ANÁLISIS DE SENTIMIENTOS SOBRE EL POSCONFLICTO COLOMBIANO UTILIZANDO HERRAMIENTAS DE MINERIA DE TEXTO

Presentado por:
Lina Andrea Torres Samboni
Ingeniera de sistemas

Director:
Ignacio Pérez Vélez

Ingeniero industrial y matemático de la Universidad de los Andes, magister en Investigación de Operaciones de Université de Grenoble (Francia) y doctor de tercer ciclo de la Université de Lyon (Francia).

Escuela Colombiana de Ingeniería Julio Garavito
Maestría en Gestión de Información
Bogotá, Diciembre de
2015

TABLA DE CONTENIDO

ANÁLISIS DE SENTIMIENTOS SOBRE EL POSCONFLICTO COLOMBIANO

1. DESCRIPCIÓN DEL PROBLEMA DE INVESTIGACIÓN

Colombia ha estado en conflicto armado alrededor de 50 años y cada Gobierno ha hecho el esfuerzo por terminar la guerra haciendo uso de diferentes estrategias, sin embargo ninguno ha logrado poner fin al conflicto. El actual gobierno ha optado por el recurso del diálogo y ha creado una mesa de conversación, ya han transcurrido 3 años de diálogos y se han establecido algunos acuerdos que a Diciembre de 2015 no se han hecho públicos. A partir de la existencia de dichos acuerdos, se han generado múltiples interrogantes sobre lo que será el posconflicto y la capacidad que tiene el país de asumirlo, entendiéndolo desde los diferentes frentes: sectores políticos, económicos, sociales, entre otros.

Todos los interrogantes que tiene la población colombiana sobre el posconflicto se reflejan en información subjetiva, es decir que cada persona tiene su propio juicio, valoración e interpretación sobre el tema, de tal forma que su opinión es diferente a la de los demás. Entre esos puntos de vista, la valoración se puede clasificar en sentimientos positivos, negativos o neutros; siendo los sentimientos "la evaluación consciente que los seres humanos hacen de la percepción de su estado corporal durante una respuesta emocional" (R. García, 2012).

Dada la naturaleza y sensibilidad del tema en Colombia, se va a realizar un análisis de sentimientos a personajes provenientes de diferentes sectores de la vida pública, porque día a día manifiestan su opinión a temas de la actualidad y además cada uno tiene un perfil o postura sobre las negociaciones que está haciendo el Gobierno. Los perfiles se examinan en los siguientes sectores de interés: guerrilleros, políticos, medios de comunicación y generadores de opinión.

El desarrollo del proyecto se va a hacer con una metodología denominada minería de opinión o como también es conocida, análisis de sentimientos (Cortizo, 2011). Esta tecnología permite extraer contenidos implícitos de recursos como blogs, Twitter, páginas web, etc., los cuales son considerados información que no tiene un modelo formal de dato (Fernández, Miranda, Guerrero, & Piccoli, 2014), en otros términos, están representados en información no estructurada. Asimismo, el uso de esta tecnología permite identificar la polaridad.

La polaridad se encarga de asignar un valor a los términos que expresan una opinión, dependiendo del significado lingüístico de la palabra. (Turney, 2002). Es así como si una palabra tiene un significado positivo o negativo, va tener un valor de acuerdo a un rango numérico especificado y el rango lo define quien esté realizando el estudio. Un ejemplo de ello es un rango de -4 a 4, siendo -4 la asignación a una palabra muy negativa y 4 la asignación a una palabra muy positiva.

Otra consideración importante de la minería de opinión es que crea un escenario de mayor alcance en la obtención de información y resulta más potente para el estudio de los resultados que se arrojan donde su tratamiento incluye series de tiempo, optimización, indexación y búsqueda, big data, algoritmos de

procesamiento del lenguaje natural, reglas de agrupación de datos, entre todos (Fernández, Boldrini, Gómez, & Martínez-Barco, 2011).

2. OBJETIVOS

2.1 OBJETIVO GENERAL

Perfilar la opinión positiva, negativa o neutra del posconflicto a partir de información no estructurada proveniente de sectores políticos, medios de comunicación, guerrilleros y generadores de opinión colombianos, como método de análisis de tendencias en opinión.

2.2 OBJETIVOS ESPECÍFICOS

Definir las fuentes de las cuales se va a recolectar toda la información para el análisis.
Crear un conjunto de datos de entrenamiento que determine la polaridad de las palabras.
Crear un modelo especializado para el procesamiento de la información.
Analizar los resultados de las etiquetas de sentimientos en los documentos utilizando los recursos gráficos de la herramienta de minería de texto.

3. MARCO REFERENCIAL

3.1 MINERÍA DE TEXTO

Antes de iniciar con la definición de la minería de texto es importante abarcar conceptos claves como dato, información, conocimiento y minería de datos, para lograr un mejor entendimiento y contextualizar al lector desde una primera instancia.

Dato es un conjunto discreto de valores, por ejemplo: 5, María, Abril. E información es el conjunto de datos procesados y que tienen un significado (relevancia, propósito y contexto) dado por el observador, como puede ser, María cumple años el 5 de Abril. Luego, el conocimiento será la capacidad de transformar la información y la experiencia de las personas en la toma de decisiones sobre una acción.

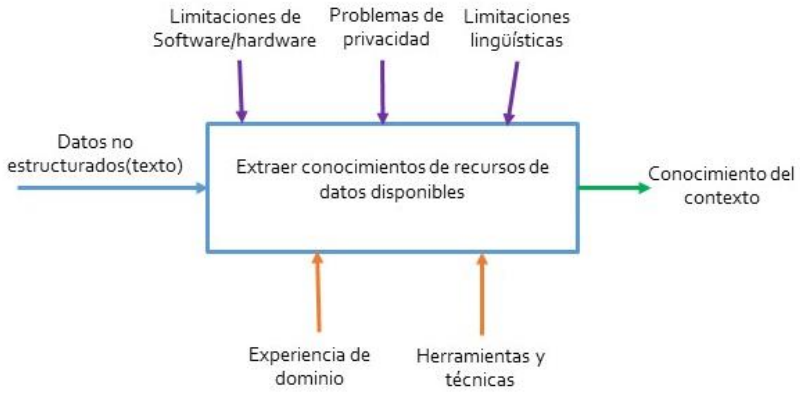
En el mundo de las tecnologías de la información, los datos son almacenados normalmente en bases de datos y su volumen puede llegar a ser muy grande, a partir de esto surge la minería de datos para ayudar a la comprensión de los contenidos en dicho almacenamiento. Las bases de datos tienen una estructura y un esquema de organización conocido, luego a partir de ellas se trata de adquirir conocimiento de datos originales, lo que hace más fácil la extracción de información.

La minería de texto (text data mining TDM) es una aplicación de la minería de datos (Hearst, 2003) y consiste en descubrir o hallar, a partir de cantidades de información no estructurada el conocimiento del cual no existe ningún registro escrito. La información no estructurada es aquella que no está contenida en un "almacén" (base de datos), de forma organizada para luego ser encontrada y utilizada fácilmente para distintos propósitos, lo que dificulta su extracción. Esta información puede estar representada en textos como los mensajes de correo electrónico, presentaciones en power point, documentos en word, mensajes instantáneos (Twitter, WhatsApp), software de colaboración (conferencias de video, salas de chat), entre otros; o se encuentra de forma no textual en imágenes de formato JPEG, archivos de audio MP3, correo de voz, etc.

La principal característica de la minería de texto es que trabaja con base en el lenguaje natural. Este lenguaje es el que hablamos los humanos todos los días, es espontáneo, no es artificial y no ha sido programado de ninguna manera. Es así, como en los textos ya descritos se representa el lenguaje natural, los cuales son objeto de extracción de conocimiento.

Se enfoca en el descubrimiento de patrones interesantes o sucesos recurrentes, su objetivo es descubrir tendencias, desviaciones y asociaciones en la gran cantidad de información textual disponible. Algunas aplicaciones de los sistemas de minería de textos son la identificación y re direccionamiento del contenido de e–mails; los sistemas de vigilancia tecnológica, análisis de información en artículos y libros, búsqueda relevante de contenido en artículos, etc.

Los elementos que intervienen en el proceso de la minería de textos se pueden observar en el siguiente diagrama de contexto:



Fuente. Estudiantes universidad San Carlos

Dado que los datos no estructurados son el insumo principal para realizar minería de textos, un ejemplo de ello es el siguiente Tweet:

"Militares se alistan en las aulas para posconflicto. Capacitaciones en #SENA Formulación de proyectos. Vía @ELTIEMPO http://app.eltiempo.com/colombia/cali/posconflicto-en-el-valle/16400807…".

Este texto no tiene un estándar porque contiene varios símbolos, direcciones web sin ningún significado y si se tomara otro Tweet también se representaría de forma diferente. No tiene una directriz donde se especifique una estructura como por ejemplo, que siempre debe ir al inicio del tweet el nombre de la noticia, luego las etiquetas de otros usuarios y la dirección URL del sitio web donde se puede consultar la información, de esta manera un tweet siempre tendrá una organización y disposición totalmente diferente..

En la salida, cuando se finaliza el estudio se obtiene el conocimiento extraído a partir del análisis de tendencias, clasificaciones o agrupaciones de datos, es aquí donde se alcanzan los objetivos que se definen al iniciar la minería de textos.

En este proceso también intervienen una serie de restricciones, encabezadas por las limitaciones impuestas por el software como los límites de descarga de registros que tienen los navegadores de

internet como Google, Internet Explorer y las redes sociales como Facebook y Twitter. Otras limitaciones son de hardware y cuestiones de privacidad para obtener los permisos de usuario para acceder a grandes volúmenes de información.

La minería de textos cuenta con una metodología propia para llevar a cabo el proceso de extracción de información. Esta metodología puede ser entendida como el proceso mediante el cual se llevan a cabo una serie de tareas ordenadas orientadas para la consecución de tres objetivos principales:

1. Establecimiento del corpus (conjunto de textos).
2. Creación de la matriz de términos.
3. Extracción del conocimiento.

Metodología de extracción de la información:

1. Establecer el corpus

En esta primera etapa surge el concepto de la lingüística de corpus, que consiste en una rama de la lingüística que realiza investigaciones a partir de datos obtenidos de un corpus (plural córpora), es decir, muestras reales del uso de la lengua. El término corpus se usa en dos sentidos y el que aplica en el contexto de la minería de texto es aquel en el que es la recopilación de material lingüístico hecho con un propósito de investigación concreto, como muestras de oraciones, enunciados o textos. Los córpora de gran tamaño en formato digital constituyen una fuente muy robusta de información sobre el uso de la lengua, ya sea información gramatical, semántica, léxica, discursiva o de otro tipo. Se utilizan, por ejemplo, como fuente de información para la redacción de diccionarios y para el procesamiento automático del lenguaje natural (Cervantes, 2015).

El propósito principal del establecimiento del corpus es recolectar todos los documentos relacionados con el contexto a estudiar. La colección resultante puede estar compuesta por documentos de texto, XML (eXtensible Markup Language), correos electrónicos, páginas web, comentarios de redes sociales, RSS (Really Simple Syndication, es un formato XML para sindicar o compartir contenido en la web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos.), entre otros.

Una vez terminado el proceso de recolección, los documentos se organizan y transforman de tal manera que al final todos estén en el mismo formato (por ejemplo en ficheros de texto ASCII), según como lo pueda manejar la herramienta de minería de textos, para que luego puedan ser procesados por una máquina. La organización puede ser sencilla y consistir en una colección almacenada en un directorio o más elaborada y consistir en múltiples registros de una base de datos.

2. Crear la matriz de términos

En este paso, se utiliza el corpus para crear la matriz de términos del documento, también conocida como term-document matrix (TDM). En dicha matriz, las filas representan los documentos, mientras que las columnas representan los términos, es decir las palabras del documento. La relación entre términos y documentos se expresa mediante índices, es decir, medidas relacionales que pueden ser tan simples como el número de ocurrencias de cada término en cada documento o el número de apariciones del término. La meta consiste en representar la esencia del corpus a través de la matriz de términos.

Sin embargo, no todos los términos que se encuentran en un documento lo caracterizan como es el caso de los artículos, verbos auxiliares, signos de puntuación, pronombres, etc.; por ejemplo en las siguientes frases:

He de hacerlo cuanto antes.
Hoy **he** comido a las dos.

El verbo "he" es auxiliar y no tiene poder de diferenciación, por lo tanto no debe formar parte de la matriz de términos. A esta lista de términos que no tiene poder de diferenciación se le conoce con el nombre de "stop words". La construcción inicial de una matriz de términos debe incluir todos aquellos términos identificados en el corpus (columnas) a excepción de aquellos presentes en la lista de "stop words", todos los documentos del corpus (filas) y la ocurrencia de cada término en cada documento (intersección fila-columna o celda).

Además de la técnica de "stop words", también se lleva a cabo aquel conocido como stemming o lematización. El stemming consiste en reducir un término a su raíz (stem o lema) mediante un algoritmo para que diferentes formas gramaticales o declinaciones verbales se identifiquen con un mismo término; por ejemplo, el resultado de stemming para las siguientes palabras; perros, perras, perrito, perrote, es perr; en el cual las diferentes formas que podían adoptar esas palabras son reducidas a una forma común.

Lo siguiente es calcular la frecuencia inicial de los términos para poder aplicar filtrar aquellos que no son interesantes. La frecuencia es el número de apariciones de un término dentro de un documento lo cual significa la importancia que tiene en dicho documento, pero no es razonable pensar que la frecuencia tiene igual importancia en un documento u otro. Por ejemplo, si un término aparece una vez en el documento A y tres veces en el documento B no es apropiado concluir que ese término es tres veces más descriptivo del documento B que del documento A. Para construir adecuadamente los índices existen los siguientes procesos de estandarización:

Frecuencias logarítmicas: esta transformación disminuye el impacto de las frecuencias originales (raw frequencies) y cómo afectan al resultado de los análisis y cómputos.

Frecuencias binarias: la matriz de términos resultante sólo contendrá 1s y os para indicar la presencia o ausencia de las respectivas palabras o términos. Esta transformación disminuye el impacto de las frecuencias originales en los análisis y cómputos que se realicen.

Frecuencias inversas (Inverse document frequencies): consiste en una transformación muy útil porque ayuda a medir la relevancia de los términos, dado que refleja el número de apariciones de un término en el documento y tiene varias versiones, por lo general basadas en una función inversa del número de documentos en el que aparezca el término en cuestión.

3. Extraer el conocimiento

Tras la construcción de la matriz de términos, se extraen patrones en el contexto del problema específico que está siendo estudiado. Por ejemplo en el ámbito comercial, donde resulta interesante encontrar patrones ocultos de consumo de los clientes para poder explorar nuevos horizontes. Un caso específico es el saber que un vehículo deportivo corre un riesgo de accidente casi igual al de un vehículo normal, cuando el dueño tiene un segundo vehículo en casa; esta afirmación ayuda a crear nuevas estrategias

comerciales para ese grupo de clientes. Asimismo, predecir el comportamiento de un futuro cliente, basándose en los datos históricos de clientes que presentaron el mismo perfil, ayuda a poder retenerlo durante el mayor tiempo posible (Ollero Fernández, 2015).

Las principales cuatro técnicas para la extracción de información son: clasificación, clustering, asociación y análisis de tendencias.

3.1 Clasificación: la clasificación, como su propio nombre indica, consiste en agrupar las instancias pertenecientes a los datos en un conjunto de categorías o clases conocidas. La clasificación de textos asigna a varios documentos una o más categorías, etiquetas o clases basadas en el contenido, esto se conoce como aprendizaje supervisado porque la función a partir de los datos de entrenamiento. Un ejemplo de esto es que dado un conjunto de temas y una colección de documentos de texto se tenga que encontrar cuál es el tema de cada documento a través de modelos construidos mediante conjuntos de datos de prueba. En la actualidad, la clasificación de textos automática se aplica a filtros de spam, clasificación de páginas web, generación automática de metadatos, detección de género y muchos otros, un caso puntual se produce cuando se toma un documento que va ser publicado en una página web y el clasificador de textos ubica el documento dentro de una categoría, como deportes, política, cultura o automóviles.

3.2 Clustering: también se conoce como agrupamiento y consiste en la división de los datos en grupos de objetos similares. De forma general, las técnicas de clustering son las que utilizando algoritmos matemáticos que se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se clasifican en clases que son muy similares internamente (entre los miembros de la misma clase) y a la vez diferente entre los miembros de las diferentes clases (Leyva Abreu, 2012).

Este tipo de análisis de aplica cuando se desconocen las etiquetas asociadas a los datos. Al hacer clústeres, se pueden identificar espacios de características, y por lo tanto, descubrir distribuciones de patrones y correlaciones entre los atributos. A diferencia de la clasificación, el clustering o aprendizaje no supervisado clases predefinidas (ni conjuntos de entrenamiento), por esta razón, el clustering es un ejemplo de aprendizaje por observación, mientras que clasificación es un aprendizaje por ejemplos.

El clustering está especialmente indicado en aquellas aplicaciones donde se trata de detectar relaciones entre varios textos, de distribuirlos dinámicamente en agrupaciones naturales o de descubrir los temas más relevantes que emergen de sus contenidos y expresarlos en sus propios términos. En particular, el clustering se aplica donde se requiere descubrimiento de temas no predefinidos ni previstos en encuestas y reclamaciones (que permita una gestión más proactiva y una respuesta más eficaz); agregación y descripción de la reproducción exacta de una frase utilizando "sus propias palabras"; análisis de la voz del cliente, empleado, ciudadano, etc., gestión de ideas, la gestión de la experiencia del cliente.

