results si se pretende hacer una consulta sobre resultados más antiguos.

max_id_str: Representación en forma de String del parámetro max_id.

next_results: Solicitud preparada para obtener los datos que permita devolver Twitter hacia atrás en el tiempo dentro de sus limitaciones. Se puede obtener información de entre 6 y 15 días de antigüedad.

query: El término o términos de los que se ha realizado la búsqueda.

refresh_url: Solicitud preparada para obtener los datos que permita más recientes de Twitter sobre la "query" buscada.

since_id: Identificador del Tweet del que parte refresh_url para obtener resultados recientes.

since_id_str: Representación en forma de String del parámetro since_id.

statuses: Es un array con tantos objetos como número de resultados se hayan obtenido. Equivale al mismo número que tiene el campo search_metadata[count]. Estos objetos son tweets, en la Figura 2.11 se ve a detalle la estructura de un tweet. Para una información más exhaustiva, consultar el anexo A o la documentación oficial de twitter (Twitter, 2015)

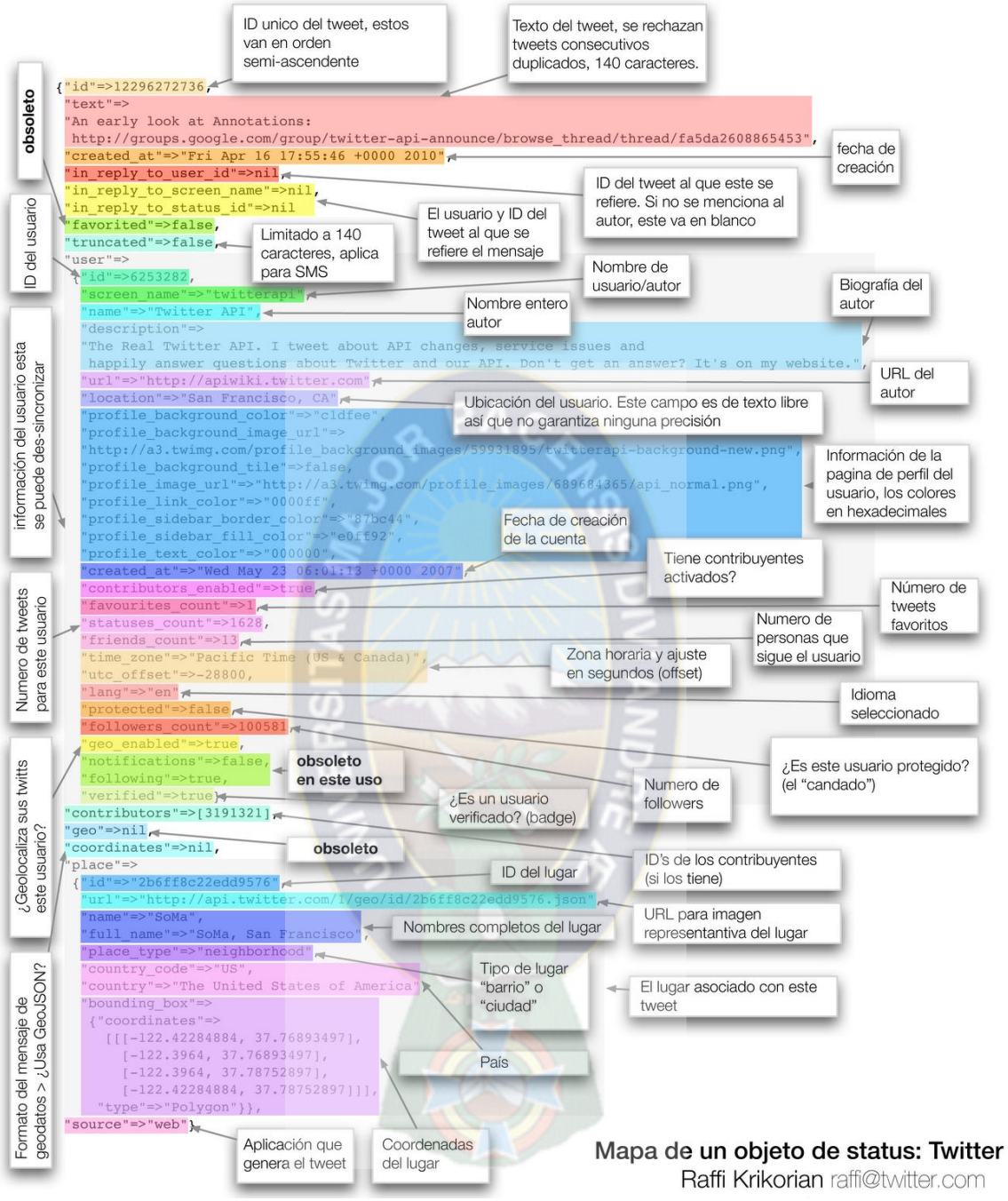


Figura 2.11. Mapa de un objeto de estatus en Twitter

Fuente: Conecti (sf)

2.5.6. Información extraíble a través de la API Rest de Twitter

Es importante destacar que la API Rest proporciona acceso a información ya existente en twitter en el momento de hacer la llamada. A continuación se presenta un listado de las

posibles peticiones a la API Rest de Twitter. Sólo incluimos aquellas peticiones relacionadas con la obtención de información, pasando por alto aquellas que estén relacionadas con el manejo de cuentas de usuario o su configuración. Debido a que es una lista moderadamente extensa, se presenta dividida en secciones dependiendo del dato de entrada necesario para hacer la petición, distinguiendo entre un usuario registrado, una lista de usuarios, un tweet o una consulta. Dentro de esta última posibilidad, distinguimos entre consultas orientadas a la búsqueda de tweets, a la búsqueda de usuarios o consultas de ámbito general.

2.5.6.1. Peticiones que reciben como entrada un usuario

- **Últimos tweets publicados por el usuario.** Hasta 150 tweets.
- **Últimos tweets que contienen una mención al usuario**. Hasta 20 tweets.
- **Últimos tweets marcados como favoritos por el usuario.** Hasta 20 tweets.
- **Últimos tweets que han sido retweeteados.** Hasta 20 tweets sin incluir los retweets.
- **Lista de usuarios a los que sigue el usuario.** Lista de amigos.
- **Lista de usuarios que siguen al usuario.** Lista de followers.
- **Timeline del usuario.** Tweets publicados por los usuarios seguidos.
- **Últimos mensajes directos recibidos por el usuario.**
- **Últimos mensajes directos enviados por el usuario.**
- **Lista de usuarios de los que el usuario no quiere recibir retweets.**
- **Relación entre el usuario y una lista (máximo 100) de usuarios.** Esta relación puede ser: “ambos usuarios se siguen entre sí”, “un usuario sigue al otro”, o “no existe ninguna relación”
- **Detalles de configuración de usuario.** Ejemplos de estos detalles son el idioma en el que tiene configurada la interfaz o la zona horaria en la que se encuentra.
- **Lista de usuarios que el usuario ha bloqueado.** Bloquear a un usuario implica dejar de seguir a ese usuario, y fuerza al usuario bloqueado a que deje de seguir al que ha hecho el bloqueo, siempre que sea posible. Además, Twitter no permitirá que el usuario bloqueado vuelva a seguir al usuario que le bloqueó.
- ***Listas a las que está suscrito el usuario.*** Definimos como una lista a un grupo de usuarios creado por algún otro usuario de la red para organizarlos, y poder ver en el timeline de la lista todos sus tweets. Dependiendo de la visibilidad, una lista puede ser pública (visible en toda la red) o privada (visible sólo para el usuario que la creó).

- ***Lista de usuarios recomendados.*** A partir de la clasificación que hace Twitter internamente de cada usuario, Twitter recomienda al usuario otras cuentas a las que puede interesarle seguir. Twitter no revela los criterios usados por este sistema de recomendación.

2.5.6.2. Peticiones que reciben como entrada una lista de usuarios

- ***Timeline de los tweets publicados por cualquier usuario perteneciente a la lista.***
- ***Comprobación de si un usuario determinado está suscrito a la lista.***

2.5.6.3. Peticiones que toman como entrada un tweet

- **Los últimos retweets del tweet especificado.** Hasta un máximo de 100.
- **Los últimos usuarios que han retweetado el tweet especificado.** Hasta un máximo de 100.

2.5.6.4. Peticiones que reciben como entrada una consulta

Consultas orientadas en la búsqueda de tweets. Las posibilidades para la construcción de este tipo de consultas son las siguientes, pudiéndose además combinar los criterios de búsqueda.

- ***Buscar tweets dependiendo del contenido.*** Es posible definir una lista de términos que queremos que aparezcan en el texto del tweet, cadenas exactas de texto, hashtags, nombres de usuarios o enlaces a direcciones de internet. Además, también se puede indicar la actitud en el mensaje diferenciando entre mensajes con una actitud positiva y mensajes con actitud negativa. Este último punto implica un procesamiento interno de lenguaje natural transparente a nosotros.
- ***Buscar tweets dependiendo de la fecha de publicación.*** Es posible definir un rango de fechas para determinar tweets que hayan sido publicados antes, durante o después de esas fechas.
- ***Buscar tweets dependiendo del lenguaje en el que están publicados.*** El sistema que sigue Twitter para clasificar el lenguaje de los tweets es generalizar a partir del lenguaje que el usuario ha marcado como predeterminado para mostrar la interfaz de la página web twitter.com.

Consultas orientadas en la búsqueda de usuarios. La única posibilidad que nos da Twitter es indicar una cadena de texto simple, sobre la que no se permite ningún tipo de operación.

Consultas de ámbito general.

- **Usuarios recomendados para una categoría predeterminada.** Estas categorías están definidas previamente por Twitter y las denomina slugs. De nuevo, no tenemos información de cómo ha establecido las categorías, ni cómo funciona el sistema de recomendación.
- **Información conocida sobre un área geográfica.** El tipo de información que podemos encontrar es, principalmente, a qué país corresponde al área definida, y si se trata de una ciudad o de "terreno abierto".
- **Los 10 trending topics en una localización concreta.** Es posible solicitar las tendencias en un área geográfica concreta, usando identificadores. Estos identificadores son o de país o de ciudad (sólo están soportadas las más importantes).
- **Lista de localizaciones de las que Twitter tiene información relativa a algún trending topic.** Esta consulta puede entenderse como la inversa del punto anterior "top-10 de trending topics para una localización concreta".

2.6. HERRAMIENTAS DE SOFTWARE UTILIZADAS

Se utilizara para este propósito las siguientes herramientas de software:

2.6.1. Librería Twitter4j de Java

La aplicación desarrollada necesita acceder a la información de Twitter para recolectar los tweets que posteriormente se analizan y clasifican. Para ello se ha utilizado una librería gratuita y open source llamada Twitter4j[6] escrita en Java. Aunque es una librería no oficial de Twitter, se presenta en la documentación para desarrolladores[7] existente en su web. Ésta permite realizar de una forma sencilla la integración de nuestra aplicación con la API de Twitter. La versión utilizada ha sido la 4.0.1.

Twitter utiliza el protocolo de autorización OAuth (Open Authorization). Este protocolo abierto permite que una aplicación externa acceda a Twitter en nombre de un usuario sin que ésta conozca las credenciales de su cuenta. De este modo hace de llave de entrada a Twitter de

[6] http://twitter4j.org/en/index.html
[7] https://dev.twitter.com/docs/twitter-libraries

una forma segura para el usuario, ya que su contraseña no necesita ser conocida por un tercero para actuar en su nombre. Twitter4j permite compatibilidad con OAuth.

Se ha creado una cuenta de Twitter que será la que se utilice para acceder y recolectar los tweets. Por otro lado, es necesario registrar la aplicación en Twitter[8] para que cuando nos conectemos sepa qué aplicación es la que está intentando acceder a sus datos. Para ello, al realizar el registro nos asigna un par de identificadores llamados consumer key y consumer secret que identifica inequívocamente la aplicación. Además, obtenemos el access token y el access token secret, que permiten realizar peticiones a la API de Twitter en nombre de la cuenta creada. Estos identificadores son necesarios para realizar la conexión con la API de Twitter y recolectar los tweets.

Una vez realizada la conexión a la API se pueden realizar búsquedas de tweets por diferentes criterios. Nuestro sistema realiza las búsquedas utilizando el nombre de usuario de las cuentas oficiales de los periódicos. Para ello, hay que tener además en cuenta que Twitter impone unas restricciones[9] en cuanto al número de peticiones que pueden hacerse a su API en un determinado periodo y en cuanto a la antigüedad[10] de los tweets obtenidos.

2.6.2. WEKA plataforma de aprendizaje automático y minería de datos

WEKA (Waikato Environment for Knowledge Analysis) es un software gratuito de aprendizaje automático y minería de datos distribuido bajo licencia GNU-GPL. Es una plataforma desarrollada en Java por la Universidad de Waikato, en Nueva Zelanda.

Weka contiene una buena colección de algoritmos para tareas de minería de datos y modelado predictivo. Éstos pueden aplicarse directamente a un conjunto de datos a través de su interfaz gráfica o bien pueden ser llamados desde un programa externo a través de las API proporcionadas. En este trabajo se ha tomado la segunda opción, concretamente se ha utlizado la API de Java, ya que es el lenguaje que se ha utilizado para programar todo el sistema.

Weka incluye herramientas para el preprocesamiento de los datos (filtros), clasificación (árboles, tablas), clustering, reglas de asociación, y adicionalmente, diversas formas de

[8] https://apps.twitter.com/
[9] https://support.twitter.com/articles/160385
[10] https://dev.twitter.com/docs/api/1.1/get/statuses/user_timeline

visualización de los datos, tanto en el inicio del proceso de carga de datos, como después de haber aplicado un algoritmo.