Además, el clustering juega un papel muy importante en aplicaciones como la exploración de datos científicos, agrupación de documentos, aplicaciones sobre bases de datos espaciales (tales como GIS o datos procedentes de astronomía), aplicaciones Web, marketing, diagnóstico médico, análisis de ADN en biología computacional y muchas otras.

3.3 Asociación: consiste en la generación de reglas que identifiquen las relaciones directas entre términos o conjuntos de conceptos. Las reglas de asociación son utilizadas para analizar la literatura que se ha publicado (noticias y artículos académicos publicados en la web), entre otros. El propósito principal es identificar automáticamente las asociaciones entre diferentes conjuntos.

Las reglas de asociación se generan a partir de un conjunto de m textos T = {t1,t2,..., tm} y un conjunto de n términos claves K = {k1, k2,..., kn} asociados a los textos, una regla de asociación se define como X ⇒ Y , donde X e Y son conjuntos de términos claves. De acuerdo a estas consideraciones, las reglas de asociación del tipo X ⇒ Y tienen ciertos criterios "tradicionales" de medición que son utilizados para reducir la cantidad de reglas descubiertas y para establecer ciertos niveles de relevancia:

1. Support: corresponde al número de documentos que contienen ambos términos claves X e Y, y se puede considerar como la probabilidad conjunta de X e Y, o sea P(X IY).

2. Confidence: corresponde a la probabilidad que el término Y aparezca en el texto dado que X ya apareció. También se puede considerar como la probabilidad condicional de Y dado X, o sea P(Y/ X).

Un problema con las métricas anteriores corresponde a que no son muy efectivas para la reducción de reglas que no necesariamente aporten conocimiento relevante o interesante. Una de las razones para ello, es su naturaleza estadística, dado que no toma en consideración ningún otro factor aparte de la frecuencia con que aparece un término en el texto (Pérez Cárcamo, 2007).

3.4 Análisis de tendencias: los métodos más recientes de análisis de tendencias se basan en la noción de que varios tipos de distribuciones de términos son funciones de colecciones de documentos, o lo que viene a ser lo mismo, diferentes colecciones llevan a diferentes distribuciones de términos para el mismo conjunto de términos. Por tanto, es posible comparar dos distribuciones que son idénticas a excepción de su procedencia. Por ejemplo, tener dos colecciones de la misma fuente pero en diferentes momentos en el tiempo: Delene y Crossland (2008) aplicaron el análisis de tendencias a una gran cantidad de artículos académicos para identificar la evolución de los conceptos clave en el campo de los sistemas de información.

3.2 METODOLOGÍA DE MINERÍA DE DATOS

El desarrollo del proyecto se hizo utilizando una metodología de minería de datos (CRIPS-DM) y de minería de textos. A continuación se describe detalladamente la primera:

El modelo de referencia CRISP-DM

Existen varios modelos de proceso que se han tomado como base para el desarrollo de proyectos de Data Mining, como SEMMA (Sample, Explore, Modify, Model, Assess), DMAMC (Definir, Medir, Analizar, Mejorar, Controlar), o CRISP-DM (Cross Industry Standard Process for Data Mining), sin embargo ésta es la guía de referencia más utilizada, según una publicación del 2007 de kdnuggets.com (Rodríguez Rojas, s.f.).

CRISP-DM, está dividida en 4 niveles de abstracción organizados de forma jerárquica (Figura 1) en tareas que van desde el nivel más general, hasta los casos más específicos.

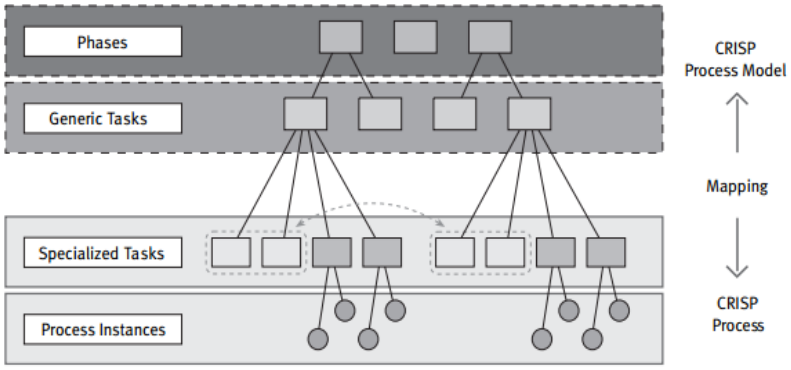


Figura 1. Esquema de los cuatro niveles de CRISP-DM. (Chapman, Clinton, & Kerber, 2000)

Organiza el proyecto de minería de datos en seis fases, como se muestra en la Figura 2. La secuencia de las fases no es rígida.

El movimiento hacia adelante y hacia atrás entre fases diferentes siempre es requerido. El resultado de cada fase determina que la fase, o la tarea particular de una fase, tienen que ser realizados después. Las flechas indican las más importantes y frecuentes dependencias entre fases.

El círculo externo en la Figura 2 simboliza la naturaleza cíclica de la minería de datos. La minería de datos no se termina una vez que la solución es desplegada. Las informaciones ocultas (lecciones ocultas) durante el proceso y la solución desplegada pueden provocar nuevas, a menudo más preguntas enfocadas en el negocio. Los procesos de minería subsecuentes se beneficiarán de las experiencias previas. A continuación cada fase (Chapman, Clinton, & Kerber, 2000):

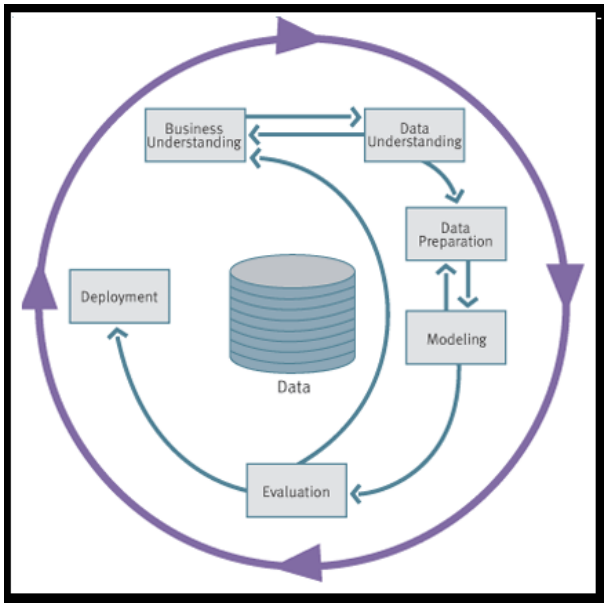


Figura 2. Fases del modelo CRISP-DM. (Chapman, Clinton, & Kerber, 2000)

Comprensión del negocio

Esta fase inicial se enfoca en la comprensión de los objetivos del proyecto y las exigencias desde una perspectiva de negocio o de la definición del problema, luego se define un problema de minería de datos tomando el conocimiento de los datos y se hace una planeación preliminar para lograr dichos objetivos.

Comprensión de los datos

La fase de entendimiento de datos comienza con la recolección de datos inicial y continua con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Preparación de datos

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos en las herramientas de modelado) de los iniciales datos en bruto. Las tareas de preparación de datos probablemente van a tener varias iteraciones y no tienen un orden establecido. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que los modelan.

Modelado

En esta fase, se selecciona la técnica de modelado que se va utilizar aunque es posible que en la fase de comprensión del negocio ya se haya seleccionado. Sin embargo, existen varias técnicas de modelado para el mismo problema, sino que cada una depende de unos requerimientos específicos sobre cómo deben estar preparados los datos.

Evaluación

En esta etapa del proyecto, se ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos.

Antes de proceder al despliegue final del modelo, es importante evaluar y revisar las actividades que se hicieron para crearlo, y determinar si se están consiguiendo los objetivos de negocio evaluados en la primera etapa. En el final de esta fase, se debe decidir el uso que se le dará a los resultados obtenidos de la minería de datos.

Desarrollo

La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento adquirido tendrá que ser organizado y presentado de tal forma que el cliente pueda usarlo. Dependiendo de los requerimientos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo.

Para finalizar, la figura 3 presenta un contexto de cada una de las fases acompañadas por tareas genéricas y sus respectivas salidas. La figura se muestra a modo ilustrativo para tener una idea de las tareas que tiene cada fase y la dimensión total que tiene un proyecto de minería de datos.

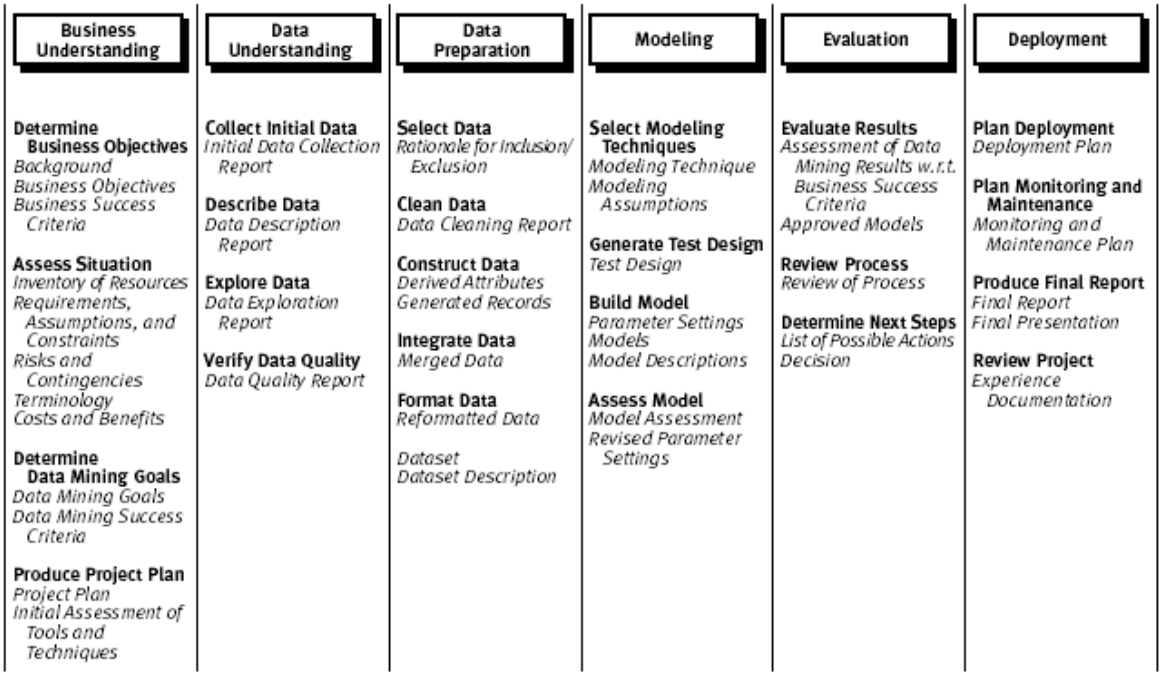


Figura 3. Tareas genéricas (negritas) y salidas (cursivas) del modelo de referencia CRISP-DM. (Chapman, Clinton, & Kerber, 2000)

3.3 ¿QUÉ ES EL ANÁLISIS DE SENTIMIENTOS? (OPINION MINING)

El análisis de sentimientos o minería de opinión (opinion mining) es una extensión de la minería de textos, donde a partir del análisis del procesamiento del lenguaje natural que puede encontrarse representado en opiniones, sentimientos, puntos de vista, emociones, etc. (Cortizo, 2011); genera una valoración positiva o negativa al corpus seleccionado mediante el aprendizaje supervisado.

Una de sus principales características consiste en que el tipo de información analizada además de ser subjetiva, generalmente es representada de forma escrita y sin un estándar o un formato pre establecido. Hablar de minería de opinión es hablar de un campo cada vez más extenso, relacionado con el análisis de los componentes que están implícitos en los contenidos generados por los usuarios de internet, por ejemplo. Dentro de este campo, existen aplicaciones que realizan un análisis más o menos profundo de los contenidos textuales, en función de la tarea o problema que se quiera resolver. En general, se encuentran dos tipos de tareas relacionadas con la minería de opinión (Cortizo, 2011):

- Detección de la polaridad: es la capacidad de determinar si una opinión es positiva o negativa. Más allá de una polaridad básica, también se puede obtener un valor numérico dentro de un rango determinado, que de una determinada forma trate de obtener una valoración objetiva asociada a determinada opinión. Su principal característica es el uso de diccionarios semánticos y de la estructura sintáctica de las oraciones para clasificar un texto, sin embargo esta técnica es muy dependiente de la calidad, el tamaño y dominio de los datos de entrenamiento.

- Análisis del sentimiento basado en características: es la capacidad de determinar las distintas características de un producto mencionadas en la opinión o en una reseña escrita por un usuario desde un sitio web, por ejemplo los comentarios que los usuarios de Facebook escriben sobre productos para la

salud, y para cada una de esas características mencionadas en la opinión, tener la capacidad de extraer una polaridad. En otras palabras, se trata de determinar qué orientaciones definen las opiniones de los usuarios en su conjunto, y cuáles tratan de identificar un conjunto de características en las opiniones de los usuarios y proporcionan un informe detallado sobre la polaridad de cada característica (Peñalver Martínez, 2015).

La importancia de este estudio se debe al interés que genera en varios sectores porque se vuelve información muy valiosa, en parte se debe a la inmensa cantidad de información disponible como es la evaluación de: productos de consumo, programas informáticos, deportes, películas, etc., (Rangel, Sidorov, & Suárez-Guerra, 2014).

La minería de opinión inicialmente tuvo fines informáticos, pero a medida que esta investigación fue evolucionando tomó tal fuerza que del área de informática se ha extendido a las ciencias políticas y ciencias sociales, debido a la importancia para los negocios y la sociedad en su conjunto y, se han creado aplicaciones muy interesantes como (Esquivel Gámez, 2009):

- Análisis de mercados, en el cual se recogen estadísticas sobre la ocurrencia de palabras, frase o temas que serán útiles para estimar mercados, curvas de demanda y demográficas, revisión de detalles de productos desde catálogos en línea.
- Manejo de relaciones con el cliente (CRM), mediante la minería de correos de entrada de las quejas y retroalimentación de clientes, aunque también se puede realizar en llamadas telefónicas, foros, listas de correo, cartas, encuestas de opinión.
- Administración de recursos humanos, en la cual con la minería de reportes de la compañía y sus correspondientes actividades, condiciones y problemas, se pueden detectar a tiempo las variables que pueden indicar el estado del clima organizacional.
- Análisis de bases de datos de patentes para detectar los mayores participantes, tendencias y oportunidades.

Esta metodología aplica en varios ámbitos empresariales y sociales porque las opiniones fundamentan las actividades humanas y son factores de influencia en los comportamientos de las personas; de esta forma las decisiones que la gente toma están condicionadas en experiencias de otros individuos y cuando se va a hacer una elección, se revisan las opiniones de otros para confirmar su experiencia.

La creciente importancia del análisis de los sentimientos coincide con el crecimiento de los usuarios de internet y más específicamente de las redes sociales. Los medios sociales aportan cada día un gran conjunto de datos que por primera vez en la historia humana permiten tener acceso a un enorme volumen de datos almacenados en un formato digital para un análisis (Pang & Lee, 2008). Como consecuencia de ello, se ha incrementado la atención por identificar el contenido emocional, es decir, las emociones que se generan de esta cantidad de publicaciones, por ejemplo, como producto de la gran cantidad de comentarios que se generan en Twitter, en los foros de discusión, blogs, micro-blogs, y las redes sociales es notorio que los usuarios tienden a dividirse entre aquellos que están a favor o en contra de algún acontecimiento (Gálvez-Pérez, y otros, 2015). Debido a esto los especialistas se centran en el estudio de estos datos con herramientas de data mining para extraer los patrones y tendencias con respecto a temas determinados.

Aunque existen varias metodologías para estos análisis como la detección de la polaridad o el análisis basado en características, entre otros, uno de los métodos más comunes para evaluar un sentimiento es clasificar las palabras de acuerdo a una escala (polaridad), como se definió anteriormente.

3.4 ANTECEDENTES DE ANÁLISIS DE SENTIMIENTOS

La minería de opinión ha despertado el interés de los investigadores para su análisis y es por ello que actualmente se encuentran varios estudios enfocados en diversas especialidades. A continuación se mencionan algunos de ellos, en los que se destaca el análisis de sentimientos con información que proviene de la Web 2.0 y que se distinguen por hacer un procesamiento del lenguaje natural en español:

""Linguistic Inquiry and Word Count" (LIWC) es una herramienta de software que analiza textos y fue diseñado por James W. Pennebaker, Roger J. Booth, y Martha E. Francis. El LIWC es capaz de calcular cómo las personas usan diferentes categorías de palabras a través de una gran diversidad de textos. Ya sea en correo electrónicos, discursos, poemas o la trascripción de cualquier diálogo cotidiano, el LIWC permite determinar el grado en que autores/hablantes usan palabras que connotan emociones positivas o negativas, auto-referencias, palabras extensas o palabras que se refieren a sexo, comer o religión. El programa fue diseñado para analizar simple y rápidamente más de 70 dimensiones del lenguaje a través de cientos de muestras de texto en segundos.", (Associates, 2013)

La investigación "Estudio de las categorías LIWC para el análisis de sentimientos en español" de los autores (Salas-Zárate, Rodríguez-García, Almela, & Valencia-García, 2011), consiste en el análisis de varias dimensiones lingüístico- psicológicas obtenidas desde LIWC para clasificar opiniones en español en cinco categorías. LIWC tiene un diccionario en español con 7.515 palabras y cada palabra se puede clasificar en 72 categorías, además las categorías se clasifican en cuatro dimensiones: procesos lingüísticos estándar, procesos psicológicos, relatividad y asuntos personales. Este trabajo estuvo a cargo de investigadores españoles quienes definieron un corpus que fue obtenido de opiniones de productos electrónicos tales como dispositivos móviles. El corpus se procesó en LIWC y la evaluación de resultados se hizo mediante clasificadores con los siguientes algoritmos J48, SMO, y BayesNet de WEKA. Estos algoritmos se basan en árboles de decisiones, que consisten en representación de funciones lógicas (if-then).