En este proyecto WEKA ha sido utilizado para generar el clasificador que determina la polaridad de un tweet, es decir, lo clasificará como positivo, negativo, neutro o ninguna de las categorías anteriores. En la Fugura 13imagen se puede ver su interfaz gráfica.

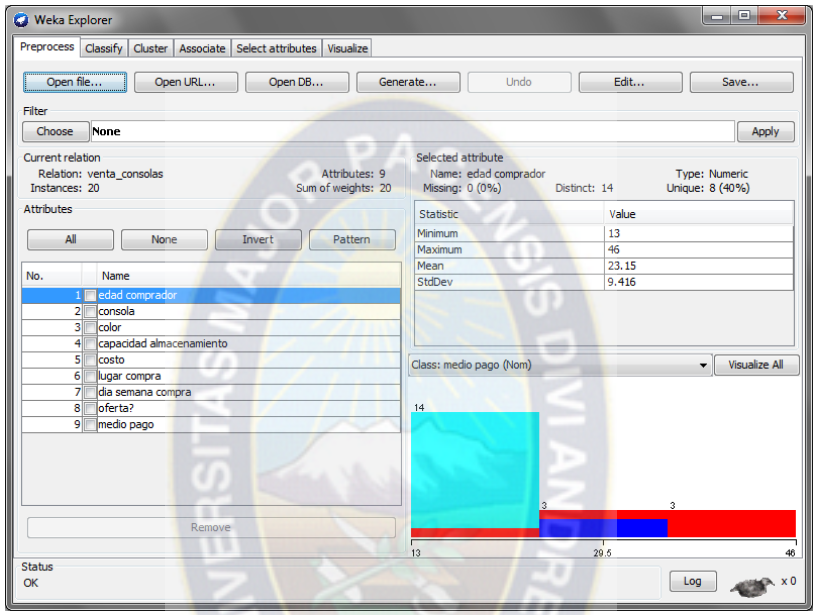


Figura 2.12. Interfaz gráfica del software WEKA

Fuente: Weka (sf)

3. DISEÑO METODOLÓGICO

En esta sección se da una explicación detallada sobre el proceso de diseño del modelo que consiste en realizarlo bajo el estudio de minería de contenido. Se describe su funcionamiento incluyendo los distintos componentes y decisiones que son fundamentales para alcanzar la propuesta de modelo.

En la búsqueda de poder detectar tendencias y/o preferencias a través de información de lo que se está discutiendo en Twitter es necesario enfocarse en la recuperación de información y de tweets debido a su gran valor para el presente modelo, además de los datos que rodea al momento de que un usuario publica un tweet.

Por lo tanto este capítulo consiste en dar respuesta a las cuestiones que plantean las distintas funcionalidades introducidas en el capítulo 1, de manera que se obtenga un modelo más cercano a la implementación, en la Figura 3.1 proponemos el esquema de modelo.

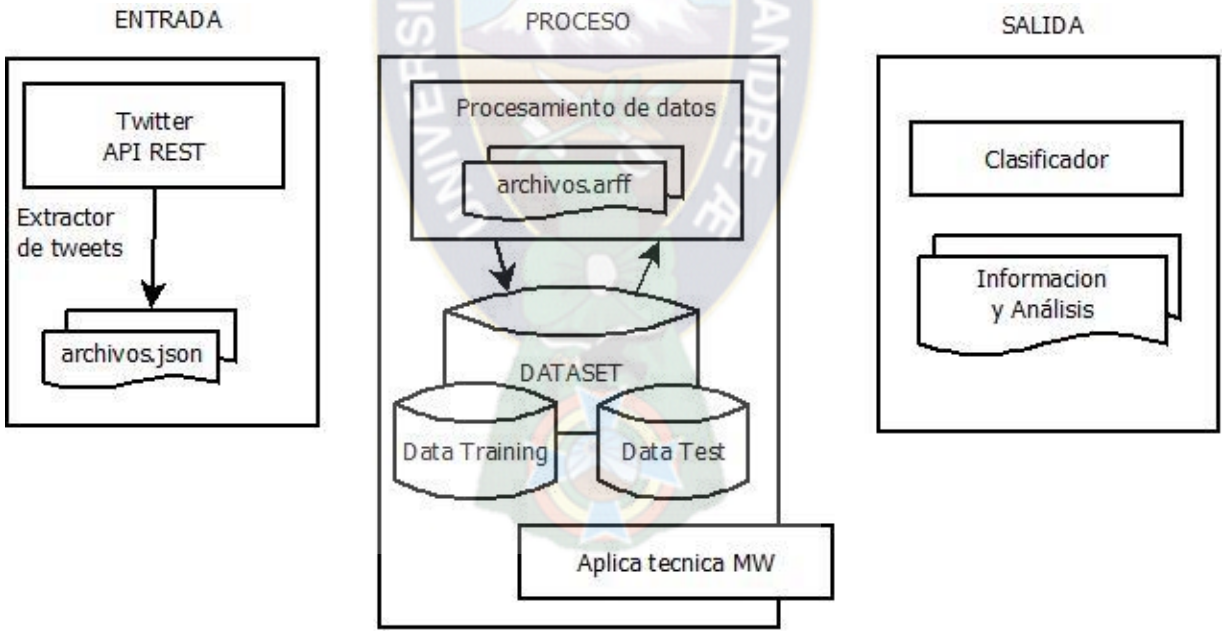


Figura 3.1. Esquema del Modelo de extracción y análisis de información de la red social Twitter

Fuente: Elaboración propia

Entrada: la entrada está compuesta de toda la información recolectada para el análisis de datos, para el caso de entrada está representada por archivos en formato JSON obtenidos

mediante la librería twitterJ4 a través del Api Rest de Twitter, la cual nos permitió recolectar muestras de tweets.

Proceso: es el procesamiento de los datos es el proceso de minería web de acuerdo a las etapas descritas en el sector 2.2.4, básicamente entendida en cuatro fases: recuperación de información, extracción y procesamiento, generalización y analisis. Logrando obtener nuestra base final de datos para la construcción del clasificador, el cual se realizara un minado aplicando técnicas y teorías de minería de datos.

Salida: consiste en los resultados obtenidos del minado, obtenidos a través del clasificador de contenido, determinando las preferencias actuales de los usuarios de Twitter.

3.1 ESTRUCTURA

Para la implementación del modelo se ha optado por dividir el problema en distintas fases ya definidas por la metodología de MW, de tal manera que se pueden distinguir cuatro fases principales, cada uno de ellos dedicado a resolver una tarea específica.

La primera etapa de recuperación de información se basa en la puesta a disposición del investigador de la información sin ningún tratamiento. La segunda etapa se centra en la extracción de la información a partir de la información adquirida en la primera etapa. Estos datos se depuran y se configura la información para su procesamiento. La tercera etapa está vinculada a la generalización de esa información, lo que supone llevar a cabo la conversión de la información obtenida en las etapas anteriores para su estandarización, manejabilidad y equiparación. Por último, en la cuarta fase se lleva a cabo el análisis de las muestras que se realiza gracias a diferentes herramientas y que permiten la obtención de conclusiones a partir de los datos obtenidos como resultado.

En la Figura 3.2 se puede apreciar la estructura de nuestra aplicación que se divide en siete módulos (todos ellos relacionados de alguna manera con la base de datos creada):

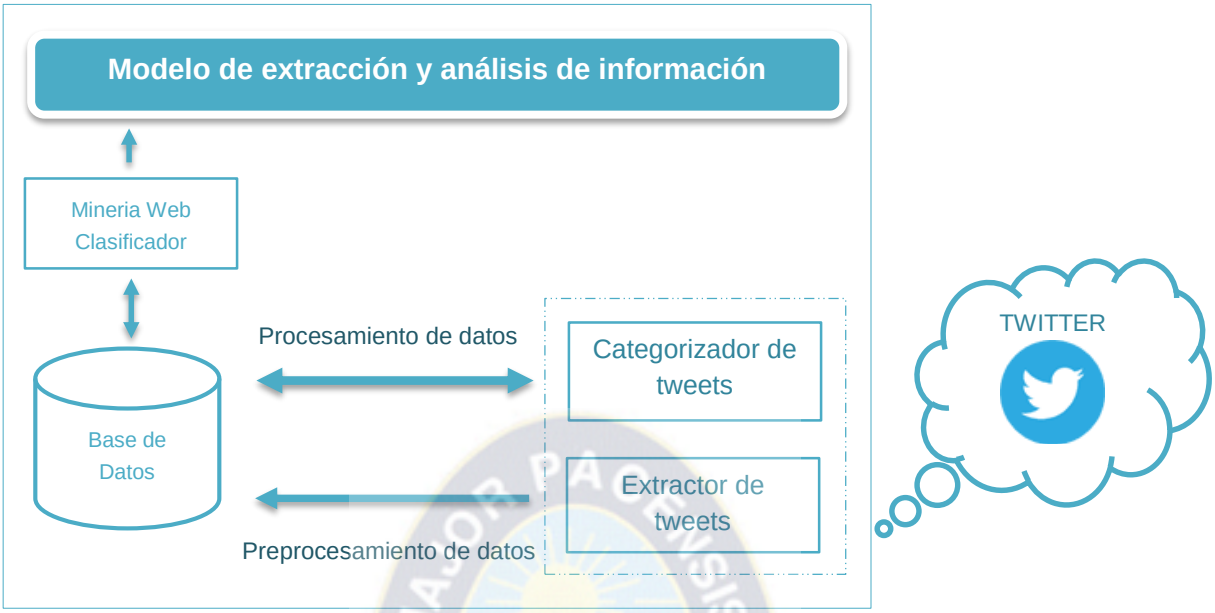


Figura 3.2. Estructura del modelo

Fuente: Elaboración Propia

3.2. RECUPERACIÓN DE INFORMACIÓN

Como ya se mencionó en el anterior capítulo, acerca de la información accesible de Twitter de forma pública, indicando que existen dos formas de consumir dicha información: la API REST y la Streaming API, debido a que el objetivo de este trabajo es obtener y almacenar información sobre twitter de manera estructurada, se estudia lo que brinda el API Rest.

Es importante destacar que la API Rest proporciona acceso a la información ya existente en Twitter en el momento de hacer la llamada. De esta manera, es importante resaltar que al recibir una respuesta por parte de Twitter, se obtiene la representación formal de la información o modelo definido por Twitter completo.

El primer paso a la hora de comenzar el desarrollo del proceso encargado de la extracción de datos desde Twitter es autenticarse en la API, para ello es necesario seguir el proceso descrito en el *Anexo 1* “Proceso de autentificación en la API de Twitter”. En este anexo se detallan tanto los pasos necesarios para obtener las claves de autenticación necesarias, como el código necesario para autenticarse por medio de la librería Twitter4J utilizando dichas claves.

Una vez autenticados en la API de Twitter, la librería Twitter4J pone a nuestra disposición varios métodos que nos permitirán buscar en Twitter a través de su API. A continuación se muestra la implementación de este proceso:

Diseño del modelo de datos

Para cumplir nuestro objetivo de extraer y analizar información es necesario definir un modelo mediante el cual seamos capaces tanto de desgranar la información presente en lo que nos envía Twitter como de almacenarla para su posterior uso. Por otro lado, para ser capaces de permitir la visualización de la información, nuestro modelo debe estar dividido en entidades de relevancia similar, de forma que sea sencillo extraer información de él.

Observando el análisis realizado en el marco teórico llegamos a la conclusión de que nuestro modelo de datos debe centrarse en los tweets, ya que analizando la información que estos contienen tendremos disponible información de todas las categorías que se han definido previamente (contenido, social, interés y temporalidad).

Una vez decidido que nuestro enfoque se basa en los tweets, nuestro siguiente paso ha sido definir cómo desgranamos la información en diferentes entidades dentro del modelo para poder guardarlas de forma independiente. De esta forma se da la posibilidad de consultarlas mediante nuestra API Rest. Las entidades que finalmente componen el modelo de este proyecto son las siguientes:

1. **Tweet.** Datos de un mensaje de texto escrito por un usuario en Twitter. Contiene los mismos datos que el original. En un principio se planificó guardar solo la información relativa al contenido que íbamos a usar, pero más adelante nos dimos cuenta que podría ser interesante tener todos los datos, ya que podríamos extraer más información a posteriori si fuese necesario
2. **Usuario.** Datos de un usuario de Twitter, en nuestro caso de cuentas oficiales de medios de comunicación de los que se quiere obtener la información.
3. **Hashtag.** Datos de un hashtag de Twitter, entendiendo por ello las palabras de un tweet que comienzan por #. Contiene la palabra en sí.

Es importante destacar que con este modelo nos asegurarnos de que cubrimos cada una de las categorías que se definieron en esta fase del modelo: los tweets aportan contenidos

actuales y los hashtags de interés. De esta forma el modelo propuesto cubre las necesidades de nuestro proyecto así como está preparado para dar soporte a futuras ampliaciones.

3.3. EXTRACCIÓN Y PREPROCESAMIENTO

El fin de esta fase es la recolección y el almacenaje de los datos. Uno de los requisitos fundamentales es que debe ser capaz de comunicarse con la REST API de Twitter, puesto que los datos solo pueden obtenerse a través de ella.

El primer paso para el correcto desarrollo del proyecto es conseguir la extracción de información. En esta fase se debe obtener y almacenar tweets de manera estructurada, Para ello, se apoya en el API Rest de Twitter.

3.3.1. Extracción de tweets

Una de las partes fundamentales en el proyecto es la extracción de los datos necesarios desde de Twitter. La interacción con esta plataforma se realiza a través de una API proporcionada por dicha plataforma, es decir, ofrece un conjunto de funciones y métodos para ser utilizados por otro software agregando una capa de abstracción sobre los mismos. Para comenzar a recoger la información, por cada usuario tratado viene dada en los siguientes puntos diferentes:

- La información del perfil del usuario, dado que uno de sus campos se corresponde con la ubicación.
- Los n-últimos tweets publicados por cada usuario.

A continuación los nombres de cuentas de medios de comunicación Seleccionados para este estudio

Cuentas	CantidadTweets
pagina_siete	529
LaRazon_Bolivia	692
LosTiemposBol	529
ATBNoticias	529

Figura 3.3. Nombre de usuario en Twitter de medios de comunicación locales

Fuente: Elaboración propia

La información recogida a través de las peticiones a la API Rest va acompañada de otros datos que no serán de utilidad para este caso de estudio. Este proceso no filtrará los datos no útiles, sino que los almacenará en bruto tal y como vienen dados por la respuesta de la API en formato JSON. Todos los datos se almacenarán en ficheros de salida.

3.3.1.1. Autenticación a través de OAuth

La API Rest funciona con peticiones de tipo GET y POST, por lo que es necesario establecer una comunicación HTTP desde la aplicación hacia la API. Puesto que se desea recolectar la información de la forma más rápida posible y debido a las restricciones que impone el propio API en cuanto al número de peticiones por hora, la autenticación a través de OAuth permite que este límite sea menos restrictivo, por lo que la aplicación necesitará autenticarse.

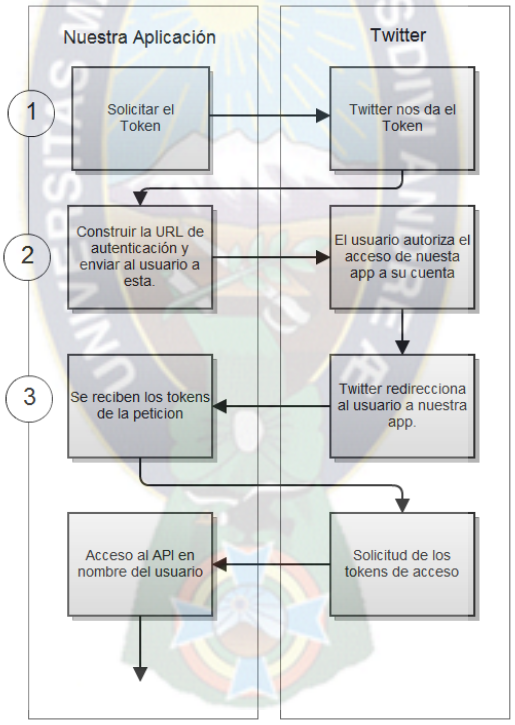


Figura 3.4. Diagrama de flujo de la autenticación vía OAuth v1.0

Fuente: Elaboración Propia

Además, la autenticación permite el acceso a perfiles de usuarios de Twitter que hayan sido bloqueados contra usuarios no autenticados, por lo que gracias a la autenticación se disminuirá el número de respuestas de tipo "no autorizado" obtenidas. Como parte de la autenticación

siguiendo este protocolo, son necesarias unas claves (o tokens) para firmar las peticiones (en la ilustración se muestran sombreadas por cuestiones de seguridad).

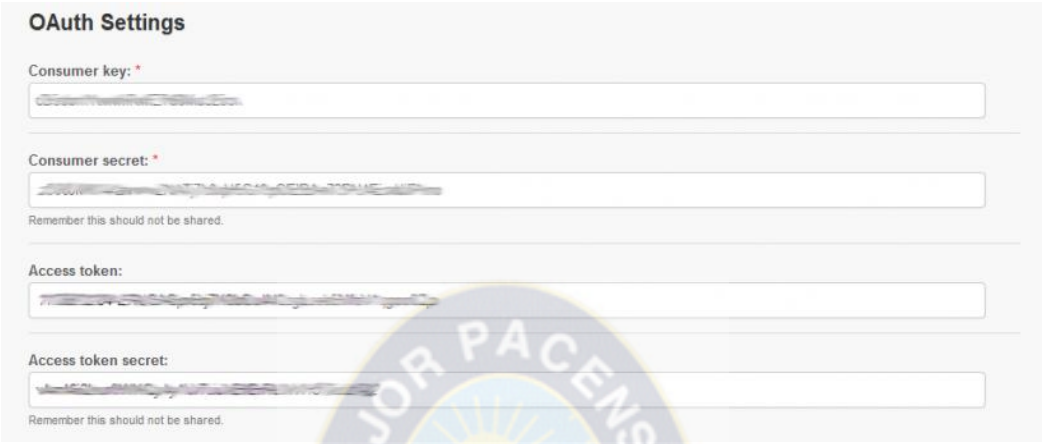


Figura 3.5. Consumer key y Consumer secret para una aplicación de Twitter

Fuente: OAuth, sf

Finalmente es necesario incluir estas claves en nuestra aplicación al comienzo de la clase encargada de extraer la información desde Twitter a través de la librería Twitter4j para así poder utilizar la API de Twitter.

3.3.1.2. Recolección de tweets y almacenaje de los datos

Esta fase se encarga de recolectar tweets al azar y de cuentas oficiales de medios de comunicación escogidos. Para ello se utiliza la librería twitter4j, una librería escrita en Java que permite fácilmente integrar una aplicación Java con Twitter, lo podemos descargar desde la siguiente dirección: http://twitter4j.org/en/index.html

Al tener todo preparado para hacer consultas al servicio Rest haciendo uso de la API 1.1 de Twitter. Debemos tener en cuenta que la contestación de las respuestas se lleva a cabo en formato JSON. Vamos a ver cuáles son las posibilidades que tenemos accediendo a la documentación de la misma[11].

[11] https://dev.twitter.com/docs/api/1.1

```
1  public void Tweet() throws TwitterException{
2        Twitter twitter;
3         ConfigurationBuilder cb = new ConfigurationBuilder();
4         cb.setDebugEnabled(true)
5                 .setOAuthConsumerKey("Consumer Key")
6                 .setOAuthConsumerSecret("Consumer Secret")
7                 .setOAuthAccessToken("Access Token")
8                 .setOAuthAccessTokenSecret("Access Token Secret");
9         twitter = new TwitterFactory(cb.build()).getInstance();
10
11        Paging pagina = new Paging();
12
13        Status tweetEscrito = twitter.updateStatus(Mensaje);
14
15    }
```

Figura 3.6. Autenticación en la API de Twitter a través de la librería Twitter4J

Fuente: Elaboración propia

Los mensajes o status son guardados por el sistema en ficheros JSON. Se genera un fichero por cada uno de los usuarios de Twitter indicados en el fichero de configuración.

El fichero encargado de capturar la información de Twitter lanzará los métodos de captura e inserción de tweets discontinuamente, ya es importante mencionar que debido a las restricciones impuestas por Twitter no pueden ser constantes (180 peticiones cada 15 minutos)[12], a continuación se muestra el número total individual de tweets para el análisis de nuestro estudio.

Tras almacenar en objetos los tweets capturados desde la API de Twitter, estos se introducirán en la de base de datos. En el método implementado encargado de insertar los tweets en la base de datos destaca la captura de excepciones provocadas por una violación de la clave primaria, en cuyo caso se realiza una actualización de la información asociada almacenada (en lugar de la inserción).

A continuación se muestra un ejemplo de uno de estos ficheros JSON.

[12] https://dev.twitter.com/docs/using-search

```
{
    "created_at": "Thu Nov 05 15:35:33 +0000 2015",
    "id": 662292172474097700,
    "id_str": "662292172474097664",
    "text": "#LTahora Alberto Gonzales promulgó ley de #referendo2016:
    Es un momento hermoso, somos unos favorecidos por la historia.",
    "source": "<a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>",
    "truncated": false,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,

    "user": {"id": 94438031,
    "id_str": "94438031",
    "name": "Los Tiempos",
    "screen_name": "LosTiempos Bol",
    "location": "Cochabamba",
    "description": "Prensa/Noticias/Publicaciones",
    "url": "http://t.co/oPS9IcaWxC",
    "entities": {
    "url": { "urls"...},
    "favorited": false,
    "retweeted": false,
    "lang": "es"
}
```


Figura 3.7. Fichero JSON con los tweets recolectados

Fuente: Elaboración propia

Para cada tweet se trata la información y se recogen los atributos que serán necesarios para implementar nuestro modelo, para almacenar dichos atributos se ha creado una clase Java (“Tweet”) cuyas variables coincidirán con los campos de nuestra base de datos. Aunque se obtiene mucha información relacionada con el tweet, la única información que se guarda en la base de datos de cada uno es la siguiente:

TWEET
Id_tweet. Tweet RTs que tiene el tweet FAVs que tiene el tweet Nombre_Usuario Followers del usuario Followings del usuario

Figura 3.8. Atributos de tweet almacenado

Fuente: Elaboración Propia

3.3.2. Preprocesamiento de tweets

Este proceso es el centro del modelo y es el más complejo. Puesto que definiremos los ficheros ***.arff*** para el estudio correspondiente y de test para crear el clasificador que encontrará los tweets de los medios de comunicación más populares recolectados por la fase explicada en la anterior sección.

Utiliza diferentes librerías para realizar sus funciones, entre las que se encuentran:

- WEKA: Para crear el clasificador y evaluarlo. Además, transforma los tweets recolectados en un formato reconocido por el clasificador para realizar la predicción de tendencias.
- MySQL-Connector: Necesaria para conectar y trabajar sobre la base de datos MySQL.

3.2.2.1. Transformación de Hashtag y links

Por este motivo se ha realizado un tratamiento de la información existente en la base de datos. Este tratamiento ha consistido en realizar búsquedas por palabras clave en el contenido del tweet que pudiesen ayudar a catalogarlo dentro de alguno de los temas de actualidad política ya comentados. De este modo, se han conseguido etiquetar alrededor de 1.000 mensajes que no contenían hashtag,

Por ejemplo, el siguiente tweet no contiene ningún hashtag pero está relacionado con el caso Economía:

> *Tweeet: "Evo Morales reitera el ofrecimiento de seguridad jurídica a inversionistas https://t.co/mcVfp1Yxdr https://t.co/Dazejvrbxa".*

Ante un caso como el ejemplo anterior, se decide asignar un valor manual al campo hashtag de la tabla Tweets de la base de datos. Para ello se ejecuta una sentencia SQL como la siguiente:

```
UPDATE tweets SET hashtag="#Economia" WHERE hashtag IS NULL AND
texto LIKE"%inversion%";
```

De este modo, se consigue disminuir el número de tweets que no contienen hashtag y que no podemos relacionar con uno de los tópicos a analizar. Algunas de las palabras claves que se

han utilizado en las búsquedas para relacionar tweets con los tópicos tratados, han sido por ejemplo:

Tópico: Encuentro restos arqueológicos en la avenida Bush. Palabras clave:

- restos + arqueológicos
- hallazgo + arqueológicos

Aparte de este problema que ha sido tratado, se han detectado casos en los que sobre un mismo tema existen diferentes hashtags que lo hacen referencia. Por ejemplo:

Tópico: Declaraciones de Evo contra Ministra de Salud. Hashtags relacionados:

- #Evo
- #EvoMorales

Por este motivo se ha decidido agrupar en un mismo tema varios hashtags que lo hacen referencia.

Topic	Hashtag
1-E	#Evo
1-E	#EvoMorales

Tabla 3.1. Contenido parcial de la tabla Tópicos de la BBDD

Fuente: Elaboración Propia

Se tuvo que realizar transformaciones a algunas de las variables escogidas anteriormente, para modificarlas y crear otras nuevas a partir de ellas. Las modificaciones realizadas fueron:

Tweets - links, para realizar esta categorización de tweets se asigna con un valor de {0,1} por cada contenido de tweet, estableciendo si el tweet contiene un link o dirección URL asignamos el valor **1** caso contrario se le asigna **0**.

Tweets - hashtag, para realizar esta categorización de tweets se asigna con un valor de {0,1} por cada contenido de tweet, estableciendo si el tweet contiene algún hashtag asignamos el valor **1** caso contrario se le asigna **0**.

Tweets	Contenido	Valor
con links	*"Evo a ministra de Salud: No quiero pensar que es lesbiana https://t.co/jFCXuRSlos https://t.co/QO59xmh7x5"*	**1**

sin links	*#Evo: En Chile algunos conservadores siguen con la razón o la fuerza. Ahora estamos con la fuerza de la razón*	**0**
con hashtag	*#Evo pide a Chile reparación de daño histórico con base al diálogo*	**1**
sin hashtag	*Evo Morales le pide a Chile una propuesta escrita sobre el diálogo marítimo https://t.co/JmKBAYyKm8*	**0**

Tabla 3.2. Categorización contenido de tweets

Fuente: Elaboración propia

3.2.2.2. Categorización de tweets por contenido

Uno de los objetivos del proyecto es el procesamiento de tweets para obtener información procesada. En este punto se trata la categorización de tweets por contenido. Definido un conjunto de categorías, nuestro objetivo ahora es clasificar cada tweet en una de las categorías de dicho conjunto basándonos en su contenido.

Este apartado trata de cómo se ha abordado el problema de la categorización de textos cortos mostrando no sólo la solución final adoptada, sino la evolución y las decisiones adoptadas en el proceso de desarrollo. A rasgos generales, los pasos que se han seguido para incorporar la categorización de textos a nuestro proyecto han sido los siguientes:

a) Se ha creado un conjunto de entrenamiento que sirva como base para clasificar textos cortos.
b) Se han implementado dos algoritmos categorizadores que se basan en herramientas especializadas en la recuperación de información.
c) Se ha implementado un sistema evaluador que fuese capaz de determinar la precisión de los categorizares creados partiendo de un conjunto de tweets clasificados manualmente.
d) Se ha realizado un análisis de los resultados obtenidos para determinar la versión del categorizador de textos que se ajusta mejor a las necesidades del proyecto.