Por medio de aprendizaje automático se infiere un árbol de decisión a partir de un conjunto de instancias o ejemplos. El algoritmo "J48" de Weka utiliza un método heurístico para inferir el árbol, donde se realiza la selección del atributo en cada nivel del árbol en función de la calidad de la división que produce. (Kohavi, R. & Quinlan, J. R., 2002). Por otro lado, el algoritmo SMO es un clasificador hace uso del algoritmo secuencial de optimización minimal de John Platt para entrenar el clasificador de soporte vectorial.

Finalmente los resultados expresaron que la clasificación de opiniones con dos categorías (positiva, negativa) obtuvo mejores resultados, siendo el clasificador SMO el que tuvo un mejor comportamiento.

Por otro lado, la investigación "Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog Sentiment Analysis and Opinion Mining: The EmotiBlog Corpus" (Fernández, Boldrini, Gómez, & Martínez-Barco, 2011), se basó en la carencia de recursos, métodos y herramientas para un análisis efectivo de la información subjetiva. El software utilizado fue *EmotiBlog*, el cual es una colección de entradas de blogs creado y anotado para detectar expresiones subjetivas en los nuevos géneros textuales

nacidos con la Web 2.0. El principal reto fue demostrar que la creación de un corpus y con el uso de la implementación tecnológica *EmotiBlog*, se podían superar esos desafíos. Además, demostrar que con la ayuda de: SentiWordNet y WordNet (Baccianella, Esuli, & Sebastiani), se aumentaría la cobertura de los resultados sin disminuir la precisión, incluyendo métodos de procesamiento del lenguaje natural (PNL) *como stemming o lematización, bolsa de palabras, stop words*, etc.

En la medida que se recolectó información sobre estudios previos en minería de opinión, se encontró que en gran parte de los experimentos se genera el corpus por medio de comentarios de Twitter, haciendo un filtro de cada palabra en el corpus de entrenamiento y prueba, dejando únicamente aquellas palabras que cumplan con una y solamente una etiqueta morfológica y el proceso de etiquetado lo hacen usando el etiquetador *TreeTagger* (Jasso-Hernández, Pinto, Vilariño, & Lucero, 2014).

El último estudio a considerar es titulado "Herramienta para el Análisis de Sentimiento en el Proceso de Paz Colombiano" (Guio Fonseca, 2014), que consistió en la construcción de una herramienta para hacer el seguimiento de los comentarios generados en Twitter, con respecto al tema específico del proceso de paz colombiano. Los autores utilizaron las metodologías CRISP-DM y SCRUM, debido a que además de hacer minería de texto también se involucraba el desarrollo de un software. Cabe anotar que SCRUM es una metodología ágil y flexible para gestionar el desarrollo de software, cuyo principal objetivo es maximizar el retorno de la inversión. Se basa en construir primero la funcionalidad de mayor valor para el cliente y en los principios de inspección continua, adaptación, auto-gestión e innovación.

Los autores construyeron un lexicón o un diccionario de términos de forma semiautomática a partir de palabras semillas, estas palabras son aquellas que para el autor están cargadas de significado, en este caso son las palabras definidas por un experto de dominio y, utilizaron calificaciones obtenidas de un diccionario suministrado por la Sociedad Española para el Procesamiento del Lenguaje Natural. Para medir los resultados utilizaron las métricas de precisión y exhaustividad, que son empleadas en la medida del rendimiento de los sistemas de búsqueda, recuperación de información y reconocimiento de patrones. En este contexto se denomina precisión como a la fracción de instancias recuperadas que son relevantes, mientras recall o exhaustividad es la fracción de instancias relevantes que han sido recuperadas. Haciendo el cálculo a partir de la siguiente fórmula:

$$Exhaustividad = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ total\ de\ documentos\ relevantes\ presentes\ en\ el\ fondo\ documental} \times 100$$

$$Precisión = \frac{Número\ de\ documentos\ relevantes\ recuperados}{Número\ total\ de\ documentos\ recuperados} \times 100$$

De esa forma, el análisis concluyó en un modelo que alcanzó una exhaustividad promedio aproximada de 57% y una precisión promedio de 61%, cuyos valores son aceptables según el autor para el modelo hibrido construido (Guio Fonseca, 2014).

3.5 HERRAMIENTA DE MINERÍA DE TEXTO

3.5.1 KNIME

KNIME es una herramienta que integra diversos componentes para aprendizaje automático y minería de datos a través de su concepto de fraccionamiento de datos (data pipelining) modular. La interfaz gráfica de usuario permite el montaje fácil y rápido de nodos para procesamiento de datos (ETL: extracción, transformación, carga), para el análisis de datos, modelado y visualización. Desde sus inicios es utilizado en la investigación farmacéutica, pero también se utiliza en otras áreas, como: análisis de datos de cliente de CRM, inteligencia de negocio y análisis de datos financieros.

Como otros entornos de este tipo (WEKA, RapidMiner, etc.), su uso se basa en el diseño de un flujo que plasma las distintas etapas de un proyecto de minería de datos. Es una plataforma modular de exploración de datos, que permite al usuario crear flujos de datos, de forma visual e intuitiva por medio de nodos; lo cuales son módulos que encapsulan distintos tipos de algoritmos y que implementan diferentes tipos de acciones. Además, permite ejecutar de forma selectiva algunos de los pasos creados o todo el flujo desarrollado. Luego, los resultados se pueden investigar mediante vistas interactivas tanto de los datos como de los modelos.

KNIME es una herramienta de código abierto que puede ser descargada y utilizada gratuitamente bajo los términos de la licencia GPLv3 con una excepción que permite que otros usuarios utilicen el bien definido nodo de API para añadir extensiones de propiedad.

Su primera versión fue lanzada en Julio de 2006 y su propósito es la ejecución de técnicas de minería de datos. Fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania, bajo la supervisión del profesor Michael Berthold.

En la actualidad, la empresa KNIME.com, radicada en Zürich, Suiza, continúa su desarrollo, además de prestar servicios de formación y consultoría. KNIME está desarrollado en Java sobre Eclipse, la cual es una potente y completa plataforma de programación, desarrollo y compilación de elementos tan variados como sitios web, programas en C++ o aplicaciones Java.

Knime tiene nodos que implementan distintos tipos de acciones que pueden ejecutarse sobre diferentes tipos de fuentes de datos:

- Manipulación de filas, columnas, como muestreos, transformaciones, agrupaciones.
- Visualización (histogramas).
- Creación de modelos estadísticos y de minería de datos, como árboles de decisión, regresiones.
- Validación de modelos, como curvas ROC.
- Scoring o aplicación de dichos modelos sobre conjuntos nuevos de datos.
- Creación de informes a medida gracias a su integración con el proyecto BIRT (Business Intelligence and Reporting Tools) de Eclipse.
- El carácter abierto de la herramienta hace posible su extensión mediante la creación de nuevos nodos que implementen algoritmos a la medida del usuario.

Debido a la experiencia y dominio de la herramienta por parte de los autores, el desarrollo del proyecto se hace sobre Knime.

4. DISEÑO METODOLÓGICO

4.1 COMPRENSIÓN DEL NEGOCIO

En esta etapa inicial se comprende el objetivo del proyecto y se definen las fuentes de información. En esta metodología se seleccionaron personajes que generan interés de acuerdo a las referencias dadas por los siguientes artículos "21 Personajes de la Política Colombiana con más de 50.000 Followers" (Suarez, 2013), el artículo de la revista Semana "20 mejores líderes de Colombia" (Semana, Semana, 2013) y el artículo de la revista Dinero "Los más influyentes" (Dinero, 2014) y otros artículos especificados a lo largo del documento; sin embargo, se trató de usar un criterio imparcial teniendo la certeza de que este listado se puede complementar en trabajos futuros. De esta forma se agruparon los personajes dentro de 5 categorías que fueron consideradas relevantes de acuerdo al contexto nacional en el que se producen las noticias e información importante del país:

Medios de comunicación en formato RSS: en esta categoría se encuentran los medios que publican noticias diarias en el formato RSS, el cual es un formato de datos que sirve para el envío de contenidos a quienes están registrados en un determinado sitio de Internet. Esta estructura permite que la distribución del contenido se realice sin que sea necesario valerse de un navegador, ya que la acción se lleva a cabo a través de un software creado especialmente para leer esta clase de datos que se conoce como agregador.

En ese sentido, se escogieron las noticias generadas por los periódicos El Espectador y El Tiempo, ya que ofrecen este formato y además son diarios muy reconocidos a nivel nacional.

Comentarios en Twitter (Tweets) de guerrilleros colombianos: en la definición de este listado se escogieron guerrilleros de las FARC y el Ejército de Liberación Nacional (ELN) que tuvieran una cuenta en Twitter, puesto que ninguna publicación formal ha hecho un artículo o un comunicado donde filtren los personajes de este tipo con gran importancia e influencia en el país. De esa manera se seleccionaron los siguientes:

Nombre	Twitter
Timoleón Jiménez	Timochenko_FARC
Gabriel Ángel	GabAngel_FARC
Diálogos Paz FARC	FARC_Epaz
Yadira Suárez	Yadira_FARC
Iván Marquéz	IvanMarquezFARC
Mujer Fariana	MujerFariana
Ricardo Téllez	Ricardo_TFARC
Julián Subverso	Subverso_FARC
Viviana Hernández	Viviana_FARC
Pastor Alape	AlapePastorFARC
Milena Reyes	FARC_MilenaR
Olga Arenas	Olga_FARC
Sergio Marín	Sergio_FARC
Boris Guevara	BorisG_FARC

Alexandra Nariño	Tanja_FARC
Wendy Arango	wendya_FARC
Carmenza Castillo	Carmen_FARC
Raúl Urrego	Raul_UrregoFARC
Ivonne Rivera León	Ivonne_FARC
Frente Antonio Nariño	FRENTEAN_FARC
FARC-EP Resistencia	resistenciacol
Patricia Mendoza	Patricia_FARC
ELN RANPAL	ELN_RANPAL
Pablo Catatumbo	Pcatatumbo_FARC
Victoria Sandino	SandinoVictoria
Ejército de Liberación Nacional	ELN_Colombia

Tabla 1. Cuenta de Twitter de guerrilleros

Comentarios en Twitter (Tweets) de políticos colombianos: para hacer esta selección se consultó un experto en Ciencias Sociales quien dio su concepto sobre los personajes de la vida política que despiertan interés para el análisis sobre el tema de estudio, adicional se utilizó la información del artículo de Caracol radio "Revelan ranking de los 10 políticos más influyentes en Colombia por Twitter" (Radio, 2012). Es así como se seleccionaron los siguientes sujetos con su respectiva cuenta oficial de Twitter:

Nombre	Twitter
Clara López	ClaraLopezObre
Iván Cepeda Castro	IvanCepedaCast
Gloria Inés Ramírez	gloriainesramir
Alirio Uribe Muñoz	AlirioUribeMuoz
Gina Parody	ginaparody
Claudia López	CLOPEZanalista
Alfonso Prada	alfonsoprada
Juan Manuel Santos	JuanManSantos
Jose David Name	JoseDavidName
Mauricio Lizcano	MauricioLizcano
Roy Barreras	RoyBarreras
Angelino Garzón	Angelino_Garzon
Aurelio Iragorri V.	MinIragorri
Cecilia Álvarez-C	CeciAlvarezC
Gabriel Vallejo	GabrielVallejoL
Mauricio Cárdenas S	MauricioCard
Gustavo Petro	petrogustavo
Horacio Serpa	HoracioSerpa
Cristo Bustos	CristoBustos
Humberto de la Calle	HdelaCalle_CLN
Guillermo Rivera Flórez	riveraguillermo

Sofía Gaviria	SOFIAGAVIRIAC
Alejandro Gaviria	Agaviriau
Ernesto Samper	ernestosamperp
David Barguil	davidbarguil
Hernán Andrade	AndradeSenador
Marta Lucía Ramírez	mluciaramirez
Andrés Pastrana	AndresPastrana_
Juan Camilo Restrepo	RestrepoJCamilo
Carlos Fernando Galán	CFGalanprensa
German Vargas Lleras	German_Vargas
Álvaro Uribe Vélez	AlvaroUribeVel
Alfredo Rangel	AlRangelS
José Obdulio Gaviria	JOSEOBDULIO
Óscar Iván Zuluaga	OIZuluaga
Paloma Valencia	PalomaValenciaL
Francisco Santos	PachoesColombia
Ernesto Macías Tovar	emaciastovar
Carlos Holmes Trujillo	CarlosHolmesTru
Armando Benedetti	AABenedetti

Tabla 2. Cuenta de Twitter de políticos

Comentarios en Twitter (Tweets) de medios de comunicación: en esta categoría se eligieron los principales medios de comunicación colombianos con un cubrimiento de todo el territorio nacional, de acuerdo a lo mencionado por el periódico el Tiempo en su artículo "EL TIEMPO, el perfil más influyente de Twitter en Colombia" (Tecnósfera, 2015), la lista de los medios de comunicación a nivel nacional del Centro de Prensa Internacional (Internacional, 2013) y el artículo de Kien y Ke "Top 10 de Medios Colombianos en la Web" (KienYKe, 2012). Asimismo, es importante aclarar que en esta categoría se les da un tratamiento a los medios de comunicación como institución y no como persona natural, esto con el propósito de evitar relacionarlos con los periodistas que trabajan ahí. Los siguientes fueron los seleccionados con su cuenta oficial de Twitter:

Medio	Twitter
El Espectador	elespectador
El Nuevo Siglo	ELNUEVOSIGLO
Revista Dinero	RevistaDinero
La Silla Vacía	lasillavacia
La Silla Vacía	lasillaenvivo
Revista Semana	RevistaSemana
Portafolio	Portafolioco
Confidencial	confidencialcol
Kien y Ke	kienyke
La República	larepublica_co

El Tiempo	ELTIEMPO
Caracol Radio	CaracolRadio
Radio Santafe	radiosantafentc
Blu Radio	BluRadioCo
La Fm	LAFmNoticias
W Radio	WRadioColombia
Rcn La Radio	rcnlaradio
Canal RCN	CanalRCN
Caracol TV	CaracolTV
Noticias Uno	NoticiasUno
El País	elpaiscali
La Vanguardia	vanguardiacom
City TV	Citytv
Señal Colombia	SenalColombia
La crónica de hoy	lacronicadehoy
El Heraldo	elheraldoco
El Nuevo Diario	elnuevodiariord
Canal Capital	CanalCapital
Revista Cromos	Cromoscomco
Teleantioquia	Teleantioquia
Matador	matadoreltiempo
Cable Noticias	Cablenoticias
NTN 24	NTN24
La Luciérnaga	LaLuciernaga

Tabla 3. Cuenta de Twitter de medios de comunicación

Comentarios en Twitter (Tweets) de los generadores de opinión: los generadores de opinión son aquellos periodistas o personalidades públicas que son líderes en opinión y están al tanto de las situaciones del país para expresar un comentario. Los siguientes personajes fueron los seleccionados con su cuenta oficial de Twitter, de acuerdo a la información del artículo "Los 8 periodistas colombianos más influyentes en Twitter" (Barrero, 2015) y al artículo "¿Quiénes son los tuiteros más influyentes de Colombia?" de la revista Semana (Semana, Semana, 2015):

Nombre	Twitter
Felix de Bedout	fdbedout
Camila Zuluaga	ZuluagaCamila
Guillermo Prieto L.	PirryTv
Fidel Cano Correa	fidelcanoco
YolandaRCN	YolandaRuizRCN
Héctor Abad	hectorabadf
Felipe Zuleta Lleras	Fzuletalleras
Claudia Morales	ClaMoralesM

León Valencia	LeonVaLenciaA
Claudia Palacios	claudiapcnn
Ramiro Bejarano G	RamiroBejaranoG
Claudia Gurisatti	CGurisattiNTN24
Julio Sanchez Cristo	jsanchezcristo

Tabla 4. Cuenta de Twitter de generadores de opinión

4.2 EXTRACCIÓN DE INFORMACIÓN

El flujo del modelo se inició con la extracción de información de cada una de las categorías definidas en la compresión del problema. Knime ofrece múltiples utilidades para realizar esta tarea y en este caso se va extraer la información de los medios de comunicación en formato RSS de la siguiente forma:

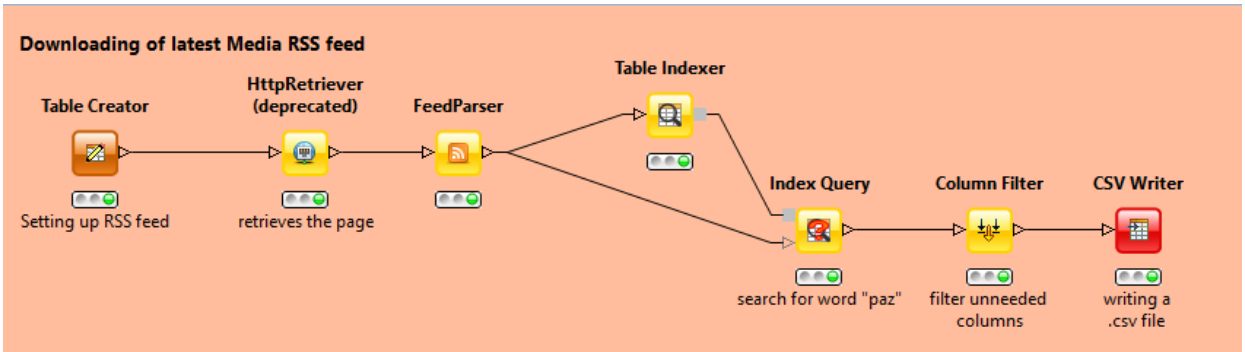


Figura 4. Extracción de información de medios de comunicación en formato RSS

En el primer nodo "Table Creator" se escriben las URL de las páginas web en formato RSS, para que Knime pueda identificar el origen de la información. Los siguientes nodos se encargan de extraer la información desde las páginas web anteriormente referenciadas y filtrar de los resultados obtenidos aquellos que tengan la palabra "paz". Luego se ajusta el documento quitando columnas que no se necesitan y se procede a guardar la información en un archivo de formato .csv, con el nombre de RSSMedia.csv. La información se capturó desde Junio hasta Diciembre de 2015 para obtener una trazabilidad en el tiempo.