A cada registro de la base de datos que representa un tweet, se le añade la categorizacion, de modo que al final de este proceso la base de datos tiene una única tabla de datos llamada Tweets con los campos que se indican en el siguiente cuadro.

Campo	Descripción
Id	Id del tweet

Usuario	Nombre de usuario que publico el tweet
Texto	Contenido del mensaje del tweet
RTs	Número de Re-tweets del tweet publicado
FAVs	Número de favorites del tweet publicado
Followers	Número de amigos
Followings	Número de seguidores
Link	Si contiene o no un texto tipo Link. Sólo guarda el primero un valor (0,1)
Hashtag	Si contiene o no un texto de tipo hashtag. Sólo guarda el valor (0,1)

Tabla 3.3. Definición de la tabla que contiene los tweets en para el análisis

Fuente: Elaboración propia

3.2.2.3. Obtención del conjunto de entrenamiento

Los sistemas basados en la recuperación de la información requieren ser entrenados. Para poder entrenar un sistema de categorización de textos, es necesario obtener un conjunto de textos clasificados por categorías. A este conjunto de entrenamiento le llamamos corpus. Escoger un corpus de textos hace plantea la siguiente cuestión: ¿Qué textos aportarían un contenido representativo de términos que son mencionados en Twitter? Simplificando en una frase más concisa: ¿Sobre qué se habla en Twitter?

Para tratar de encontrar los temas de conversación de interés, obviando todas aquellas conversaciones banales o de carácter privado, se examinó el trending topic del 17 de noviembre del presente año, fecha que corresponde a la semana de extracción de tweets para el estudio correspondiente, revisando la página web Trendsmap[13] la cual nos proporciona información de los trending topic tanto a nivel mundial como a nivel de ciudad local.

Trending topics principales el 17 de noviembre del 2015:

- Bolivia
- Tecnología
- Hallazgo
- EvoMorales
- Anf
- Arqueologicos

[13] www.trendsmap.com

Figura 3.9. Mapa de Trending Topic de la ciudad de La Paz en noviembre de 2015

Fuente: Trendsmap (sf)

Buscando el motivo de estos trending topics, aparece el tema acerca de los hallazgos de restos arqueológicos por la Empresa @MiTeleferico, además de la repercusión de las declaraciones del presidente Evo morales hacia la ministra de Salud y otros temas de los cuales los ciudadanos de la ciudad de La Paz hace mención en esta red social.

Por cada medio de comunicación, se detalla a continuación el ranking de los 10 hashtags que más ha nombrado y el número de veces que lo ha hecho. También se muestra el número de tweets que no contienen ningún hashtag, que como se puede ver, representa aproximadamente un 80% de los tweets totales.

ATB Noticias	
Sin hashtag	1009
#LoMásVisto	39
#Ahora	26
#VamosBolivia	14
#Bolivia	14
#ElAlto	14
#LaPaz	13
#Último	12
#ATBNoticias	12
#DeportesATBRadio	12
#MarParaBolivia	11

Los Tiempos	
Sin hashtag	170
#LaPaz	3
#Bolivia	128
#EliminatoriasRusia2018	87
#Ultimo	13
#París	12
#Paraguay	34
#Venezuela	65
#Cochabamba	35
#Francia	17
#LTahora	41

.

Tabla 3.0.4. Ranking de hashtag citados por los periódicos ATB Noticias y Los Tiempos

Fuente Elaboración Propia

Página Siete	
Sin hashtag	161
#Último	90
#Nacional	76
#Sociedad	58
#Economía	36
#Campeones	30
#Evo	26
#Portada	13
#Página7	10
#Bolivia	7
#CosaSeria	6

La Razón	
Sin hashtag	172
#EvoMorales	41
#EnDirecto	108
#Mundo	35
#LoÚltimo	69
#LaPaz	35
#Bolivia	46
#ultimo	10
#Francia	21
#ElAlto	10
#Marcas	157

Tabla 3.5. Ranking de hashtags citados por los periódicos Página Siete y La Razón

Fuente Elaboración Propia

Esta información, junto con un estudio más exhaustivo de los mensajes que no contenían ningún hashtag, ha sido utilizada para decidir los temas de actualidad a tratar en este trabajo:

- Atentados en Francia
- Declaraciones de Evo contra Ministra de Salud
- Partidos de Bolivia rumbo a Eliminatorias Rusia 2018
- Encuentro restos arqueológicos en la avenida Bush
- Pago de doble aguinaldo
- Construcción del Centro Nuclear

En las tablas anteriores, se puede observar que además de existir un gran número de mensajes sin hashtag en su contenido, se encuentran muchos tweets que contienen hashtags que no ayudan a identificar el tópico que se trata en el mensaje. Por ejemplo: #Ultimo, #Ahora, #EnDirecto, #LoUltimo, etc.

Viendo estos ejemplos, ¿es posible hablar de un tema que une todo, o al menos la inmensa mayoría del contenido de interés generado en Twitter? Sí, de actualidad, ya sea actualidad política, deportiva, cultural etc. Se determina que el contenido que se genera en Twitter, aparte de las conversaciones privadas o la conversación sin sentido, guarda una relación directa con algún tema que o está ocurriendo, o ha ocurrido en el mundo en un periodo de tiempo muy cercano.

Una vez que se han elegido los textos que van a conformar el conjunto de entrenamiento, se define el conjunto de categorías que se va a tratar para la categorización. Inicialmente se planteó la siguiente cuestión. ¿Es posible cubrir la totalidad del contenido que se genera en Twitter? Dado que el proyecto no se centra exclusivamente en la categorización de textos sino que es sólo una de las áreas en las que se quiere trabajar, se consideró que no será necesaria una categorización exhaustiva de todo el contenido generado.

En un principio, se optó por manejar cuatro categorías diferentes: actualidad política, actualidad economía, actualidad cultural y actualidad deportiva. Esta idea tenía su base en que es, a grandes rasgos, la división de los medios de prensa donde la información es extraída. Se percibió que al categorizar textos cortos de 140 caracteres como máximo había muchas posibilidades de que fuese muy difícil discernir, incluso para un ser humano, entre un tweet relacionado con la actualidad económica y un tweet relacionado con la actualidad política. Los siguientes dos tweets de ejemplo (Figuras 3.10 y 3.11) corresponden a personas ajenas totalmente al proyecto y pueden servir para ilustrar esta situación:

Figura 3.10. Tweet de ejemplo del 24 de marzo del 2015 por @carlosdmesag

Fuente: cuenta oficial de Carlos de Mesa G. en Twitter

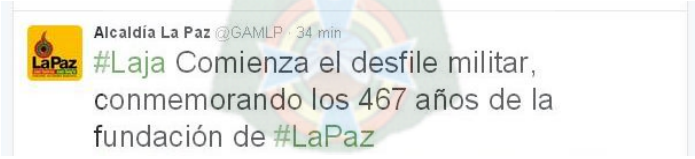


Figura 3.11. Tweet de ejemplo del 20 de octubre del 2015 por @GAMLP

Fuente: cuenta oficial de Alcaldía La Paz en Twitter

Tweets que tienen un fuerte contenido político y a la vez están tratando sobre temas cívicos. Se decidió que esta división no iba a dar categorías completamente excluyentes. Por otro lado, analizando los tweets etiquetados como culturales fue fácil notar que habría muchas posibilidades de que el porcentaje de tweets etiquetados como cultura fuese extremadamente bajo.

La solución adoptada fue reducir las categorías a dos grandes grupos bien diferenciados: actualidad política y actualidad deportiva. Estas dos categorías no abarcan todo el universo, porque un tweet podrá no pertenecer a ninguna de estas categorías. Para simplificar, todo lo que quede fuera de estas dos categorías pertenece a una tercera categoría que fue denominado contenido no categorizable o ruido.

La Figura 3.12 muestra, en forma de diagrama de conjuntos, qué conjunto de tweets se desea categorizar. Sea U el universo de temas de conversación y N el subconjunto de tweets que son calificados interesantes, U - N es todo el contenido calificado como ruido. Los dos subconjuntos de N utilizados son P (contenido relacionado con la actualidad política) y D (contenido relacionado con la actualidad deportiva). Se considera que la intersección de estos dos conjuntos es vacía. El sistema determinará dos conjuntos P' (contenido categorizado como actualidad política) y D' (contenido categorizado como actualidad deportiva). El objetivo será acercar lo máximo posible P' a P y D' a D, minimizando el número de falsos positivos y falsos negativos.

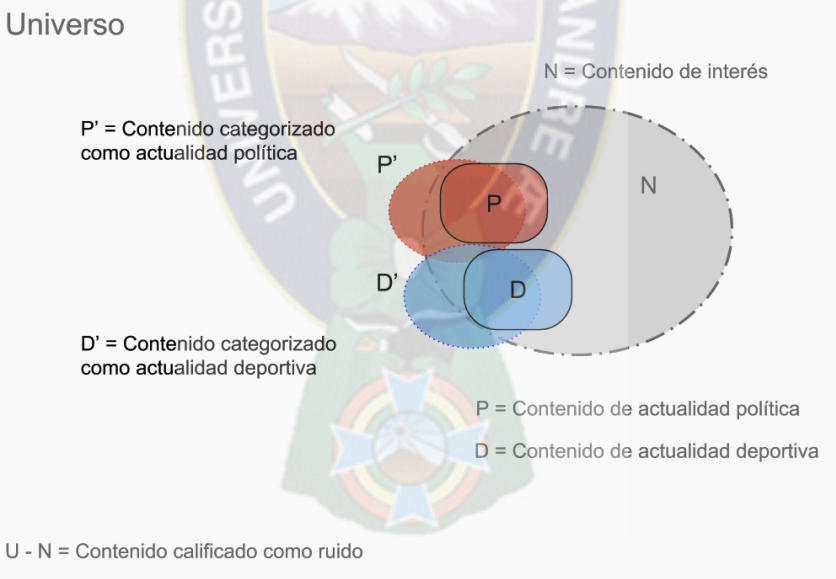


Figura 3.12. Situación abstracta de la categorización del contenido publicado en Twitter

Fuente: Elaboración propia

3.4. GENERALIZACIÓN

Para poder sacar ideas, reconocer patrones, y probar los algoritmos desarrollados, nuestro primer paso fue obtener una muestra de tweets. Definimos como muestra de usuarios a una colección de datos de tweets reales que publican en español, extraídos de Twitter.

Además se ha intentado que la muestra fuese lo más homogénea posible, extrayendo tweets a diversas horas para tener datos tanto de Bolivia como de América Latina y de un tamaño suficientemente representativo.

Para tomar la muestra se ha aprovechado que en el proyecto ya teníamos implementada la conexión a Twitter REST. Anteriormente era posible ya se podía obtener tweets con las características de palabras o lenguaje que quisiéramos, por lo que para obtener la muestra lo lanzamos con un filtro de lenguaje en español para así recibir sólo aquellos tweets escritos en español en ese momento.

La API de Twitter está diseñada de forma que cualquier tweet que se extraiga siempre contenga información básica del usuario que ha creado ese tweet, o lo ha retuiteado, por lo que cada vez que extraíamos un tweet nosotros podíamos almacenar en nuestra base de datos los datos referentes al usuario que nos interesaba. De esta forma, conseguimos tener una muestra de 3000 tweets con los datos:

- usuario
- texto
- followers
- followings
- retweets
- favoritos
- ***link***
- ***hashtag***
- ***clase***

Una vez obtenida toda la información sobre cada uno de los tweets, se procede a generar un fichero con extensión ARFF (Attribute Relation Format File), formato reconocido por la herramienta WEKA. En este fichero se indican todos los atributos existentes para cada una de los tweets (instancias), y se añade además valor (clase) como último atributo.

```
% Datos para pruebas de tweets (tesis 2015)
% Universidad Mayor de San Andres(Bolivia-LaPaz)

@relation Categorizador
@attribute usuario {'ATBNoticias', 'LaRazonDigital','Páginasiete','LosTiempos','otro'}
@attribute texto string
@attribute followers NUMERIC
@attribute Follows NUMERIC
@attribute Retweets NUMERIC
@attribute Favorites NUMERIC
@attribute link{0,1}
@attribute hashtag{0,1}
@attribute clase{0,1}

@data
'otro','#MarParaBolivia',0,0,0,0,0,1,1
'otro','#EvoMorales',0,0,0,0,0,1,1
'otro','#Referendo2016',0,0,0,0,0,1,1
'otro','#EliminatoriasRusia2018',0,0,0,0,0,1,1
'otro','Deporte',0,0,0,0,0,0,1
'otro','Gol',0,0,0,0,0,0,1
'ATBNoticias','Pobladores aseguran que continúa disminuyendo el nivel de agua del lago
Poopó ',111426,84,1,1,1,0,0
'ATBNoticias','Colapsa el alcantarillado de dos zonas en #Chuquisaca e inunda un
mercado ',111426,84,0,2,1,1,0
'LaRazonDigital','Más de un millón de peruanos padece diabetes, según el Ministerio de
Salud https://t.co/m5NVoFnEDU https://t.co/8VbS8x29Vh',158484,229,2,1,1,0,0
```

Figura 3.13. Ejemplo de fichero ARFF

Fuente: Elaboración propia

En la figura 3.13 se puede ver un ejemplo de un fichero ARFF. Cada tweet es representado por una lista separada por comas. El Primer elemento de la lista es el nombre de usuarios (ATBNoticias, LaRazonDigital, Páginasiete, LosTiempos, Otro). El segundo elemento de la lista es el contenido reducido del tweet (tipo de dato string), y el resto de elementos de tipo numérico representan los valores que tienen cada uno de los atributos generados por nuestro analizador. El último valor de la lista es el valor categorizado del tweet y únicamente puede contener los valores {0,1} por lo anteriormente explicado en el sector *3.2.2.3.*