Para las siguientes categorías se recolectó la información con el siguiente flujo:

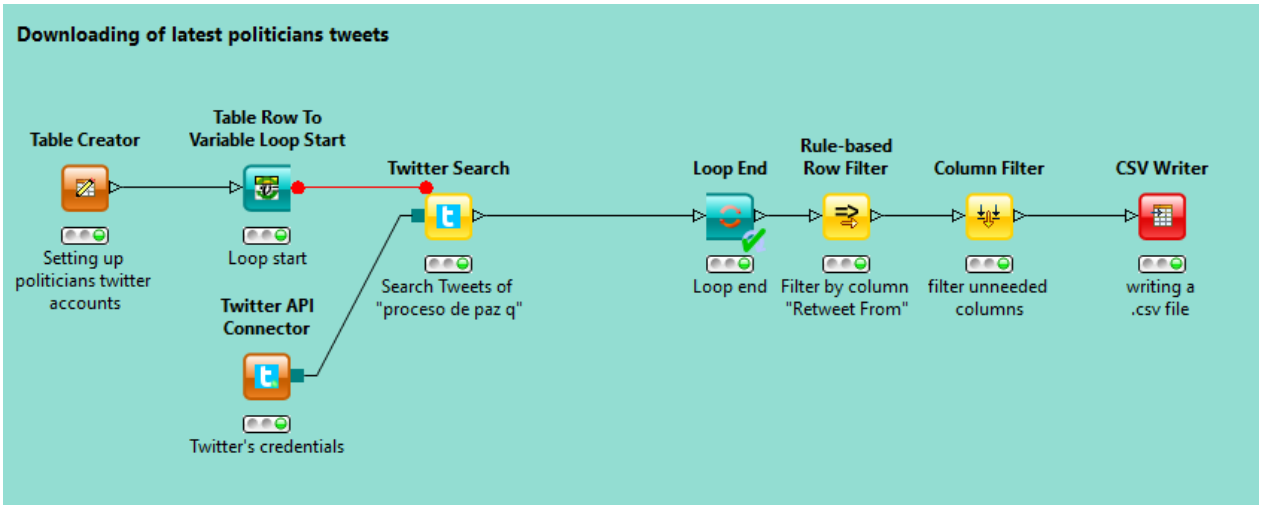


Figura 5. Extracción de información de políticos en Twitter

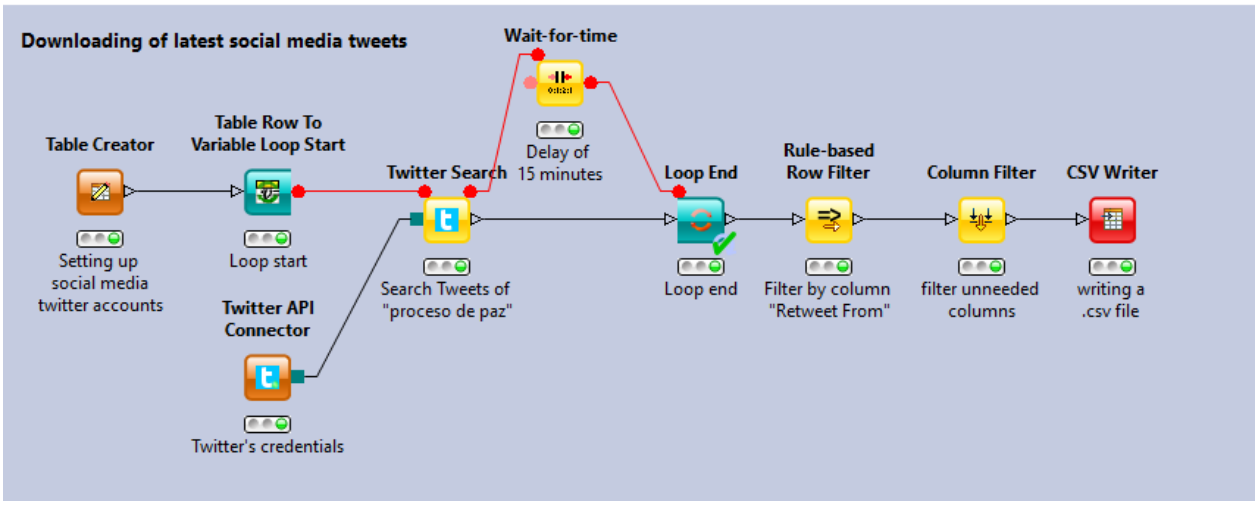


Figura 6. Extracción de información de medios de comunicación en Twitter

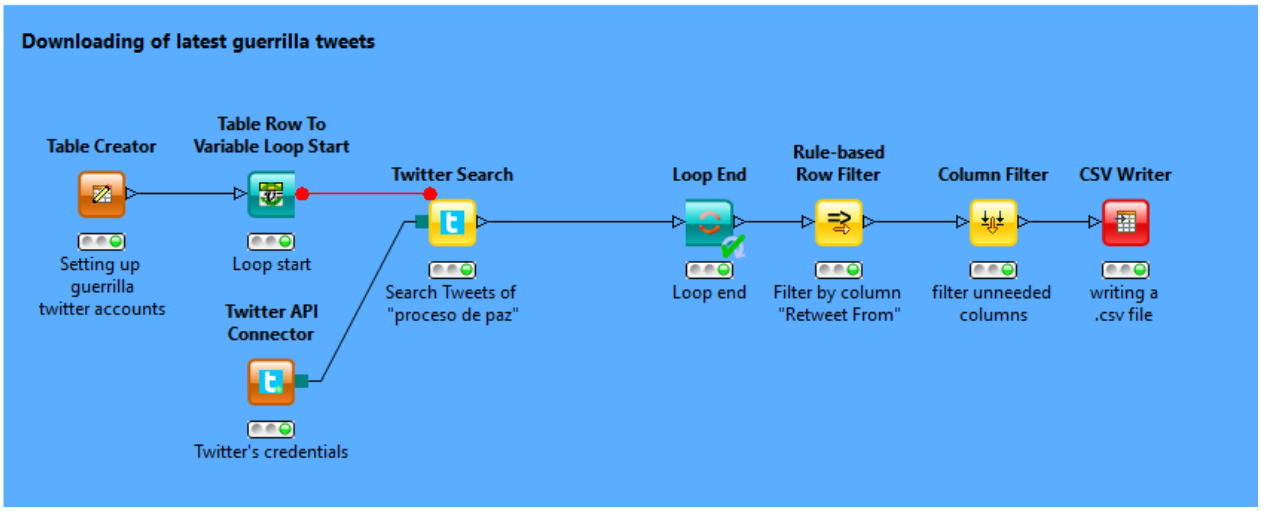


Figura 7. Extracción de información de guerrilleros en Twitter

De la misma forma que se hizo en el flujo anterior, en la tabla del nodo Table Creator y en cada una de sus filas se construyen las consultas que se van a hacer en Twitter.

Cada consulta tiene la siguiente estructura: "proceso de paz" + "Cuenta oficial de Twitter", de esta forma el texto: proceso de paz ClaraLopezObre, quiere decir que se va a hacer una consulta sobre la frase "proceso de paz" y que además coincida con la cuenta oficial de Clara López.

El flujo consiste de un ciclo (loop) en Twitter, de tal forma que se realice una búsqueda de todos los registros definidos en la tabla del primer nodo y a su vez se almacenen los resultados.

Es importante mencionar que el nodo Twitter API Connector se debe configurar con la cuenta de Twitter que va ejecutar las consultas y para este proyecto se creó una cuenta en Twitter que "seguía" (follow) a todos los políticos, medios de comunicación y generadores de opinión definidos. Además se debe crear una aplicación desde la plataforma para desarrolladores de software que provee Twitter, a continuación se muestra todo ese proceso.

Después de crear la cuenta se debe iniciar sesión en la dirección de Twitter https://dev.twitter.com con el mismo usuario y contraseña. En seguida se selecciona el botón "fabric" que es el gestor de aplicaciones como se muestra en la siguiente figura:

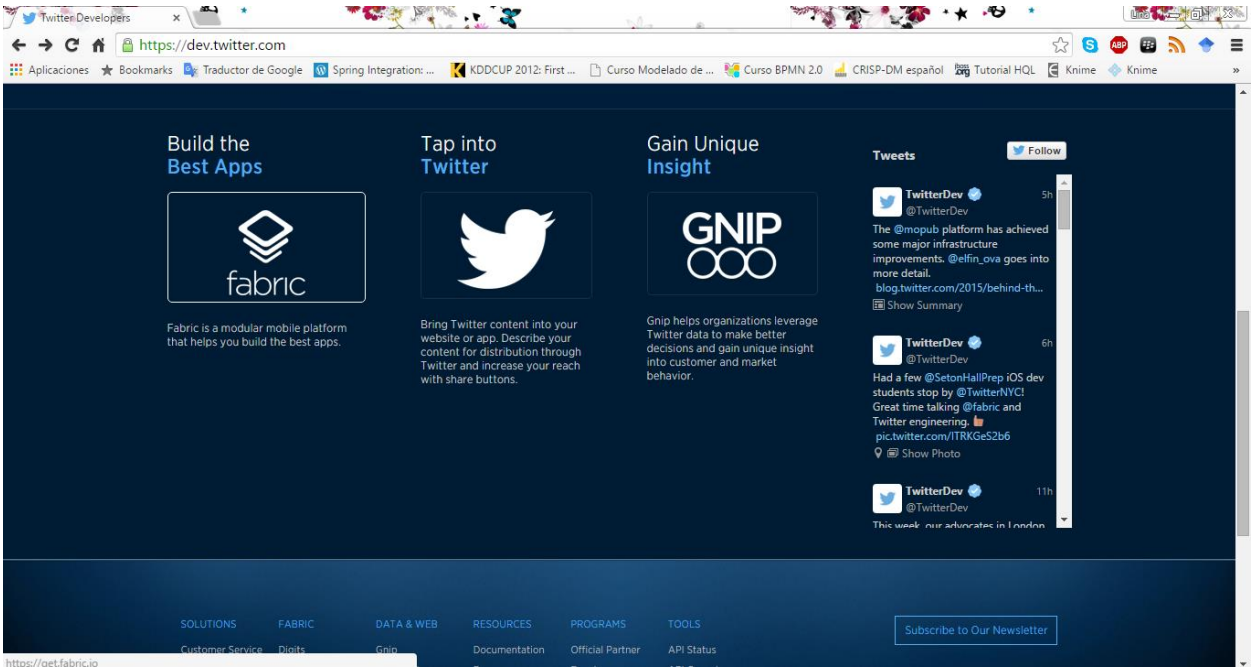


Figura 8. Ingreso al sitio web de desarrollo de Twitter

El sistema hace un direccionamiento a una nueva URL y se selecciona el botón "Get Started with Fabric".

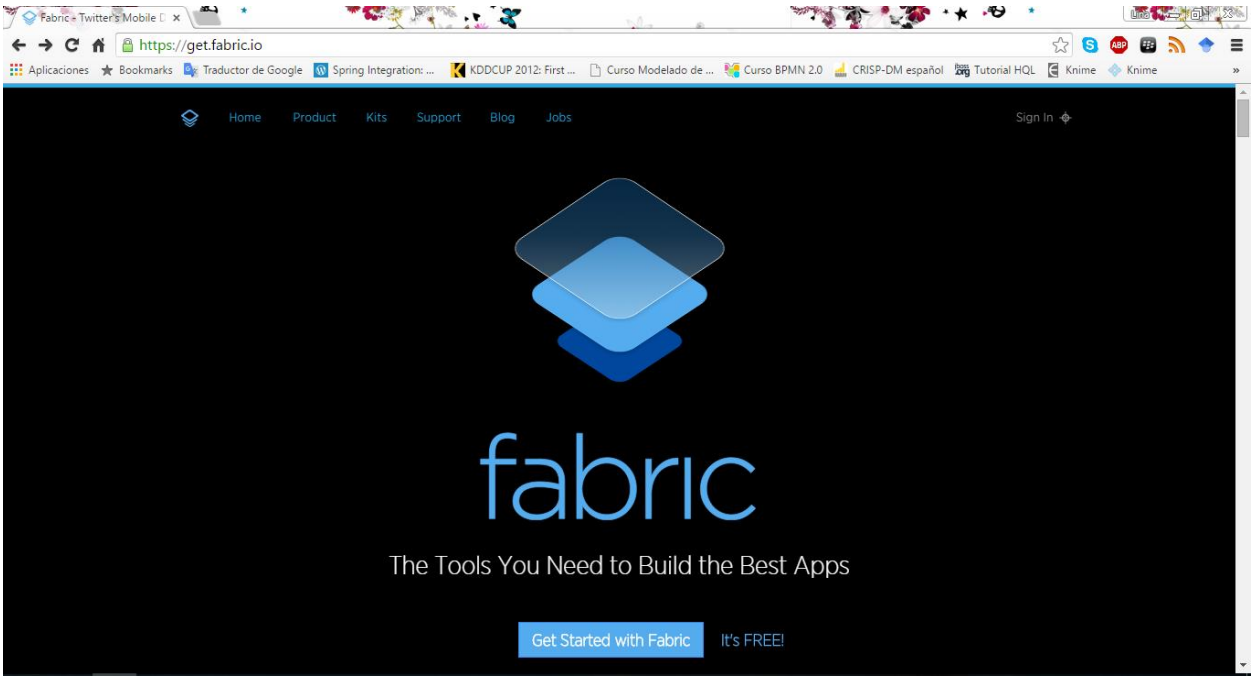


Figura 9. Interfaz fabric de Twitter

Por ser la primera vez que se va a acceder se deben ingresar las credenciales de la misma cuenta de Twitter y seleccionar la opción de chequeo para aceptar los términos y condiciones en el uso de esta plataforma.

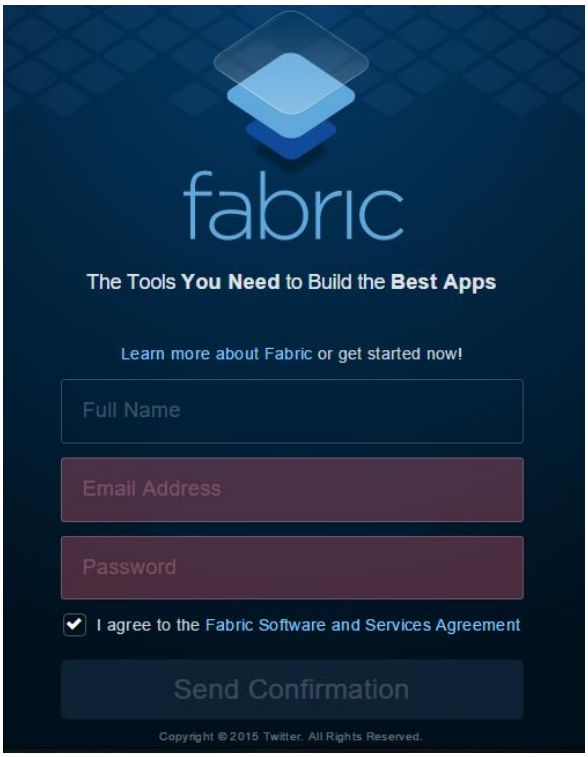


Figura 10. Inicio de sesión en fabric

Hecho lo anterior, se procede a instalar el complemento de Twitter "Fabric" en Knime yendo a la opción "Help" y luego "Install New Software".

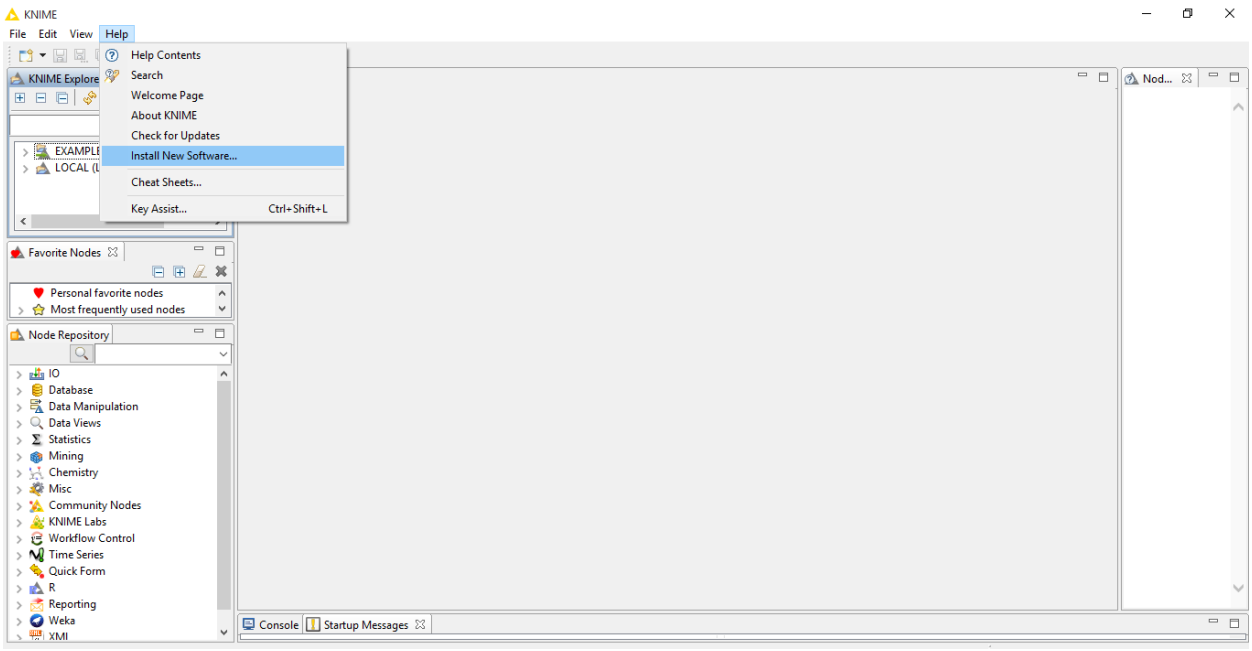


Figura 11. Instalación de Fabric en Knime

Se selecciona el botón "Add" y se añade un nuevo repositorio con el nombre Fabric y en el campo de la ubicación se digita la URL https://fabric.io/download/eclipse.

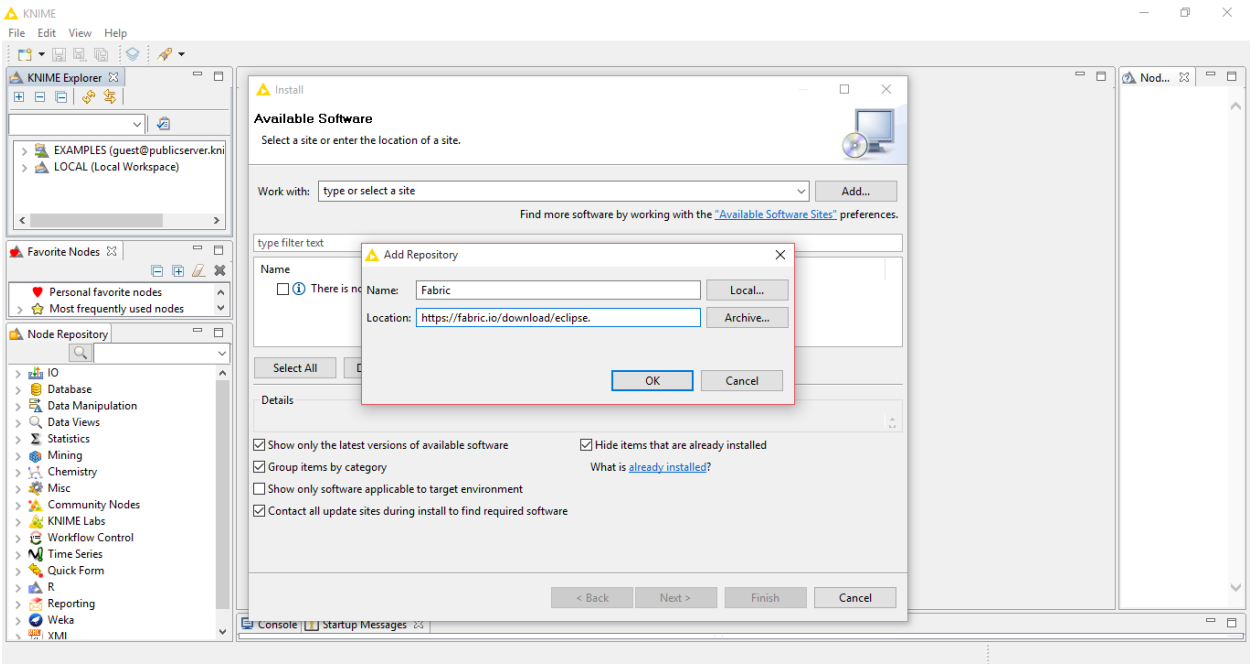


Figura 12. Configuración del repositorio

Luego se selecciona la opción "Fabric for Eclipse"

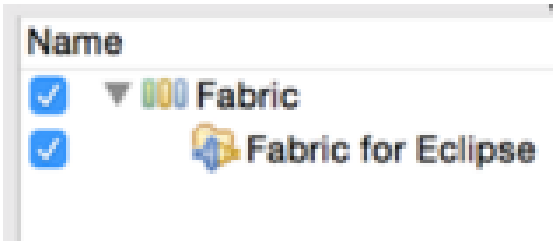


Figura 13. Instalación de Fabric en Knime

Finalmente debe aparecer el logo de Fabric en Knime.

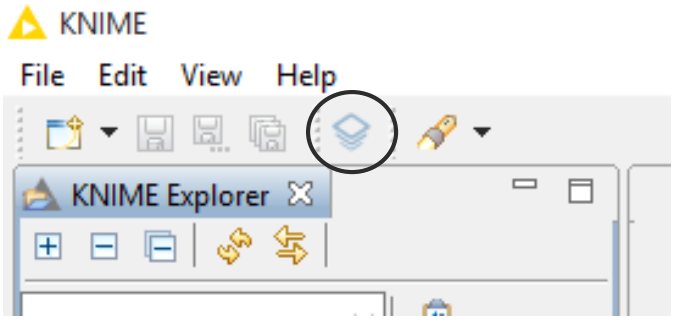


Figura 14. Fabric instalado en Knime

Cuando ya se ha instalado Fabric en Knime, se debe registrar el proyecto de minería de texto como una aplicación de Twitter accediendo a la dirección https://apps.twitter.com/ con las mismas credenciales de la cuenta inicial, se completa el nombre y la descripción de la aplicación en el formulario que aparece ahí, por último se selecciona el botón "Create your Twitter application":

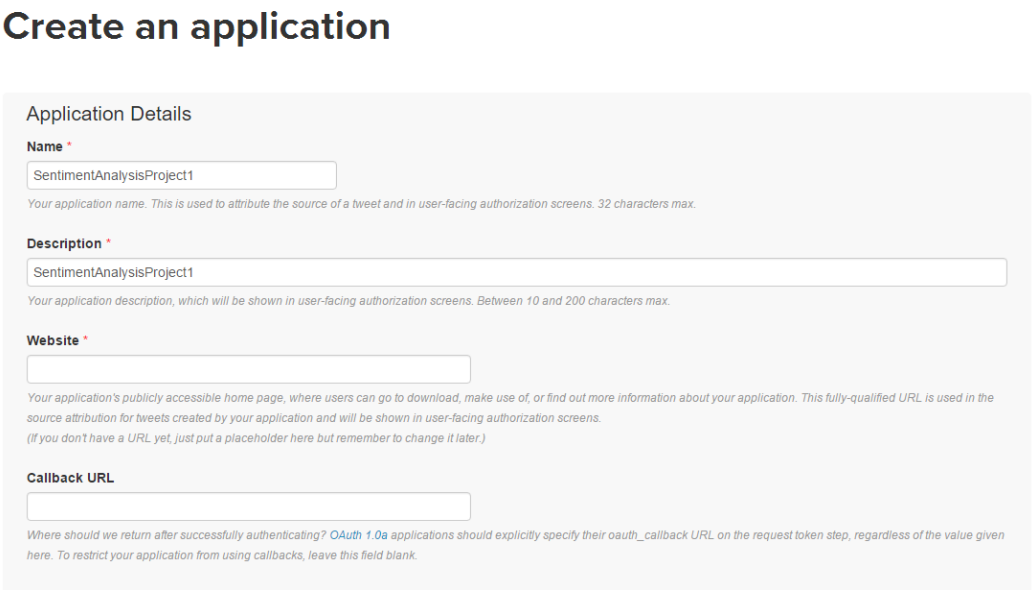


Figura 15. Crear aplicación en Twitter 1

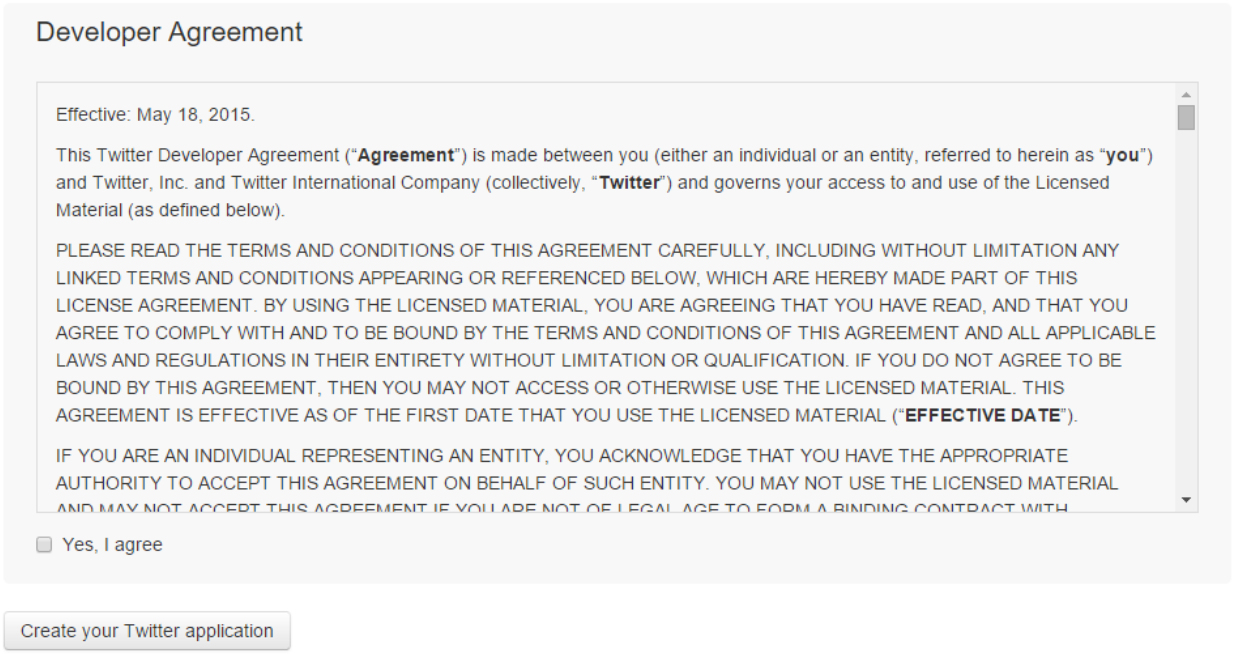


Figura 16. Crear aplicación en Twitter 2

Automáticamente se genera la siguiente información sobre la aplicación:

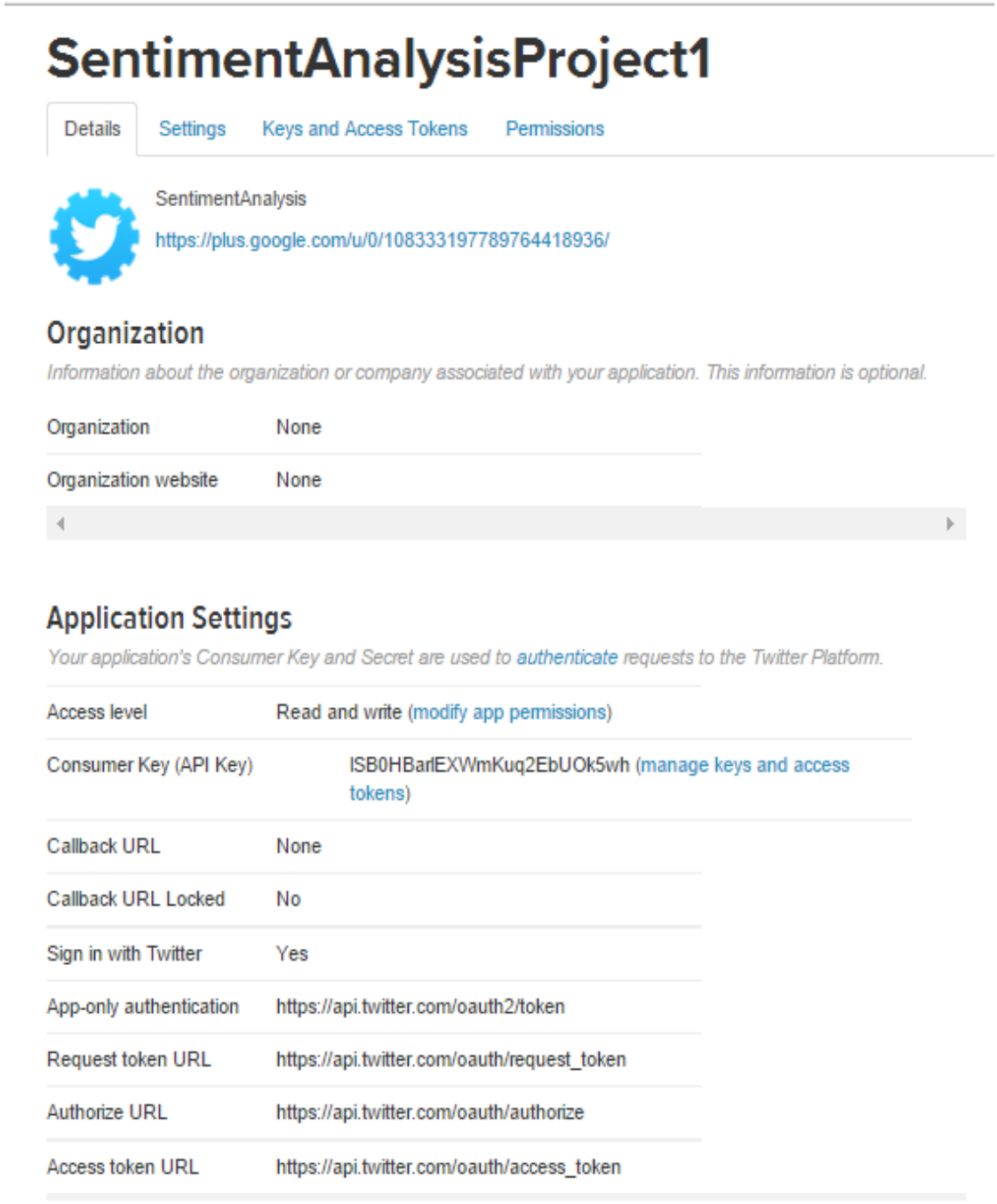


Figura 17. Detalles de la aplicación

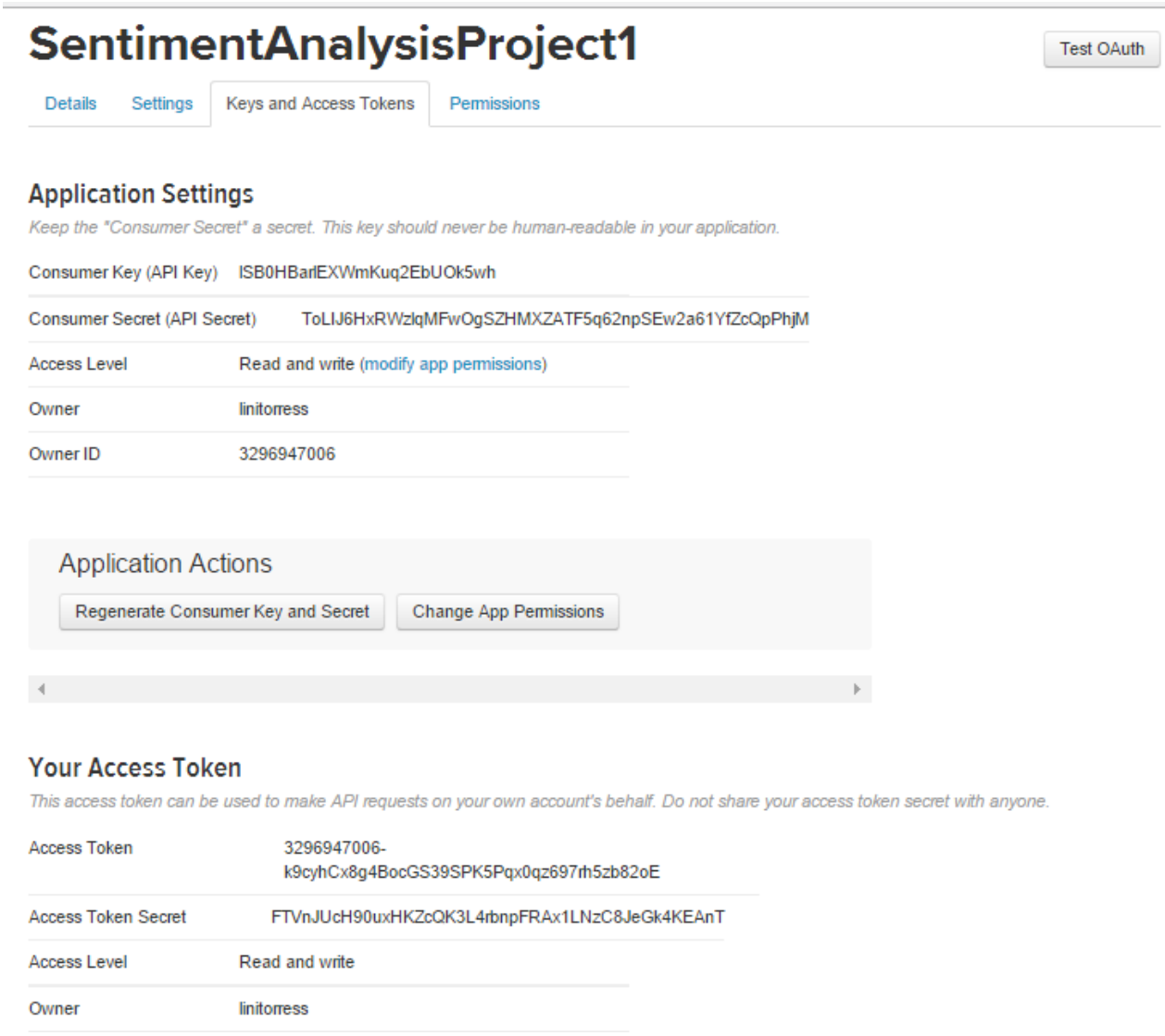


Figura 18. Llaves y Tokens de acceso

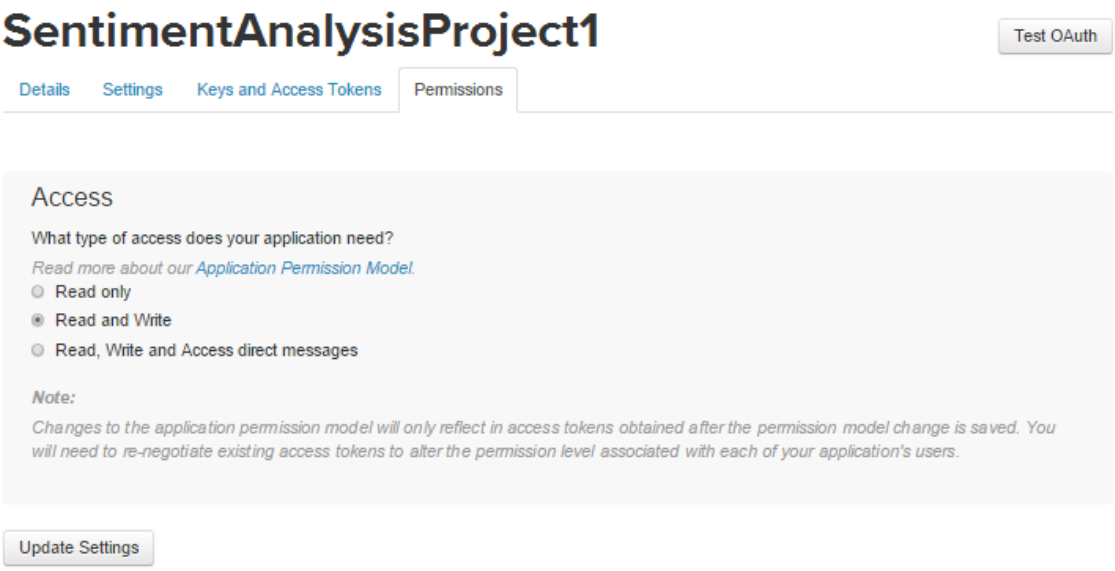


Figura 19. Permisos de lectura y escritura

Cabe mencionar que toda la configuración anterior es necesaria para que en Knime se habiliten los nodos de Twitter. De ese modo se obtiene la información de la llave y el token de acceso de la aplicación que se creó, para configurar la conexión de Twitter con a través de los nodos de conexión Twitter API Connector mostrados en las figuras 5 y 6.

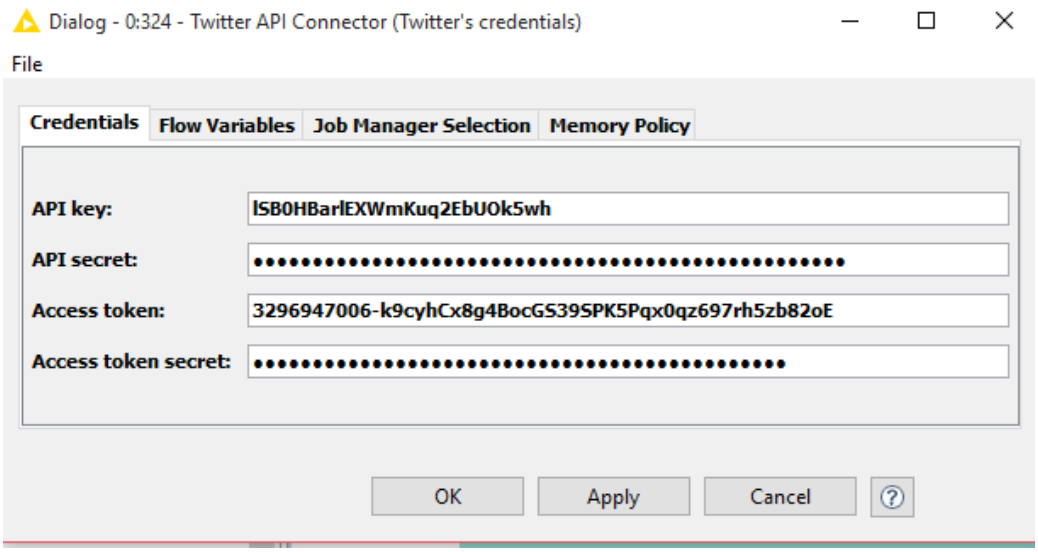


Figura 20. Configuración de credenciales de Twitter

Continuando la explicación del flujo de las figuras 5 y 6, en el nodo Twitter Search se establece la variable de la cual se obtiene el texto de la consulta y la cantidad de resultados, es decir, 10.000 filas y que además deben ser los tweets más recientes, como se muestra:

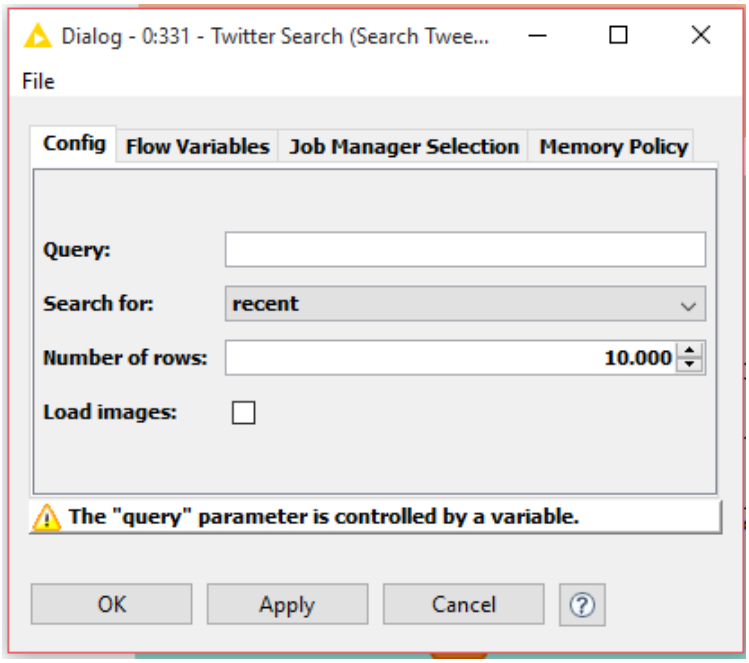


Figura 21. Configuración de nodo Twitter Search

Luego en el nodo Loop End se finaliza el ciclo y de ahí se obtiene un documento con las siguientes columnas: User, Tweet, Time, Favorited, Retweeted, Retweet from y Iteration. Así:

Collected results - 0:327 - Loop End (Loop end)

File

Table "default" - Rows: 3666 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	User	Tweet	Time	Favorited	Retweeted	Retweet from	Iteration
0#0	menriquehdezv...	RT @juiciosito: ...	2015-10-0...	0	1	juiciosito	0
1#0	juiciosito	@DanielSamper...	2015-10-0...	2	1	?	0
2#0	saularte	@ClaraLopezOb...	2015-10-0...	0	0	?	0
3#0	colombia_lider	http://t.co/hG0...	2015-09-3...	1	0	?	0
4#0	florezo_andres	Ahora @PachoS...	2015-09-2...	0	0	?	0
5#0	angelachiguasu2	RT @ClaraLope...	2015-09-2...	0	67	ClaraLopezObre	0
6#0	PepitoUsb25	RT @XiniaNavar...	2015-09-2...	0	7	XiniaNavarro	0
7#0	amanecercast	RT @XiniaNavar...	2015-09-2...	0	7	XiniaNavarro	0
8#0	CheGuevaraDiaz	RT @XiniaNavar...	2015-09-2...	0	7	XiniaNavarro	0
9#0	TransYFeliz	RT @XiniaNavar...	2015-09-2...	0	7	XiniaNavarro	0
10#0	tatislachaparra	RT @XiniaNavar...	2015-09-2...	0	7	XiniaNavarro	0
11#0	DonManrique	RT @XiniaNavar...	2015-09-2...	0	7	XiniaNavarro	0
12#0	EdwRodMon	RT @XiniaNavar...	2015-09-2...	0	7	XiniaNavarro	0
13#0	chaconparada	RT @ClaraLope...	2015-09-2...	0	67	ClaraLopezObre	0
14#0	UP_Colombia	RT @ClaraLope...	2015-09-2...	0	67	ClaraLopezObre	0
15#0	XiniaNavarro	El proceso de p...	2015-09-2...	0	7	?	0
16#0	nelsonopereira	RT @ClaraLope...	2015-09-2...	0	67	ClaraLopezObre	0
17#0	Estbanqb	RT @ClaraLope...	2015-09-2...	0	67	ClaraLopezObre	0

Figura 22. Resultados nodo Loop End

Una vez se obtienen todos los resultados del loop, en el nodo Rule-based Row Filter se crea una regla para que del documento se filtren los tweets que se repiten, es decir que permanecen aquellos que no hayan sido "re tuiteados".

El nodo Column Filter se encarga de alistar el documento y eliminar las columnas que no se van a usar, de tal forma que el documento final sólo va tener las columnas Tweet, Time y Iteration y por último, el documento se guarda en una dirección física de memoria en formato .csv.

Al final de todo el proceso de extracción se generaron un total de 27.194 registros, lo cual puede parecer un número reducido pero no es así porque Twitter establece un límite de 180 consultas por cada 15 minutos y al excederlo se genera un error HTTP 429 "Too many request" y automáticamente bloquea las consultas. Sin embargo para evitar este error y lograr extraer la mayor cantidad de consultas posibles se usó el nodo "Wait-for-time" el cual da una espera de 15 minutos para continuar el proceso, de tal forma que Twitter no genera el error.

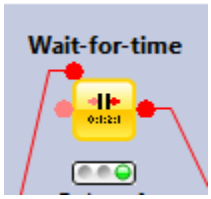


Figura 23. Nodo Wait for time

4.3 LECTURA DE INFORMACIÓN

Debido a que Knime tiene un límite para la lectura de datos desde un archivo de texto y en la extracción se generaron un total de 27.194 registros, se tuvo la necesidad de migrar todos los registros a una base de datos. Se utilizó Oracle Database 11g Express Edition por ser una licencia gratuita y se creó la instancia SentimentAnalysis con las siguientes tablas:

- OpinionMakers
- Politicians
- RssMedia
- SocialMedia

Cada tabla tiene las siguientes columnas:

- ID: De tipo Varchar2 de una longitud de 100 byte
- Tweet: De tipo Varchar2 de una longitud de 300 byte
- Time: De tipo Varchar2 de una longitud de 100 byte

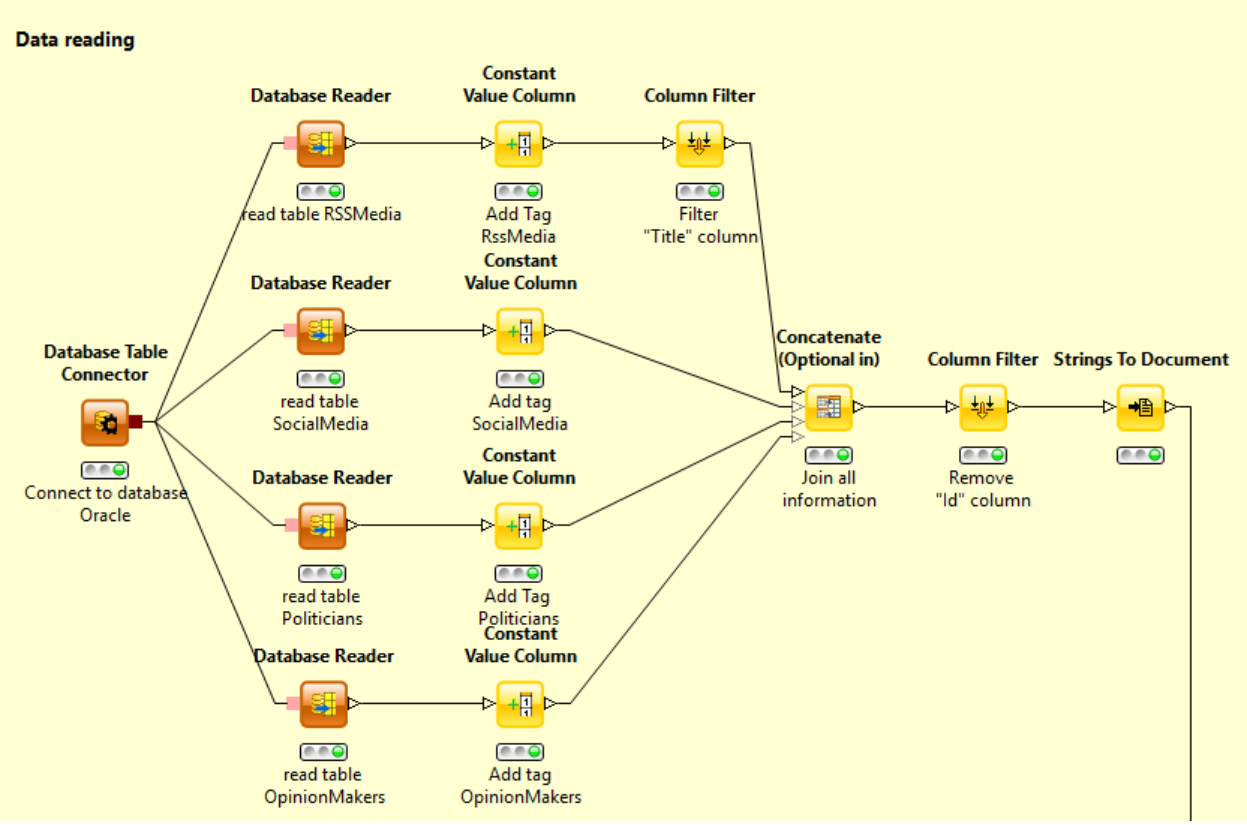


Figura 24. Lectura de información

En el nodo Database Table Connector se configuran las credenciales para acceder a la base de datos y se indican las tablas que van a ser consultadas:

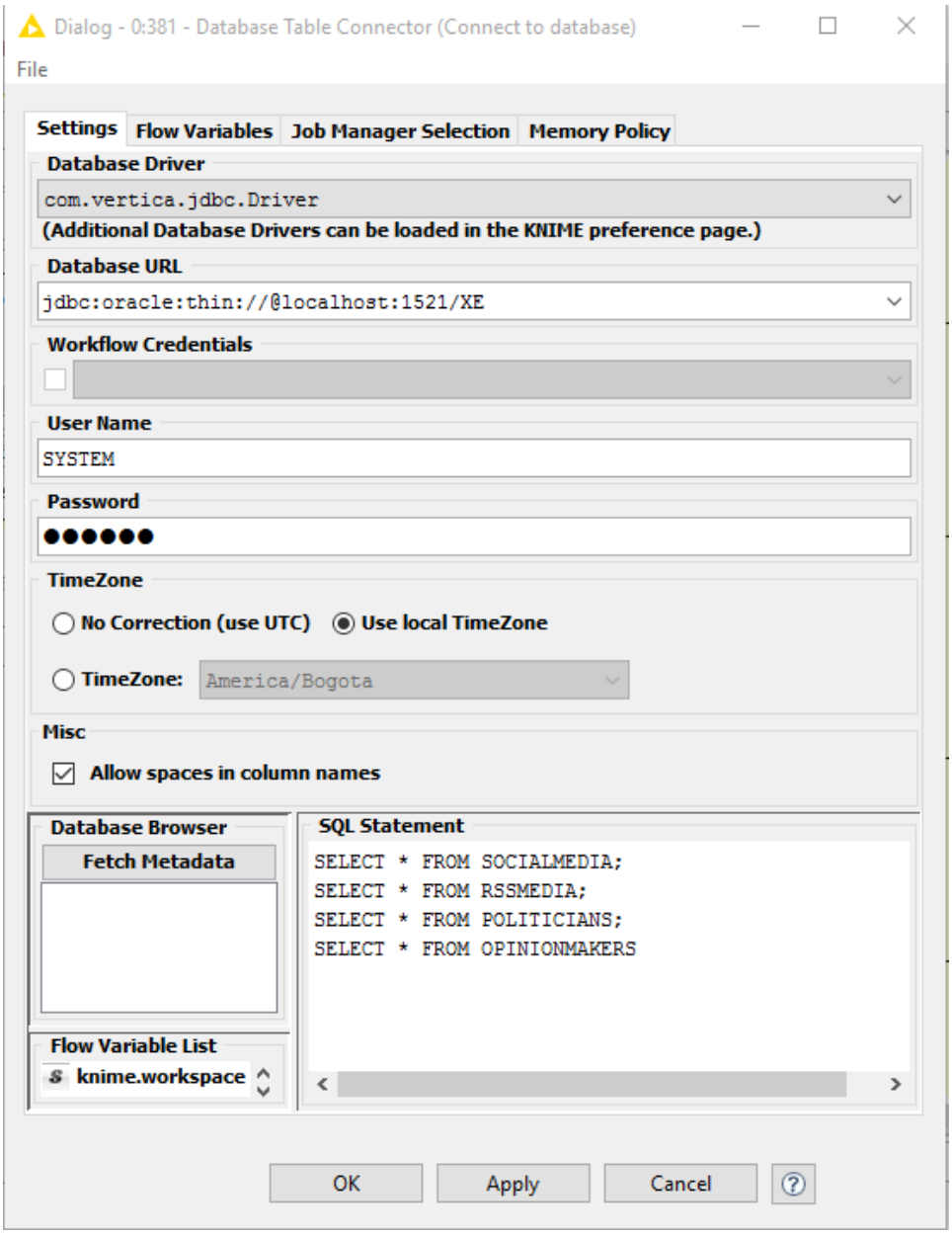


Figura 25. Configuración de credenciales de la base de datos

Luego, cada nodo Database Reader lee la tabla que le corresponde con un query sql que se debe configurar, de esta forma cada consulta será: "SELECT * FROM *NombreDeLaTabla*", por ejemplo:

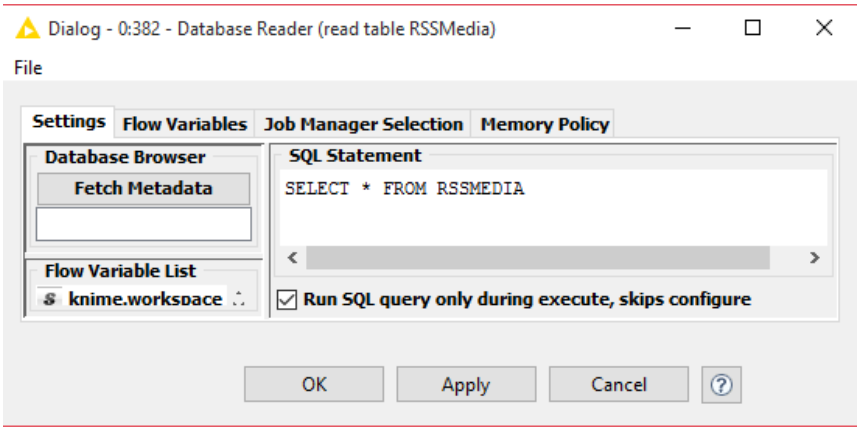


Figura 26. Query sql para leer cada tabla

El nodo Constant Value Column agrega a cada tabla una columna con un Tag, por ejemplo, la tabla RSSMEDIA va tener el tag Rssmedia. Esto con el fin de que en pasos posteriores cuando se haga el procesamiento del documento, se puedan identificar las categorías de las fuentes de origen de los datos:

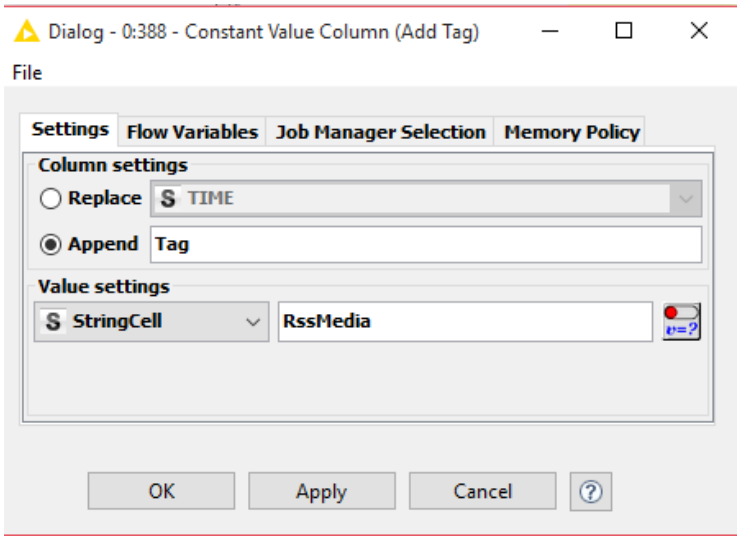


Figura 27. Configuración del nodo Constant Value Column

El nodo Concatenate (Optional in) une toda la información en una sola tabla y el nodo Column Filter elimina la columna "id", es decir que el resultado final son las columnas Tweet, Time y Tag.

Por último, Knime en su entorno utiliza los tipos de datos usuales como entero (int), doble (double), cadena (string), entre otros, pero agrega uno adicional que es exclusivamente de su propiedad y es el tipo de dato Documento (Document). Con este tipo de dato el ambiente realiza operaciones como la extracción de datos o gráficos, por ello, en esta parte del flujo se debe hacer la conversión de la columna con caracteres (string) de la tabla resultante al tipo de dato Document, de ello se encarga el nodo Strings To Document, este paso con el fin de continuar el flujo con los nodos que necesitan un tipo de dato Document.

4.4 ETIQUETAR EL SENTIMIENTO Y TRANSFORMACIÓN

En esta etapa se etiquetan los sentimientos, es decir que se obtienen las valoraciones positivas o negativas de las palabras que provienen del diccionario Spanish Emotion Lexicon (SEL).

Para obtener dicho diccionario se consultó a la Sociedad Española para el Procesamiento de Lenguaje Natural pero al no obtener una respuesta por parte de ellos, se optó por realizar una búsqueda exhaustiva en la Web tratando de obtener un diccionario lo suficientemente completo que permitiera etiquetar los sentimientos, lo cual concluyó con SEL. Esta dificultad de encontrarlo se presentó porque para el idioma español no hay un avance en este tipo de estudios y generalmente las investigaciones anteriores que se han hecho en el mundo son en otros idiomas como inglés o alemán, pero, también es muy interesante el hecho de que se presenta una oportunidad para mejorar este diccionario y por ende refinar este paso.

Este diccionario se compone de 2.036 palabras en español que están asociados con la medida del factor de probabilidad de uso Afectivo (PFA) con respecto a al menos una emoción básica: alegría, ira, miedo, tristeza, sorpresa y disgusto. De esta forma se crea un aprendizaje supervisado porque al definir una etiqueta para una palabra previamente se conoce la clase, si es positiva o negativa. Se realizó de esta manera porque muchas de las tareas necesarias para el etiquetado semántico efectivo de frases y textos se basan en una lista de palabras anotadas con algunas características semánticas léxicas que son las dadas por el diccionario.

Desde el punto de vista técnico, las palabras positivas del diccionario SEL se almacenaron en el archivo de Excel nombrado “Corpus positive words” y dichas palabras las obtiene el nodo "diccionario etiquetador" para luego buscarlas en los documentos. Los términos encontrados en los documentos se etiquetan con el tag POSITIVE. De esa forma el documento de términos y el diccionario de palabras son comparados por case insensitive por un match exacto, es decir que sin tener en cuenta si la palabra está en mayúsculas o minúsculas se comparan el documento y el diccionario para encontrar aquellas palabras que sean iguales. Entonces, se crea una bolsa de palabras y se filtra para retener sólo los términos etiquetados. De la misma manera se repite todo el proceso para las palabras negativas del diccionario SEL.

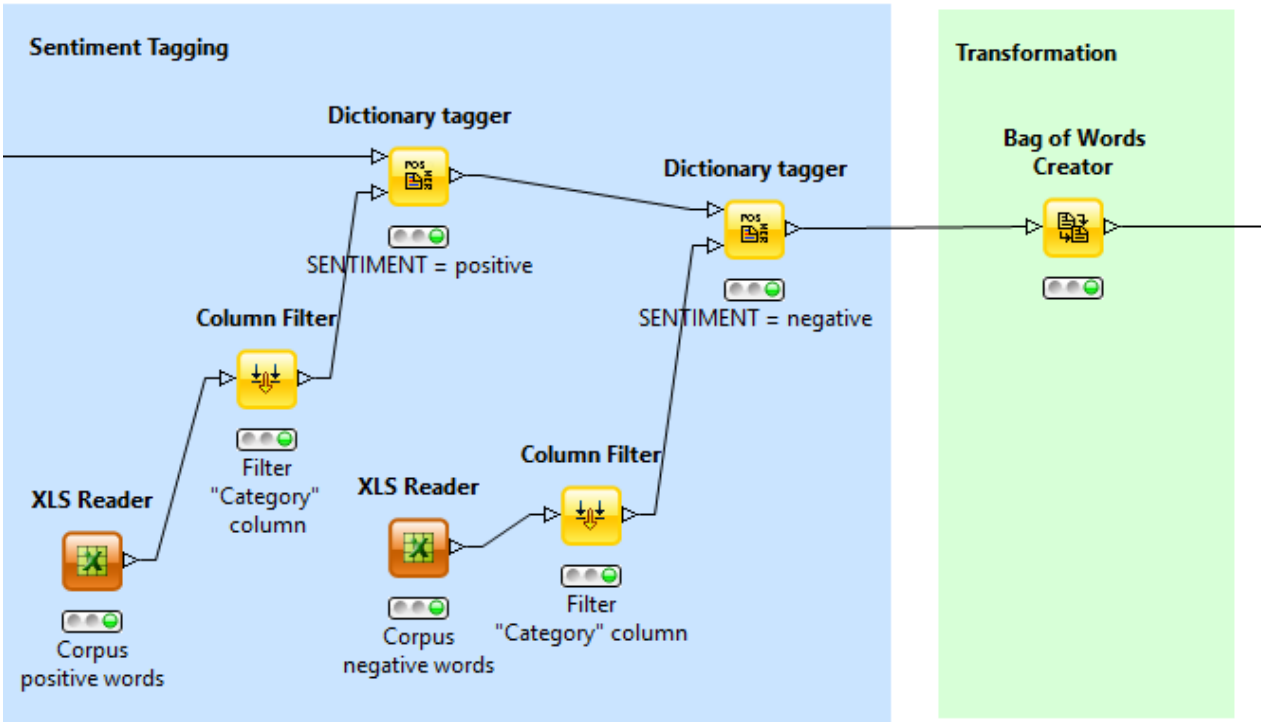


Figura 28. Diccionario de sentimientos y BoW

Inicialmente, en archivos de Excel se obtuvieron 844 palabras positivas y 1.194 palabras negativas de un diccionario (Sidorov, 2012). Los siguientes son ejemplos de algunas palabras:

Palabra	Categoría
Abundancia	Alegría
acabalar	Alegría
acallar	Alegría
acatar	Alegría
acción	Alegría
aceptable	Alegría
aceptación	Alegría
acicate	Alegría
aclamación	Alegría
aclamar	Alegría

Tabla 5. Diccionario de palabras positivas

Palabra	Categoría
abominable	Enojo
abominación	Enojo
abominar	Enojo
aborrecer	Enojo
aborrecible	Enojo

aborreciblemente	Enojo
aborrecimiento	Enojo
abusar	Enojo
acometedor	Enojo
acometer	Enojo

Tabla 6. Diccionario de palabras negativas

Es así como el nodo Dictionary tagger, reconoce las palabras dadas por el diccionario, asigna el valor a una variable y a un tipo específico. En este caso, el nodo toma la columna de nombre "Palabra" definida desde el archivo .xsl y a la variable de tipo SENTIMENT le asigna las etiquetas "POSITIVE" o "NEGATIVE".

En esta parte del flujo ocurre la transformación, se crea una bolsa de palabras "Bag of words", es decir, que la información se muestra en forma de vectores y cada documento se representa como un vector de dimensión según el número de palabras y de acuerdo con unas determinadas características, según la frecuencia con que aparezca o no en el documento. En este nodo, todas las palabras de un documento se tratan como términos índices para ese documento o que es lo mismo como un conjunto de palabras claves. Además se asigna un peso a cada término en función de su importancia, determinada normalmente por su frecuencia de aparición en el documento. De este modo, no se toma en consideración el orden, la estructura, el significado, etc. de las palabras (Vallez & Pedraza-Jimenez, 2007).

Este concepto consiste en la creación de un vector donde se muestra la frecuencia de palabras de un texto y este modelo se implementa para obtener predicciones más precisas de opiniones, se representa de la siguiente forma:

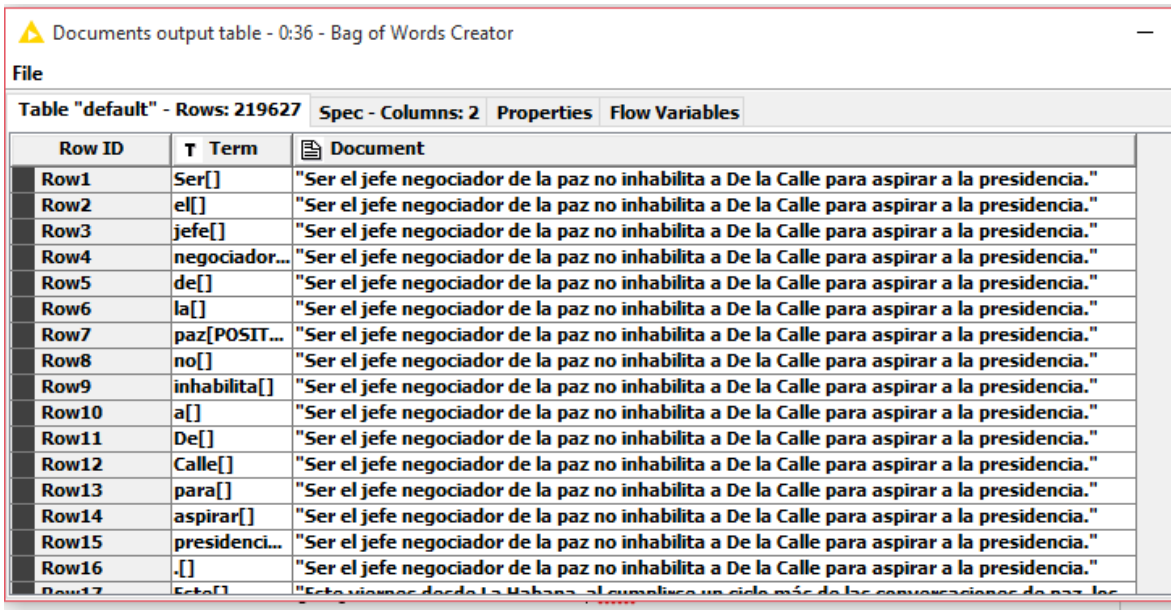


Figura 29. Ejemplo de Bag of Words

4.5 PREPROCESAMIENTO DE DATOS

El propósito del preprocesamiento de datos es principalmente corregir las inconsistencias de los datos que serán la base de análisis en procesos de minería. Con el preprocesamiento de datos se pretende que los datos que van a ser utilizados en tareas de análisis o descubrimiento de conocimiento conserven su coherencia, debido a que generalmente los datos vienen con ruido por diferentes razones, entre las cuales se encuentran (Hernández G. & Rodríguez R., 2008):

- Datos incompletos: valores faltantes para algunos atributos o sólo se tienen los datos agregados y no se cuenta con el detalle.
- Ruido: errores en los datos. Por ejemplo, manejar valores negativos para un atributo que maneja fechas.
- Inconsistencias: contiene discrepancias en los datos. Por ejemplo, edad de un empleado = 30 y fecha de nacimiento = 03/07/1998.
- Motivación cognitiva: entre estos tipos de errores se encuentran, problemas relacionados con la puesta de espacios entre los signos de puntuación, errores en la digitación de palabras (errores tipográficos), casos en que no se produce un emparejamiento correcto entre signos de interrogación, paréntesis, comillas, etc. de apertura y de cierre, errores léxicos, sintácticos, de concordancia, etc.

En muchas ocasiones, el origen de los problemas de los datos depende de la intervención humana, ya que en su momento pudieron cometer errores en la alimentación de las fuentes originales de los datos, en este caso, errores en la escritura de los Tweets.

Por otra parte, aunque la metodología CRISP-DM en su fase de preparación de los datos se ocupa de la transformación y limpieza de los datos, no desciende hasta el nivel de recomendar técnicas específicas dependiendo de la naturaleza de los datos (Uribe & Jiménez Ramírez, 2009).

Sin embargo en esta etapa se realizaron operaciones o transformaciones sobre un conjunto de documentos objetos de estudio según las razones explicadas anteriormente, algunos autores la llaman Text Refining. Este paso es muy importante ya que, dependiendo del tipo de método usado en el preprocesamiento, así mismo es el contenido de los textos que darán origen a los patrones que se descubran. En esta etapa los documentos en el flujo toman el nombre de "Forma Intermedia".