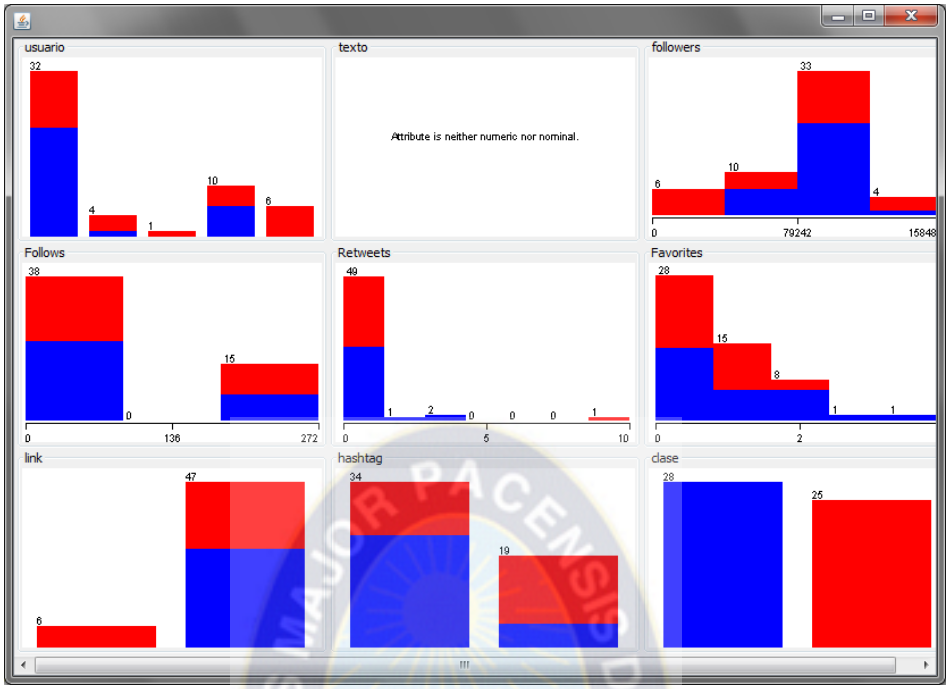


Figura 3.14. Visualización de información del archivo categorizador.arff

Fuente; Elaboración propia

3.4.1. Categorizador de textos basado en Weka

El primer acercamiento para implementar la categorización de textos cortos emplea la herramienta Weka. El objetivo es conseguir un módulo, absolutamente independiente, en el que se pudiese construir un modelo de Weka a partir de un corpus de entrenamiento conseguido de manera transparente. Una vez entrenado el sistema, se puede preguntar a qué categoría pertenece un texto.

Las respuestas dadas por el sistema pueden ser de tipo: indirecta, en la que obtendremos un vector de valoraciones entre 0 y 1 que representa el grado de pertenencia del texto a cada categoría:

- El texto pertenece a la actualidad política.
- El texto pertenece a la actualidad deportiva.

En el módulo de Weka construido se ofrece estas tres diferentes alternativas para construir el modelo interno.

- Definir de manera dinámica (en código) los datos en forma de tuplas (texto, clase).

- Leer un contenido de texto con extensión de texto que contiene los datos.
- Leer N contenidos de texto que contienen los datos. Internamente se añaden los datos de todos ellos al modelo final. Esta alternativa fue la escogida para realizar las pruebas de precisión realizadas que se detallarán más adelante.

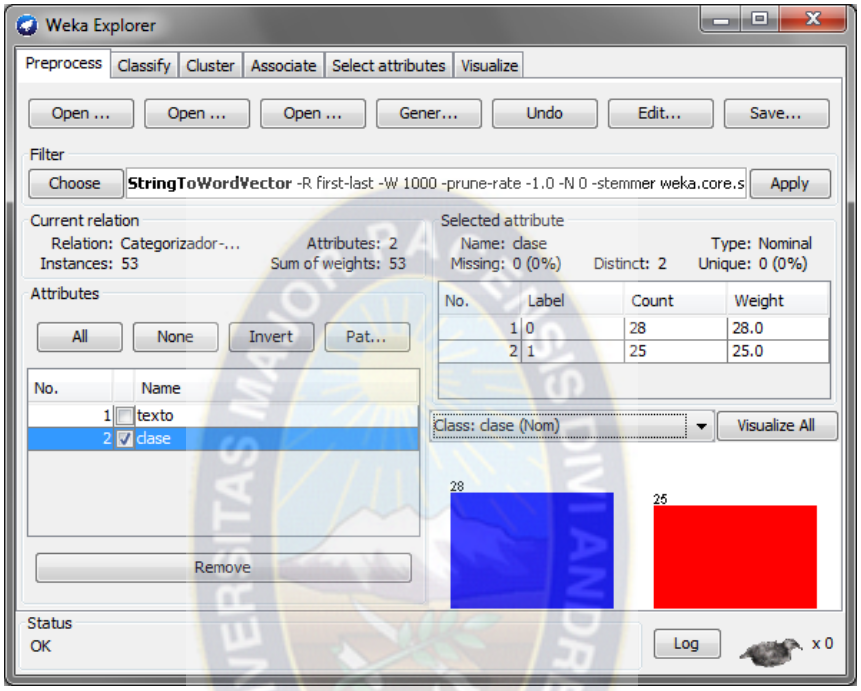


Figura 3.15. Visualización de información del archivo categorizador.arff

Fuente: elaboración propia

3.4.1. Generación de instancias

Un vez construido el modelo usando los corpus creados ya es posible de realizar la categorización de los tweets. Al pedir la clasificación de un texto, se le pide a Weka que lo clasifique no con uno, sino con tres clasificadores diferentes. El hecho de usar hasta tres clasificadores de Weka apenas representa coste en ejecución a la hora de clasificar; no así a la hora de construir los modelos a partir del corpus de palabras. Sin embargo, el coste en tiempo al construir los índices no afecta al rendimiento del sistema, ya que ese cálculo sólo hay que hacerlo una vez al iniciar el sistema.

Con las instancias generadas correspondientes al corpus de entrenamiento se procede a la creación del clasificador. Concretamente, se utiliza la técnica de clasificación que WEKA utiliza para el estudio y análisis de texto.

Una vez creado, se realiza una evaluación del mismo y se obtiene un informe detallado en el cual se distinguen tres apartados:

- **Resumen:** porcentaje global de aciertos y errores cometidos en la evaluación.
- ***Precisión detallada por clase:*** para cada uno de los dos posibles valores que puede ser predicho por el clasificador, se muestra el porcentaje de instancias correctamente predichas (TP: True positives) y el porcentaje de instancias con otros valores que son incorrectamente predichas a ese valor (FP: False positives).

Además, se muestran ciertas medidas que tienen relación con las anteriormente expuestas.

Matriz de confusión: es una matriz que muestra de una forma detallada para cada clase el número de instancias predichas. Tiene dimensiones NxN, donde N es el número de los posibles valores que puede tomar la clase. En este caso los valores son *1=pertenece al tema de actualidad determinado* y *0= no pertenece al tema de actualidad determinado*, y por tanto se obtiene una matriz de 2x2.

A continuación se muestra el resultado de la evaluación obtenida por el clasificador creado.

{categoría 1: valor 1, categoría 2: valor 2, …, categoría i-ésima: valor i}

Figura 3.16. Vector obtenido por el clasificador de textos

Fuente: Elaboración propia

Al concluir la clasificación disponemos de tres vectores, uno para cada clasificador, de valores diferentes de la forma que se puede ver en la Figura 26 Concretamente, se convertirá el campo de tipo String llamado texto que guarda el contenido reducido del tweet, en una serie de atributos. Además, la representación de los tweets o instancias pasa de ser una lista de valores, a convertirse en un vector. Este proceso se realiza utilizando el filtro StringToWordVector ofrecido por WEKA.

El algoritmo vuelve a generar otro fichero ARFF con el resultado de aplicar los filtros. En la siguiente figura se muestra el efecto de aplicar los filtros mencionados sobre el fichero de la Figura 3.17.

Para determinar los pesos finales asociados a cada categoría, se calcula la media aritmética de los **tres vectores, obteniendo de esta forma el vector solución.**

```
@relation'Rel-weka.filters.unsupervised.attribute.StringToWordVector
@attribute abrazo numeric
@attribute amigo numeric
@attribute bueno numeric
@attribute encantar numeric
@attribute gracia numeric
@attribute grande numeric
@attribute grandeza numeric
@attribute ser numeric
@attribute corrupto numeric
@attribute no numeric
@attribute numEmoPos numeric
@attribute numEmoNeg numeric
@attribute numWordPos numeric
@attribute numWordNeg numeric
@attribute numInterPos numeric
@attribute numInterNeg numeric
@attribute numUppCase numeric
@attribute numVerb numeric
@attribute numNoun numeric
@attribute numAdj numeric
@attribute numAdv numeric
@attribute numInterj numeric
@attribute classPolarity {P,N,NEU,NONE}
@data
{4 1,12 1,18 3}
{8 1,9 1,13 2,17 5,18 6,19 1,20 1,22 N}
{3 1,10 1,11 1,12 1,17 2,18 2}
```

Figura 3.17. Ejemplo de fichero ARFF filtrado

Fuente: Elaboración propia

Por último, el modulo implementado para solicitar la categoría asignada por Weka a un texto dispone de dos opciones: una que devuelve el vector solución y otra en la que el propio módulo procesa el vector solución y determina la categoría más probable. Para ello, el módulo compara los valores de cada categoría en el vector solución con su valor umbral predefinido. Si algún valor supera el valor umbral asignado a esa categoría, ésta será la categoría asignada. En caso de haber varias categorías que superen sus umbrales, predomina la de mayor valor.

3.5. ANÁLISIS

La última fase del proceso corresponde al análisis de la información extraída del servicio de red social Twitter. Para que dicha fase pueda llevarse a cabo es necesario que toda la información se encuentre disponible en el entorno local. Esto equivale a que todas las fases previas se hayan desarrollado sin problemas y en particular, que las peticiones de recuperación para cada uno de los periodos de extracción hayan descargado correctamente la información almacenada en nuestro almacén de datos". La información es contenida en ficheros planos: formato JSON para la información extraída del servicio de red social Twitter y texto para los datos de seguimiento de usuarios.

En el análisis de la información pueden distinguirse dos procedimientos, en función de la naturaleza de los datos: el seguimiento de usuarios. Para el análisis del modelo se ha empleado el módulo de weka para la manipulación y el análisis de la información.

En este sector se examina el resultado obtenido por nuestro sistema clasificador. En la evaluación del mismo, que se explica en el sector 3.4, se obtiene un porcentaje de acierto del 97%, por lo tanto sabemos que en sus predicciones existirá un número elevado de errores.

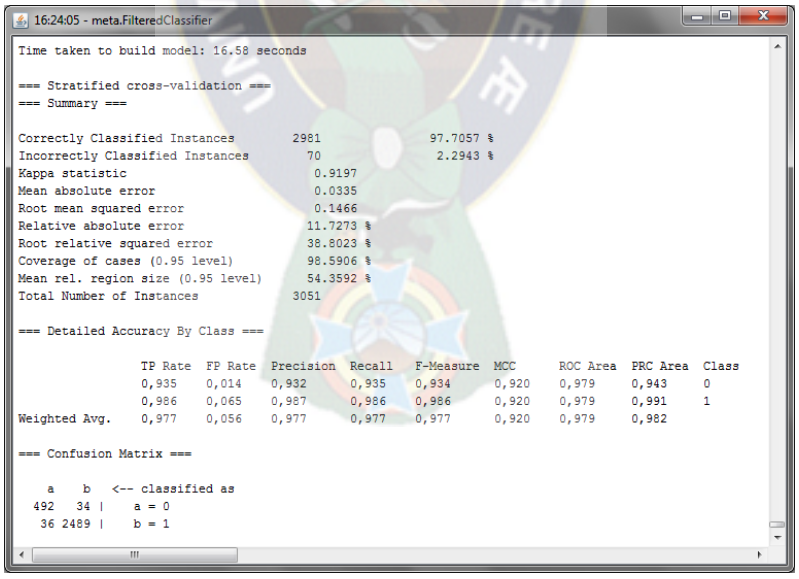


Figura 3.18. Resultado obtenido por nuestro sistema clasificador del modelo

Fuente: Elaboración Propia

Como se puede observar en la figura 3.18, alrededor de un 97% de los tweets son clasificados de forma correcta. Si nos fijamos en las dos categorías existentes, vemos que los mejores resultados se obtienen en la categoría 1 (F-Measure: 0.986) y los peores resultados en la categoría 0 (F-Measure: 0.934), donde únicamente se clasifican correctamente 34 instancias de las 526 existentes.

Analizando la matriz de confusión, averiguamos que a pesar de la cantidad de instancias predichas de forma errónea, para la clase 1 (b) se han clasificado correctamente la mayoría de las instancias, concretamente 2489. Otras 36 instancias que deberían haberse clasificado como la clase 1.

Para la clase Negativo (a) ocurre lo mismo que con la categoría anterior, es decir, la mayoría de las instancias predichas se han realizado de forma correcta (492).

4. EVALUACIÓN DE RESULTADOS

Se presentara a continuación un resumen de los descubrimientos encontrados en el capítulo anterior.

4.1. EVALUACIÓN DE CLASIFICADOR DEL MODELO

Dentro del proceso de minado esta fase se encarga de generar conocimiento, aplicando los algoritmos no supervisados de agrupamiento y asociación se llega aplicar la clasificación como un refinamiento en el análisis. Por lo tanto en el presente trabajo se decidió utilizar:

- La clasificación como tipo de modelo modelo
- El algoritmo J48 que WEKA utiliza para la filtración de texto.

En la tabla 8 se puede ver los atributos del archivo elaborado para la minería web. Este archivo contiene 3070 registros, tomando en cuenta que se recolectaron tweets durante 5 dias. Para entrenar este modelo un archivo con el 100 registro que fueron seleccionados de forma correcta y aleatoria, y el restante otra para pruebas finales.

Atributo	Descripción	Tipo
Id	ID de tweet	Numérico
Usuario	Nombre de usuario	Numérico
Texto	Contenido del mensaje del tweet	String
RTs	Número de Re-tweets del tweet publicado	Numérico
FAVs	Número de favorites del tweet publicado	Numérico
Followers	Número de amigos	Numérico
Followings	Número de seguidores	Numérico
Link	0 sin link 1 link	Nominal
Hashtag	0 sin hashtag 1 hashtag	Nominal
clase	0 pertenece 1 no pertenece	Nominal

Tabla 4.1. Atributos del archivo para el minado web

Fuente: Elaboración Propia

Se eligió al atributo **clase**:

Clase: para entrenar el clasificador y descubrir preferencias actuales de los usuarios.

Por el hecho de que la técnica de clasificación es una técnica de StringToWordVector por lo que convierte los atributos de tipo String en un conjunto de atributos representando la ocurrencia de las palabras del texto.

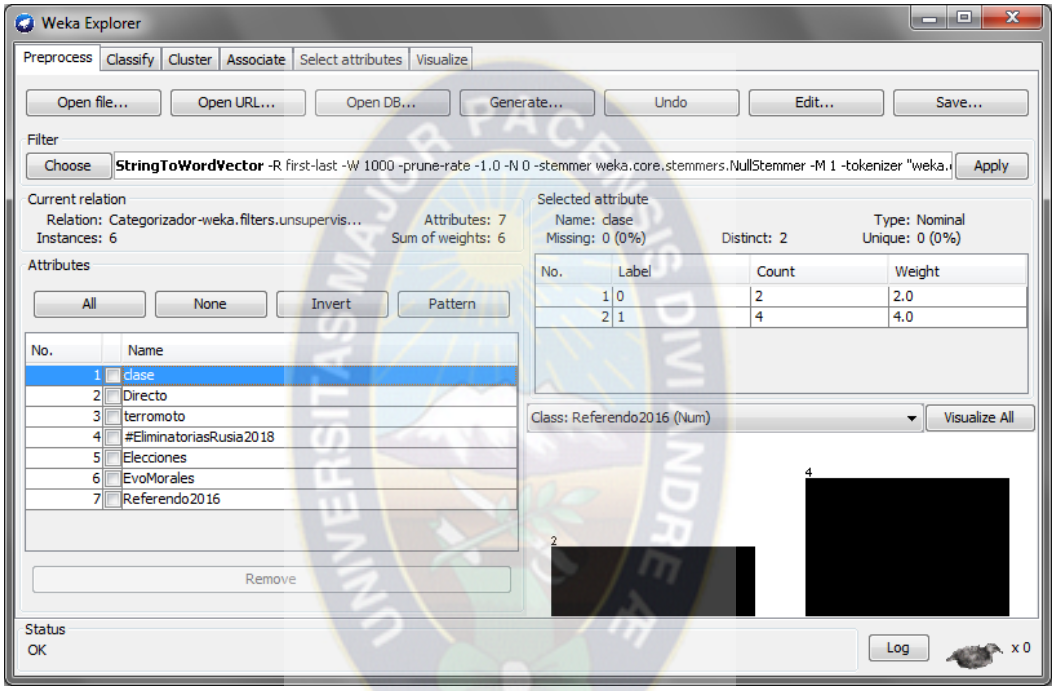


Figura 4. 1. Aplicación de la técnica de StringToWordVector al atributo clase

Fuente: Elaboración Propia

En la Figura 4.1 se puede observar los atributos del archivo elaborado para la minería de datos. Este archivo contiene 15 registros, con la finalidad de entrenarlo con la técnica de StringToWordVector, tomando en cuenta que se obtuvo esta muestra de la base de datos de tweets y la observación del monitoreo de lo que más se hablaba en Twitter en nuestra región (Trending Topic) durante 3 dias continuos para este experimento.

4.2. EVALUACIÓN DE RESULTADOS

Una vez cargado el archivo de estudio en herramienta de minado, vemos que Weka proporciona 4 modos de prueba vemos las opciones que tenemos en la técnica de clasificación:

4.2.1 Supplied test set

Con este algoritmo el fichero de datos con el que se probará el clasificador obtenido con el método de clasificación usado y los datos iniciales (archivo entrenado).

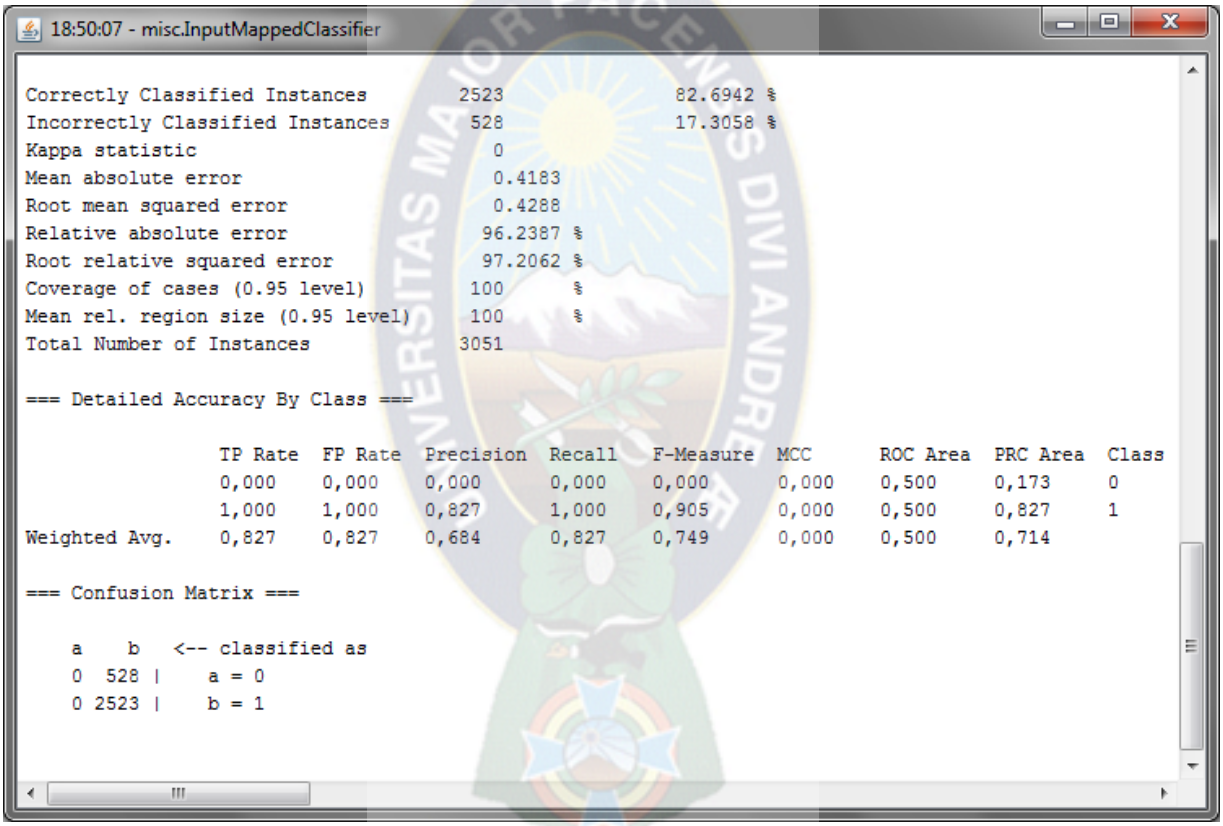


Figura 4. 2. Aplicación de la técnica de Supplied test set al atributo clase

Fuente: Elaboración Propia

Como se puede observar en la figura 4.2 alrededor de un 82% de los tweets son clasificados de forma correcta. Si nos fijamos en las dos categorías existentes, vemos que los mejores resultados se obtienen en la categoría 1 (F-Measure: 0,905) donde se clasifican correctamente las 528 instancias existentes.

Analizando la matriz de confusión, averiguamos que al no haber cantidad de instancias predichas de forma errónea, para la clase 1 (b) se ha clasificado correctamente 2501 instancias.

4.2.2. Use training set

Con esta opción Weka entrenará el método con todos los datos disponibles y luego lo aplicará otra vez sobre los mismos.

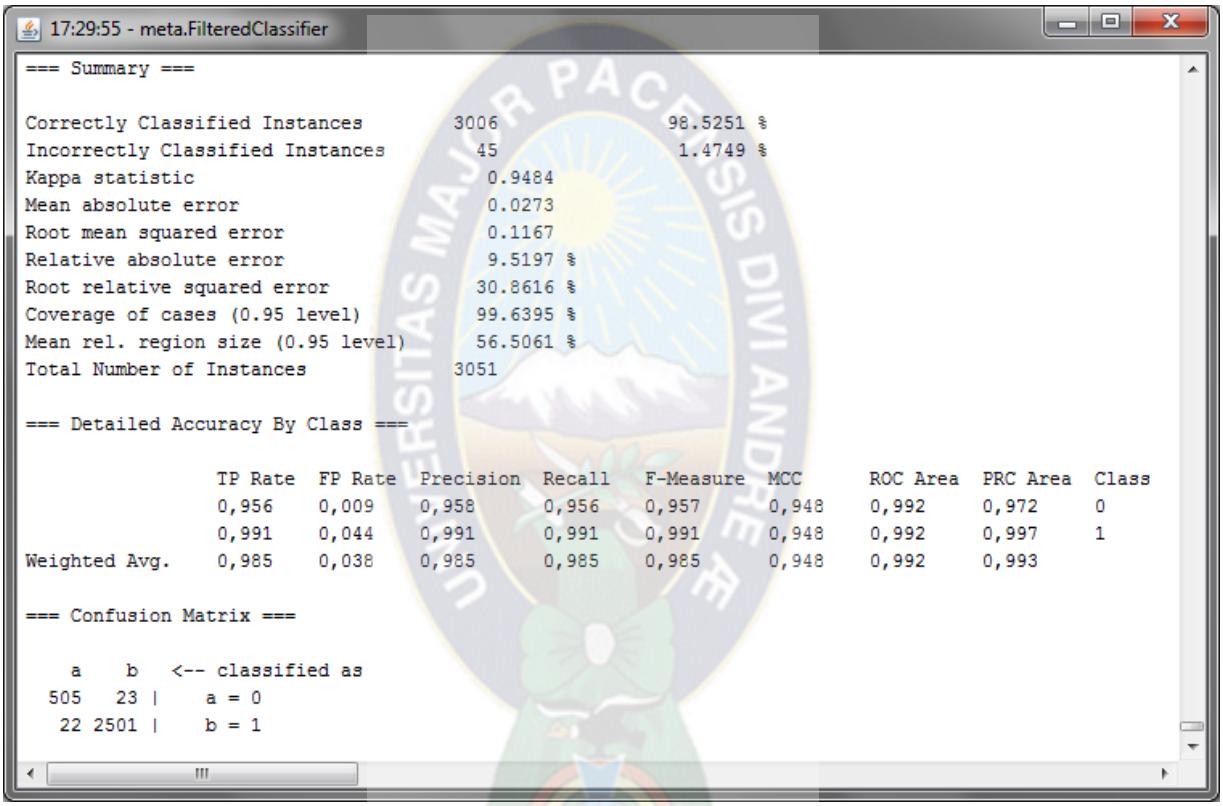


Figura 4. 3. Aplicación de la técnica de Use training set al atributo clase

Fuente: Elaboración Propia

Como se puede observar en la figura 4.3 alrededor de un 98% de los tweets son clasificados de forma correcta. Si nos fijamos en las dos categorías existentes, vemos que los mejores resultados se obtienen en la categoría 1 (F-Measure: 0.991) y los peores resultados en la categoría 0 (F-Measure: 0.957), donde únicamente se clasifican correctamente 23 instancias de las 528 existentes.

Analizando la matriz de confusión, averiguamos que a pesar de la cantidad de instancias predichas de forma errónea, para la clase 1 (b) se han clasificado correctamente la mayoría de las instancias, concretamente 2501. Otras 22 instancias que deberían haberse clasificado como la clase 1.Para la clase 0 (a) ocurre lo mismo que con la categoría anterior, es decir, la mayoría de las instancias predichas se han realizado de forma correcta (505).

4.2.3. Cross-validation

Con esta opción se realiza una validación cruzada estratificada del número de particiones dado (Folds). La validación cruzada consiste en: dado un número n se divide los datos en n partes y, por cada parte, se construye el clasificador con las n-1 partes restantes y se prueba con esa. Así por cada una de las n-particiones.

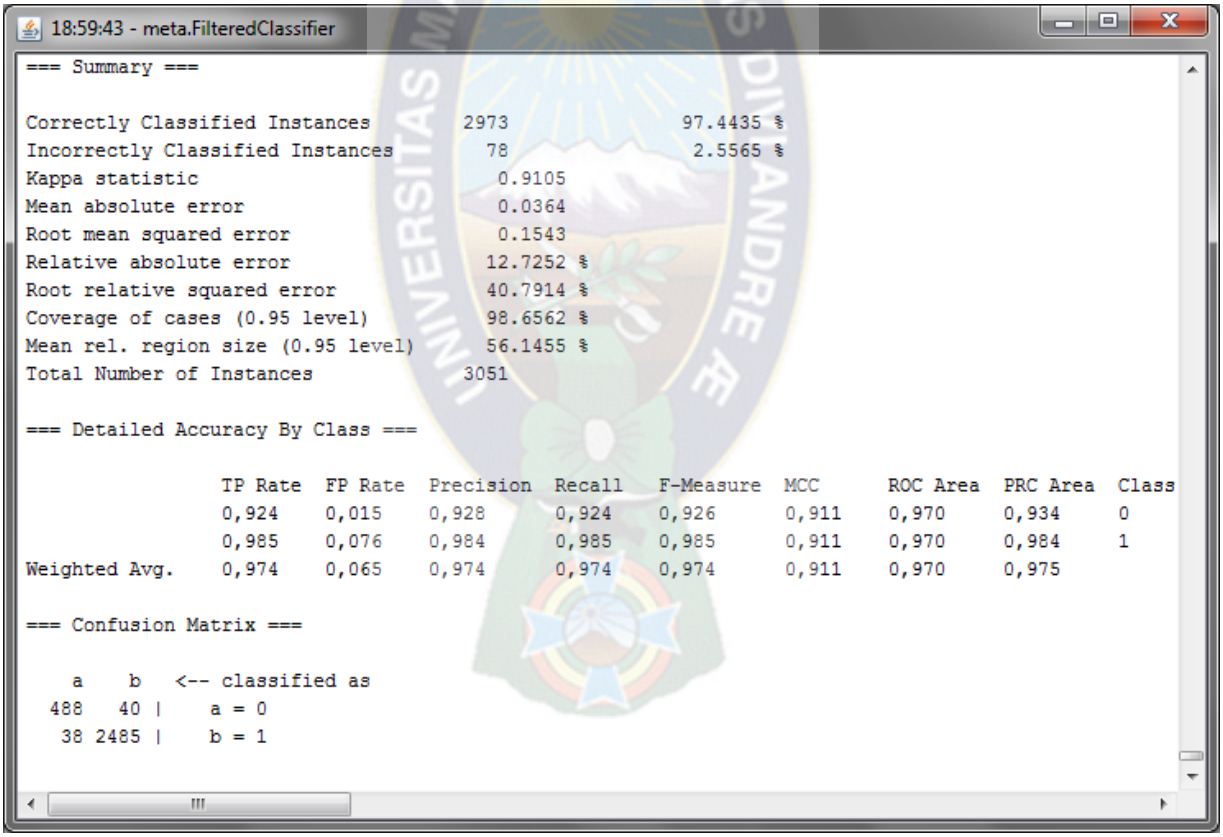


Figura 4. 4. Aplicación de la técnica de Cross-Validation al atributo clase

Fuente: Elaboración Propia

Como se puede observar en la figura 4.4 alrededor de un 97% de los tweets son clasificados de forma correcta. Si nos fijamos en las dos categorías existentes, vemos que los mejores resultados se obtienen en la categoría 1 (F-Measure: 0.985) y los peores resultados en la categoría 0 (F-Measure: 0.926), donde únicamente se clasifican correctamente 40 instancias de las 528 existentes.

Analizando la matriz de confusión, averiguamos que a pesar de la cantidad de instancias predichas de forma errónea, para la clase 1 (b) se han clasificado correctamente la mayoría de las instancias, concretamente 2485. Otras 38 instancias que deberían haberse clasificado como la clase 1.Para la clase 0 (a) ocurre lo mismo que con la categoría anterior, es decir, la mayoría de las instancias predichas se han realizado de forma correcta (488).

5. CONCLUSIONES Y RECOMENDACIONES

Problemas encontrados

Para el tratamiento de los mensajes extraídos de Twitter, se decidió extraer tweets de medios de comunicación y problema inicial acerca de los errores ortográficos disminuye, ya que no suelen realizar faltas de ortografía ni suelen repetirse letras para mostrar énfasis. Sin embargo, surge un problema diferente. Estos usuarios no escriben opiniones personales, sino titulares de noticias en las que el lenguaje suele ser más sobrio y neutro, no añadiendo interjecciones o emoticonos en ellos. Esto hace que algunos de los atributos creados en el clasificador, y que al analizar el corpus de entrenamiento tenían su importancia, dejen de tener peso, no ayudando por tanto a realizar un análisis correcto.

Conclusiones

El primer objetivo específico: "*Realizar un estudio sobre el análisis y procesado de mensajes de la red social Twitter utilizando la API pública disponible"*, este objetivo fue cumplido con éxito por el hecho de que se consigio recolectar los datos necesarios para generar una base de datos de tweets real y de considerable tamaño para un estudio de minería web.

El segundo objetivo específico: *"Obtener tweets de cuentas oficiales de medios de comunicación locales y usuarios particulares para el análisis correspondiente",* este objetivo se cumplió con éxito por el hecho de que se logró establecer una conexión con el API Rest de esta red social con la ayuda de una librería que nos almacenaba los tweets, públicos de medios de comunicación y otros usuario, en archivos JSON, para su posterior estudio.

El tercer objetivo específico: *"Clasificar la información relevante contenida en tweets recolectados en twitter"*, este objetivo se cumplió a cabalidad por el hecho que se logró clasificar la información relevante de un texto de tweet ya se realizó un proceso de depuración de aquellos tweets que: no aportaban con un texto de estudio, los que contenían iconos y emoticonos y mensajes ReTweet. Además de aquellos tweets que hacían mención de una cuenta de twitter, ya que al almacenar en un archivo .arff teníamos un problema de lectura del carácter "@".

El cuarto objetivo específico*: "Realizar un estudio de la información poco estructurada existente en los twwets"*, este objetivo se cumplió por el hecho que se alcanzó a realizar el estudio sobre la información que se genera al publicar un tweet en esta red social.

El quinto objetivo específico: *"Explicar y concluir sobre los resultados obtenidos"*, es objetivo se logró con éxito ya que después de realizar pruebas se logró determinar tendencias y preferencias actuales de la comunidad en Twitter a través de del modelo propuesto.

Conclusión del ***objetivo general***: "*Plantear un modelo de minería web que permita extraer y analizar la información de los tweets publicados por los usuarios de Twitter para conocer sus preferencias actuales locales, a partir de la masiva, dispersa y desestructurada información generada en esta red social".* Se ha alcanzado con el desarrollo del capítulo 3 y 4 de la presente tesis.

Conclusión de la **hipótesis:** "*El modelo de minería web logrará la extracción y análisis de información para conocer los intereses actuales de los usuarios de Twitter, a partir de la masiva, dispersa y poca estructurada información generada en esta red social".*

El modelo es capaz de detectar tendencias y/o preferencias con la técnica de clasificación, a partir de tweets publicados por los usuarios. Para ello se han recolectado alrededor de 3.000 mensajes de la red social Twitter durante el periodo en el que se ha desarrollado el trabajo.

Aunque la herramienta comete errores en sus clasificaciones, se consigue detectar el las preferencias de usuarios y los medios de comunicación han plasmado en los tweets publicados además.

Por otro lado, he tenido que crear un sistema completo que incluye las aplicaciones realizadas en Java, la interfaz web, la conexión con la base de datos MySQL. Y aunque al principio tuve diversos problemas, he sido capaz de ir resolviéndolos por mí misma, de lo cual estoy muy contenta. Desde el punto de vista práctico, también me ha ayudado a adquirir nuevos conocimientos de algunas de la herramienta WEKA que me ayudo a realizar el modelo,

Por otro lado, he vivido de primera mano la generosidad por parte de diferentes grupos de investigación que no conocía previamente y con los que me he puesto en contacto durante la realización del proyecto. En ningún caso han dudado en compartir conmigo algunos de sus

recursos de una forma totalmente desinteresada. En definitiva, ha sido una experiencia de final de carrera muy positiva.

Recomendaciones

Mejorar la precisión del categorizador de tweets de contenido, dados los resultados presentados en el apartado 4.2 se observa que existe un gran margen de mejora en este aspecto. En concreto, lograr determinar y separar los tweets que se pueden calificar como ruido o conversación vacía de los tweets que contienen información de interés.

Ampliación del categorizador de tweets por contenido definiendo y tratando un mayor número de categorías. Un objetivo a largo plazo podría ser cubrir un universo mayor de temas de conversación de actualidad en Twitter, y no solo actualidad deportiva y política.

Por último, mencionar la posibilidad de realizar un trabajo futuro que puede realizarse es el de añadir nuevas categorías a predecir la polaridad de un tweet, como por ejemplo: Positivo, Negativo, Neutro y Ninguno. En este caso, habría que tener en cuenta las estrategias de intensificación y atenuación existentes en gramática. Quizá añadir como atributos el número de retwitteos de un mensaje o el número de hashtags que contiene puede ayudar al clasificador.

BIBLIOGRAFIA

A. Joshi, A. R. (2011). In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations.(pág. 127:132). Oregon, USA: Portland.

Alag, S. (2008). Collective Intelligence in Action. Manning Pubn.

Anguita, M. A. & Lorenzo, R. M. (2014). Extracción, análisis y visualización de información social desde twitter. Proyecto de Sistemas Informáticos (Facultad de Informática, Curso 2013-2014).

Baeza, R. (2004). Excavando la Web. El profesional de la información, v13, n1.

Baeza, R. P. (2005). "Una herramienta de minería de consultas para el diseño del contenido y la estructura de un sitioWeb". Actas del III Taller Nacional de Minería de Datos y Aprendizaje TAMIDA2005, (págs. 39-48).

BarackObama, (sf). http://twitter.com/BarackObama. [Recuperado el 15 de 04 de 2014]

Blockeel, R. K. (2000). Web Mining Research: A Survey (Vol. 2). SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge.

Boyd, D. M. (2008). Social Network Sites: Definition, History, and Scholarship. . Recuperado el 28 de 03 de 2014, de Social Network Sites: Definition, History, and Scholarship. : http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/full

Conecti, (sf) http://conecti.ca [En línea; acceso el 02-septiembre-2015].

Cooley, R.; Mobasher, B.; Srivastava, J. (1997) "Web Mining: Information and Pattern discovery on the World Wide Web". En: ICTAI, pp. 558- 567.

L. & Gutiérrez, I. (2010). Redes sociales y otros tejidos online para conectar personas. Aprendizaje con Redes Sociales. Tejidos educativos en los nuevos entornos. Sevilla: MAD Eduforma

Enterría, A. G. (2012). El análisis de las redes sociales. Revista Española de Ciencia Política. Núm. 30, 121-131.

EsTwitter, (sf) www.estwitter.com [En línea; acceso el 02-septiembre-2015].

Etzioni, O. (1996). "The World Wide Web: Quagmire or gold mine". Commum. ACM.

Fayyad, U. P.-S. (1996). "Advances in knowlegde discovery and data mining". (pág. 6). MIT Press.

FreeLing, (sf). Demo de Freeling: http://nlp.lsi.upc.edu/freeling/demo/demo.php [En línea; acceso el 22-octubre-2015].

Hernández, J. R. (2004). Introducción a la minería de datos. España: Pearson.

Linoff, G. S. (2001). Mining the web. Nueva York: Estados Unidos: John Willey & Sons.

Liu, B. (2007). Web DataMining: Exploring Hyperlinks, Contents and Usage Data, Springer. Springer.

L. D'Monte, "Swine Flu's Tweet Causes Online Flutter." http://www.business-standard.com/article/technology/swine-flu-s-tweet-tweet-causes-online-flutter-109042900097_1.html, 2009. [En línea; acceso el 02-Diciembre-2012].

Moor, H. (2001). Privacy protection, control of information, and privacyenhancing technologies. En H. T. Moor., Privacy protection, control of information, and privacyenhancing technologies. (pág. 1:6). Computers and Society.

Moya, E. (2012). Las Redes Sociales como fuentes de información. Recuperado el 15 de abril de 2014, de www.iuisi.es

Noguera, N. G. (2002). Delitos informáticos en el código penal español. España: Delitos Informáticos.

O'Reilly, T. (sf). Wikipedia. http://es.wikipedia.org/wiki/Tim_O'Reilly [Recuperado el 05 de Abril de 2014],

OAuth, (sf) https://dev.twitter.com/docs/oauth [En línea; acceso el 20 de Agosto de 2015]

Pighin, S. (16 de abril de 2001). Data Mining. Informatica Aplicada a la Ingenieria de Procesos I.

Scotto, M. S. (April 2004). "Managing Web-Based Information". International Conference on Enterprise Information Systems (ICEIS 2004) (págs. 1-3). Portugal: Porto.

Support, T. (sf). Cómo seguir a alguien. ¿Qué es Twitter? Recuperado el 29 de 03 de 2014, de Cómo seguir a alguien. ¿Qué es Twitter? : http://support.twitter.com/groups/31-twitter-basics/topics/108-finding-following-people/articles/108082-c-xf3-mo-seguir-a-alguien

Streaming (sf). https://dev.twitter.com/streaming/overview

Rest (sf) https://dev.twitter.com/rest/public

Twitter. (sf). ¿Cuales son los números de teléfono de Twitter? Recuperado el 28 de 03 de 2014, http://support.twitter.com/articles/

Twitter. (01 de mayo de 2015). Twitter Developers. Obtenido de Parámetros que forman un Tweet.: https://dev.twitter.com/docs/platform-objects/tweets

API Twitter. (sf). https://dev.twitter.com/docs/api [Recuperado el 14 de 04 de 2014, de API]

Twitter, (sf). http://twitter.com/about [En línea; acceso el 20 de 12 de 2014]

Twitter4J, (sf). Twitter4J: http://twitter4j.org [En línea; acceso el 25 de Septiembre de 2015]

Urdaneta, E. (1997). El Data Mining.

Vallejos. (2006). Mineria de Datos.

Warehousing, D. M. (2009). Mladen W. Nadinic.

Web5design, S. a. (2010). Twitt3d. Obtenido de Twitt3d: http://www.twitt3d.com/

Weka (sf) http://www.cs.waikato.ac.nz/ml/weka [En línea; acceso el 15-septiembre-2015].

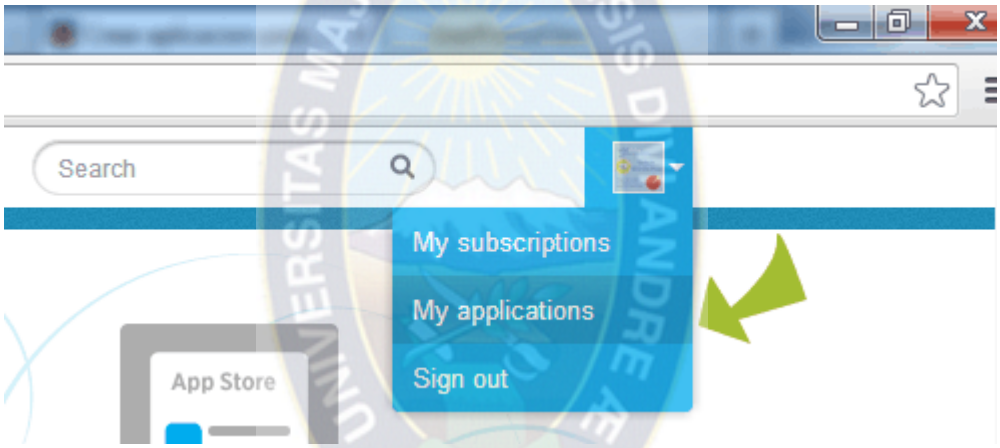
Anexos

Anexo 1

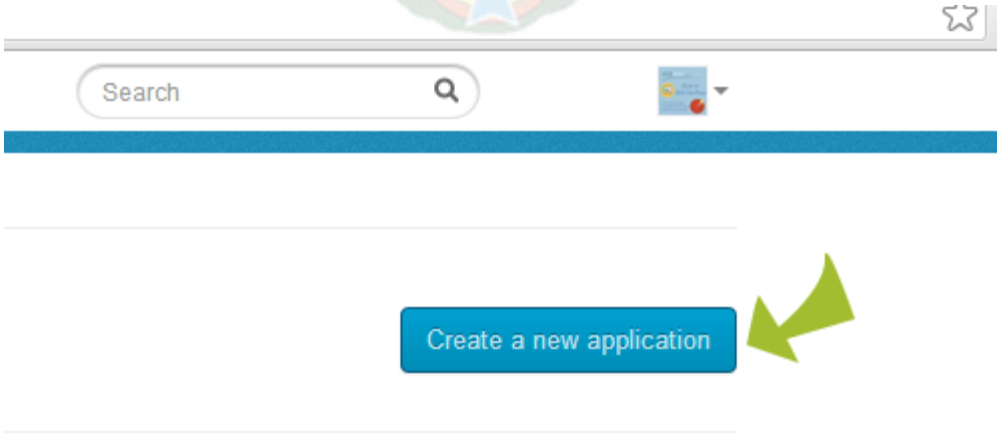
Proceso de autentificación en la API de Twitter

La obtención de tweets se realiza a través de la API de twitter, que da opción a realizar búsquedas de tweets mediante diferentes criterios de búsqueda, sólo necesitamos tener una cuenta en Twitter y acceder a la página de desarrolladores *https://dev.twitter.com* con el usuario y contraseña de nuestra cuenta Twitter.

Una vez dentro en la parte superior derecha aparece el icono de nuestro perfil y al pasar en puntero por encima nos sale un menú con la opción "My Applications".



Este menú nos lleva a *https://dev.twitter.com/apps*, desde donde podemos crear una nueva aplicación pulsando el botón "Create a new application".



Opciones de configuración al crear una aplicación en Twitter

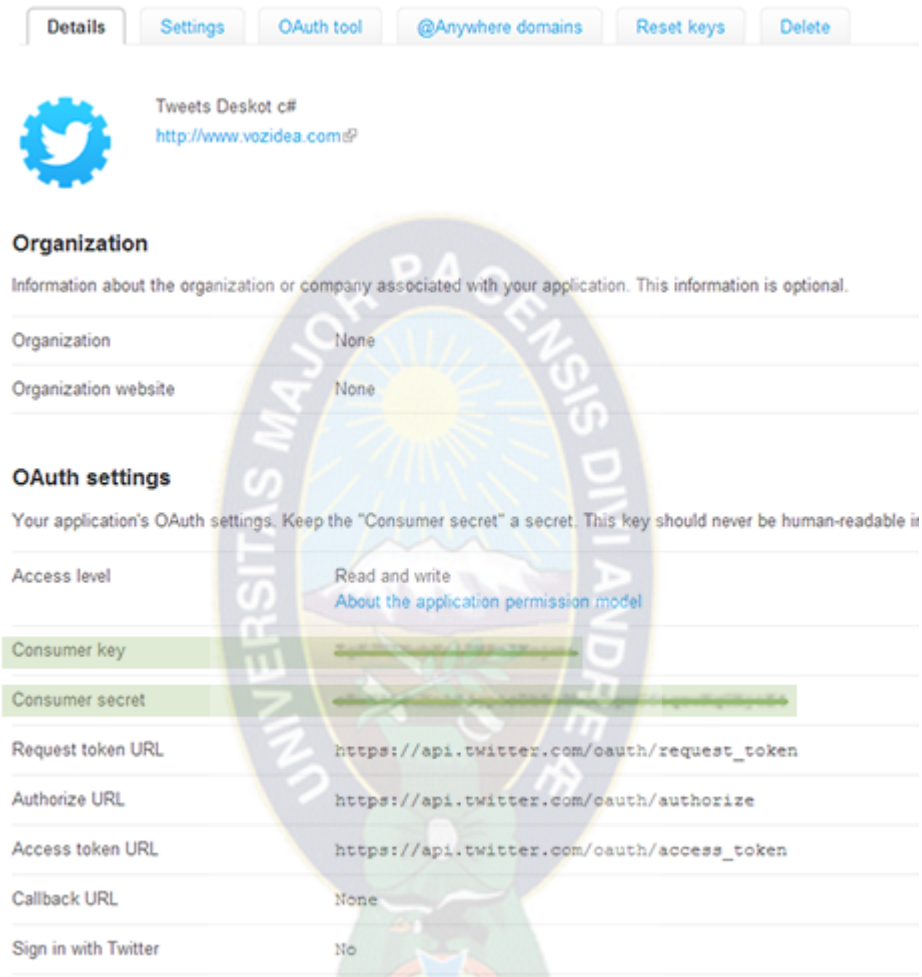
Se nos muestra en pantalla una serie de campos que debemos rellenar para crear la aplicación.

Los diferentes campos son:

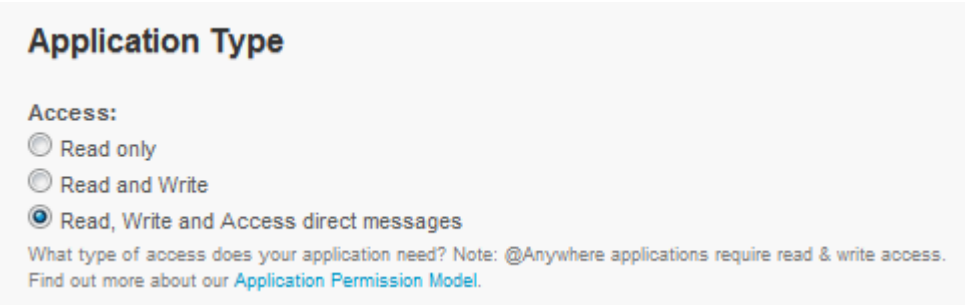
- **Name:** aquí introducimos el nombre de la aplicación. El nombre elegimos el que queramos pero teniendo en cuenta que no puede contener la palabra Twitter.
- **Description:** escribimos una descripción breve de la funcionalidad de la aplicación.
- **Website:** aquí introducimos el sitio web de la aplicación. Sino tiene, podemos escribir una y ya la cambiaremos en el futuro.
- **Callback URL:** si desarrollamos una aplicación web esta opción cobra importancia ya que es la URL a la que se retornas una vez autentificado correctamente en la aplicación. Si lo que estamos desarrollando es una aplicación de escritorio podemos dejarla en blanco.

Por último no debemos olvidarnos de marcar la casilla donde aceptamos las condicione de uso y rellenaremos el campo del captcha que nos soliciten para verificar que somos humanos. Seguidamente pulsamos el botón “Create your Twitter Application” ya tendremos nuestra aplicación Twitter creada.

Con estos pasos ya disponemos de una “Consumer Key” y una “Consumer secret”, que se podrían considerar como un usuario y contraseña que nos da acceso a nuestra cuenta Twitter a través de la aplicación, por eso nunca debemos compartir estos datos con nadie.



Otra opción importante que debemos configurar correctamente son los permisos de acceso que tendrá nuestra aplicación (por defecto sólo permite leer datos de nuestra cuenta Twitter). Los permisos de acceso se muestran en la siguiente captura y se accede desde la pestaña settings:



Si lo que necesitamos es publicar en nuestra cuenta Twitter habrá que darle permisos de lectura y escritura (Read and Write). Si ademas necesitamos enviar mensajes a otras cuentas Twitter entonces habrá que marcar la tercera opción "Read, Write and Access direct messages".

Nos fijaremos que en la pestaña "Details" tenemos un botón que dice "Create my acess token", si lo pulsamos generamos nuestros tokens de acceso y se nos mostraran en pantalla dos nuevas claves de acceso llamadas "Access token" y "Access token secret" que nos serviran para para usar la API de Twitter por ello no debemos compartirlas con nadie.

Obtención tweets para un único usuario

Para obtener tweets de un solo usuario se puede llamar al siguiente método de la API:

https://twitter.com/statuses/user_timeline/id.xml

en donde id es reemplazado por el nombre de usuario o el número identicador de este.

Por ejemplo, para el usuario @MarcAstr0 sería:

https://twitter.com/statuses/user_timeline/MarcAstr0.xml

Este método devuelve los últimos 20 tweets en el timeline del usuario en formato XML, si se quisiera obtenerlos en formato JSON, solo basta con cambiar la extensión e invocar la siguiente URL:

https://twitter.com/statuses/user_timeline/id.json

El timeline se representa como un arreglo de objetos en JSON. En XML, cada tweet se encuentra delimitado dentro de una etiqueta <status>, y el timeline se encuentra dentro de la

etiqueta <statuses>. A continuación se pueden apreciar las diferencias entre los formatos para un mismo tweet:

De toda la información que se devuelve para cada tweet, para este trabajo solo será relevante lo que aparece entre los tags <text>, que corresponde al texto del tweet:

<text>Recien me jugue un raspe y gane $100. Debe ser mi dia de suerte.</text>

Obtención de más tweets y rate limiting

El método recién expuesto muestra los últimos 20 tweets en el timeline del usuario, cantidad insuciente para el análisis que se desea hacer en este estudio. A traves de la glsapi es posible acceder a los últimos 3200 tweets (en caso de tener igual o mayor cantidad de tweets). Para ello se usa el mismo método anterior, pero se le entrega una variable adicional en la URL de la siguiente forma:

https://twitter.com/statuses/user_timeline/id.xml?page=x

Donde x es reemplazado por el número de página. Si el usuario tiene más de 3200 tweets, llegará solamente hasta la página 160. Invocando al mismo método, se puede saber la cantidad total de tweets del usuario que está dada por el tag <statuses_count> y dividiendo por 20 se puede obtener la cantidad de páginas para poder iterar.

Cuando Twitter adquirió suciente popularidad la sobrecarga de llamadas a métodos de la API producidos en sus servidores obligó a instaurar una política que limita la cantidad de llamados a la API, conocida como rate limiting. Actualmente, para la REST y Search API, se permite invocar un máximo de 150 métodos por hora. Si se desea obtener los tweets para un usuario con 3200 tweets, donde se debe invocar al método anterior 160 veces (una vez cada página) se producirá un error HTTP de tipo 400 (Bad Request) antes de obtener todos los tweets. Esta limitante está considerada en la implementación que se describirá más adelante.

Twitter provee de una API para poder desarrollar aplicaciones que interactúan con Twitter. Para acceder a la API de Twitter se hacen peticiones (requests) por HTTP y la API devuelve datos en forma estructurada para facilitar su análisis sintáctico. Esta se encuentra documentada en el portal para desarrolladores. La API consiste en tres partes:

REST API: los métodos de esta API permiten acceder a la esencia de los datos.

Search API: como su nombre lo indica, se usa para la búsqueda.

Streaming API: para acceder a los datos casi en tiempo real.

Para este trabajo solo interesan los métodos de la primera parte (REST API), debido a que solo se quiere recolectar los tweets para un conjunto de usuarios en particular, indicando el nombre de usuario de estos. Como se mencionó anteriormente, los métodos de la API se llaman mediante peticiones HTTP, que pueden ser de tipo GET, POST o DELETE. Mediante GET, se pueden acceder a métodos que no efectúan cambios en los servidores de Twitter (como obtener el timeline de un usuario), para métodos que sí efectúan cambios, se usa POST o DELETE en algunos casos. Los datos que devuelven los métodos pueden venir estructurados con los siguientes formatos:

- XML
- JSON
- RSS
- Atom

El formato JSON fue el escogido para recolectar los datos debido a que existe la biblioteca Beautiful Soup para el análisis sintáctico de documentos en XML.

Anexo 2. Detalle de arbol de Cross-Validation al atributo **clase**