Algunas de las técnicas utilizadas para la transformación de documentos en una forma intermedia pueden ser: análisis de texto, categorización, técnicas de procesamiento de lenguaje natural (etiquetado de parte del discurso, tokenización, lematización), técnicas de extracción de información (categorización, adquisición de patrones léxico sintáctico, extracción automática de términos, localización de trozos específicos de texto), técnicas de recuperación de información (indexación) (Iribarra, 2013).

Según lo anterior, se usaron técnicas de procesamiento del lenguaje natural y de extracción de información, algunas de ellas se mencionan a continuación:

Palabras que contienen números: si un texto tiene un valor numérico se filtra automáticamente.
Signos de puntuación: en los textos que aparezcan signos de puntuación, éstos se filtran.

Palabras con mínima cantidad de caracteres: si una palabra se compone por 3 o menos caracteres se debe eliminar.
Mayúsculas/minúsculas: si los textos son muy similares pero con letras mayúsculas o minúsculas indistin tamente ('CASA ' vs' casa'), se convierte todo el texto a una sola forma.
Stop words: Si el texto contiene palabras dentro de la lista de "stop words", éstas se filtran.
Stemmer: hace uso de algoritmos de lematización, los cuales consisten en extraer los sufijos y prefijos comunes de palabras literalmente diferentes pero con una raíz común que pueden ser consideradas como un sólo término.
Rule Engine: se definen reglas de procesamiento del lenguaje donde por medio del manejo de las tablas de verdad se hace el cambio de los valores en la ocurrencia de cada término dentro del documento.

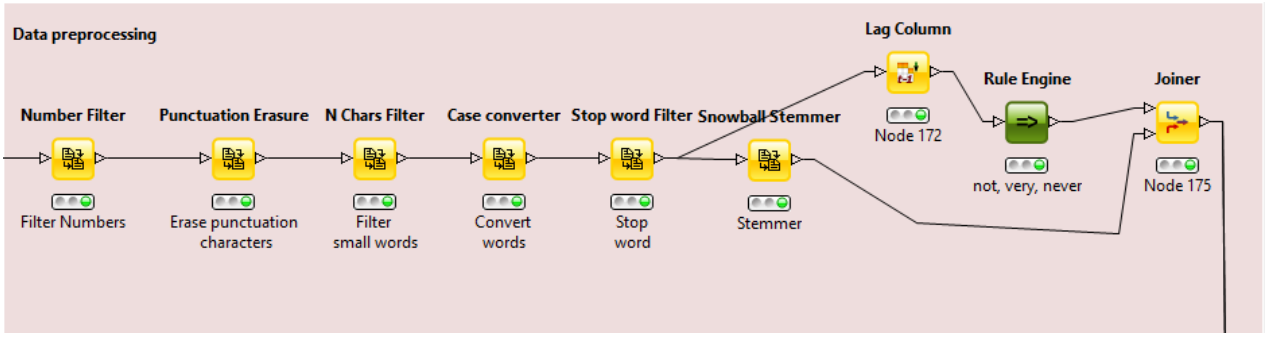


Figura 30. Pre procesamiento de datos

4.6 FRECUENCIAS

En esta parte del flujo se extraen las palabras clave relevantes de los documentos, utilizando el enfoque basado en el grafo que se describe en el KeyGraph de Yukio Ohsawa. El algoritmo de KeyGraph se basa en la segmentación de un grafo, representando la co-ocurrencia entre los términos de un documento, en grupos (clusters). Cada clúster corresponde a los mejores términos calificados por una estadística, basada en cada relación de los términos seleccionados como palabras clave (Ohsawa, Benson, & Yachida).

Luego se calcula la frecuencia de término relativo (tf) de cada término en función de cada documento y se agrega una columna que contiene el valor del tf. El valor se calcula dividiendo la frecuencia absoluta de un término de acuerdo con un documento por el número de todos los términos de ese documento.

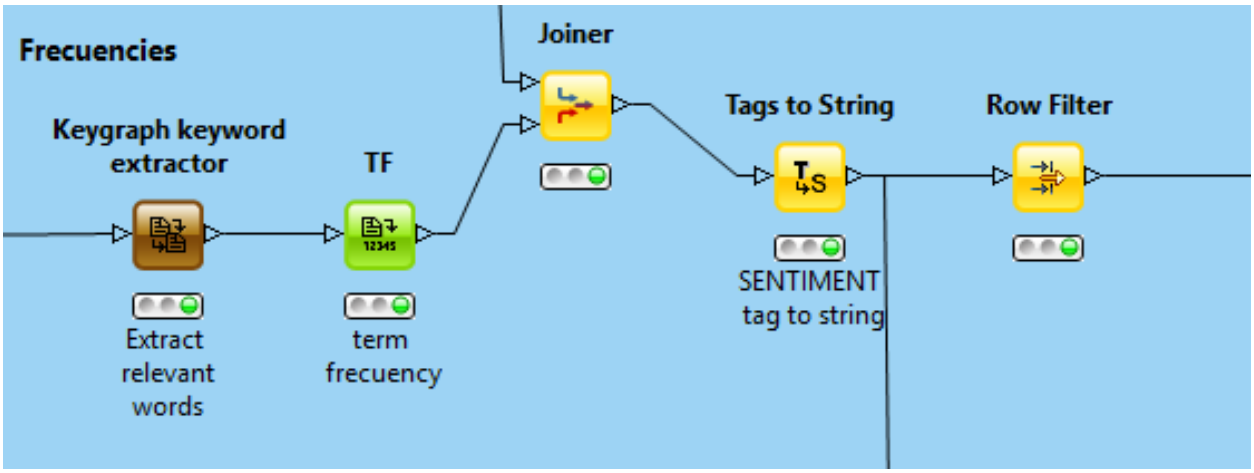


Figura 31. Cálculos de frecuencias

5. RESULTADOS

El análisis predictivo que se realizó de los perfiles de cada una de las categorías, alcanzó a recolectar un total de 27.194 registros de información entre los que se hizo una calificación de sentimientos positivos y negativos. Para obtener dichos resultados el método que se utilizó fue el aprendizaje no supervisado y fue construido a partir del etiquetado de los sentimientos con las palabras encontradas en el diccionario semántico. Haber realizado el análisis con ese modelo de supervisión hace que no haya una forma de validar con las métricas de *exhaustividad* y *precisión* los resultados, ni compararlos con otro corpus que haya sido verificado por juicio de expertos.

Por otro lado, este análisis tiene mayor valor que un pronóstico ingenuo donde se calcula la diferencia aritmética entre palabras positivas y negativas de un Tweet, porque al hacer uso del diccionario semántico se trabaja con la experiencia de los investigadores que han hecho un procesamiento del lenguaje natural serio en construirlo y en el análisis no se da lugar a la subjetividad de quien lo realiza.

En la medida que se quiera realizar la predicción con un método de aprendizaje supervisado, debe dividirse el corpus completo entre un corpus de entrenamiento y de prueba con respecto a un valor de porcentaje definido. El valor de porcentaje se utiliza para determinar el tamaño del corpus de prueba, por lo general, se toma del 70 al 80 por ciento para usarlos como corpus de entrenamiento y los registros sobrantes se toman como el corpus de prueba; de esa manera se minimizan los errores cometidos permitiendo que exista una forma de validar los resultados y obtener métricas que definan la veracidad de los resultados.

Otra forma de aplicar el aprendizaje supervisado es crear el corpus de entrenamiento a partir del juicio de expertos. El juicio de expertos es un método de validación útil para verificar la fiabilidad de una investigación que se define como "una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en éste, y que pueden dar información, evidencia, juicios y valoraciones" (Escobar-Pérez y Cuervo-Martínez, 2008:29).Tras someter el diccionario semántico a la consulta y al juicio de expertos éste debe reunir dos criterios de calidad: validez y fiabilidad. La validez de contenido se establece con frecuencia a partir de dos situaciones, una que concierne al diseño de una prueba y, la otra, a la validación de un instrumento sometido a procedimientos de

traducción y estandarización para adaptarlo a significados culturales diferentes. Es aquí donde la tarea del experto se convierte en una labor fundamental para eliminar aspectos irrelevantes, incorporar los que son imprescindibles y/o modificar aquellos que lo requieran. (Garrote & Rojas, 2016)

El juicio de expertos consistiría en etiquetar los sentimientos a partir de opiniones que puedan brindar profesionales expertos en la disciplina relacionada al tema que se desarrolla, en este caso se aplicaría un juicio de expertos del lenguaje natural y de ciencias políticas. Una vez se tenga calificado el diccionario se aplicaría el procedimiento explicado anteriormente para realizar un análisis con un método de aprendizaje supervisado. Ahora, en esta investigación no se aplicó el juicio de expertos por la ausencia de los mismos entre los recursos interesados en el desarrollo del proyecto.

A partir de la figura 32 se identificó que la mayoría de registros son originados por los medios de comunicación, sin importar que en el inicio de la extracción de información hubiera mayor cantidad de registros configurados de políticos en Twitter o de guerrilleros que de los medios de comunicación; correspondiendo estos últimos a un 86,36% de la información (23.484 registros). Según lo anterior, se puede afirmar que son la principal fuente de información y debido al volumen que generan tienen una gran presencia en el pueblo colombiano.

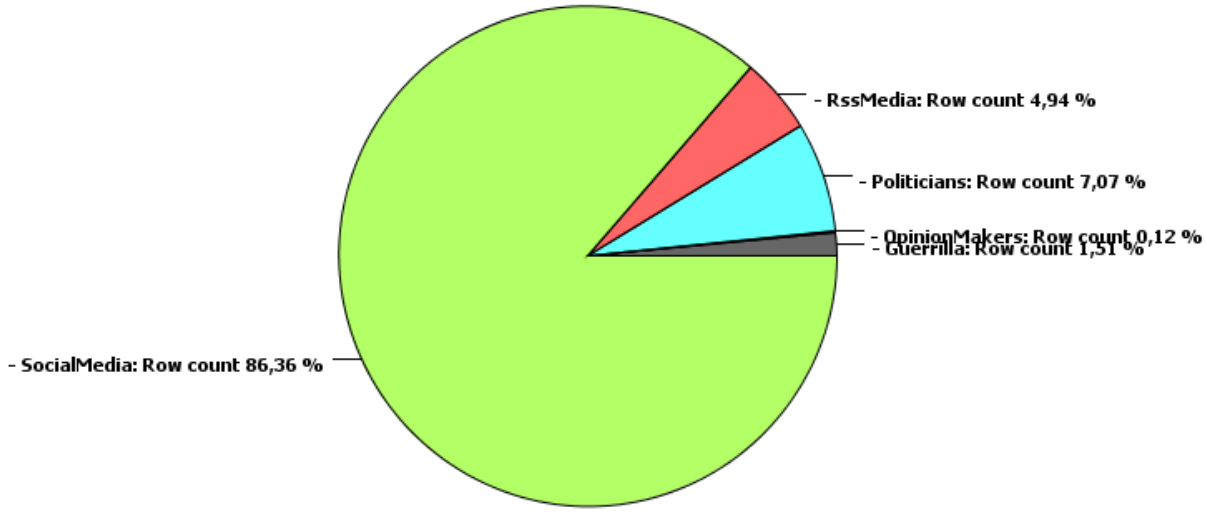


Figura 32. Porcentaje de información por categoría

Además, en la figura 33 se muestra que la mayoría de los registros producto del análisis fueron identificados como positivos con un 58.28% pero que no se alejan mucho del porcentaje restante negativo.

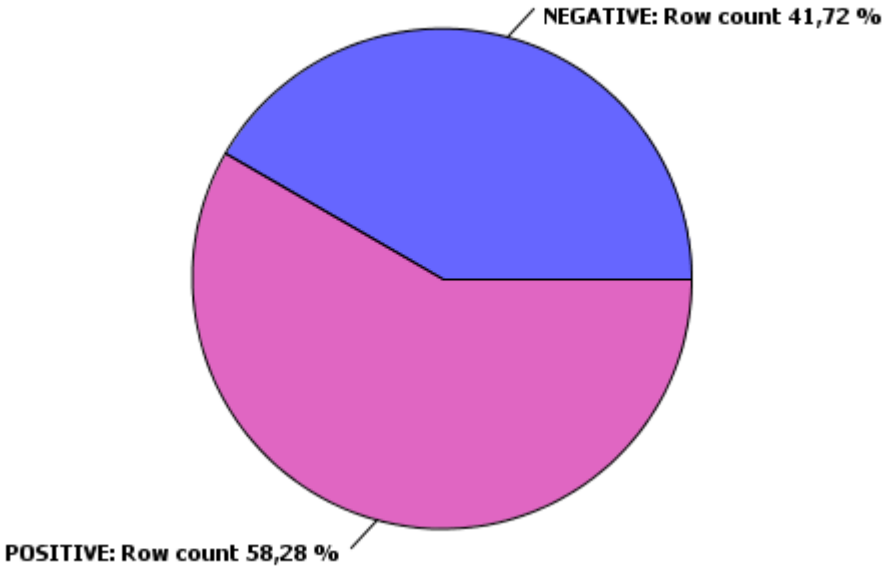


Figura 33. Porcentaje de opiniones positivas y negativas

Para obtener los perfiles de cada categoría, es decir, para poder identificar la corriente por la que los guerrilleros, políticos, generadores de opinión, medios de comunicación y medios en formato RSS tienden a manifestarse, se filtraron los sentimientos que resultaron del aprendizaje supervisado de cada categoría para obtener un tag cloud. Este proceso se muestra en las figuras 34 y 35.

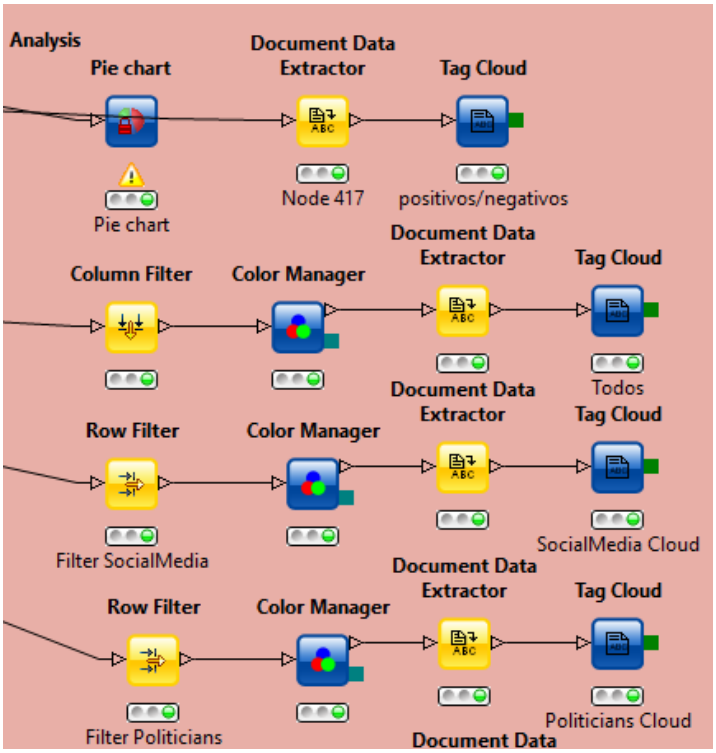


Figura 34. Obtención del tag cloud 1

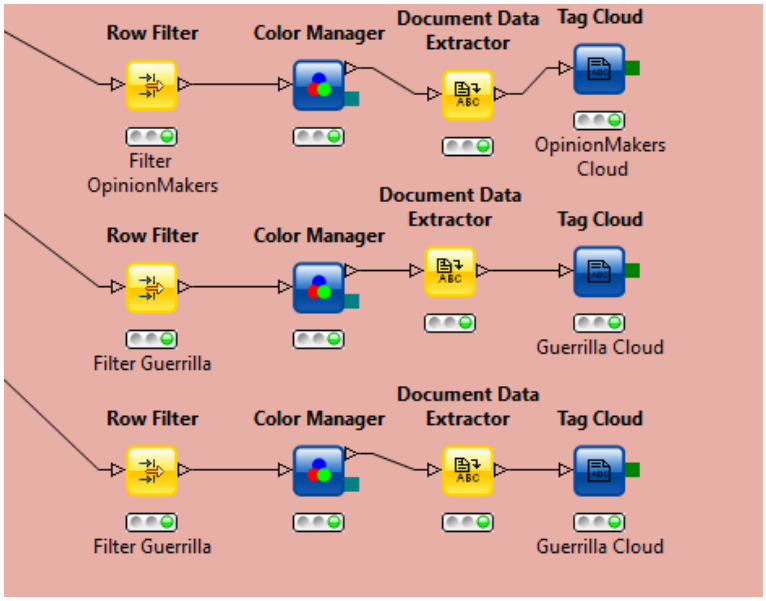


Figura 35. Obtención del tag cloud 1

Un tag cloud ("nube de etiquetas" en español) es una forma de representar la información. En este caso representa las raíces (stemming) de las palabras que el análisis de sentimientos identificó como positivas y negativas de acuerdo a la importancia (score) con que fueron calificadas al calcular la frecuencia. Las palabras se muestran mediante la manipulación de las propiedades visuales, se utiliza la distribución de tamaño de fuente, la transparencia y la audacia de la fuente.

La primera vista de la nube de etiquetas es de todo el conjunto de las palabras positivas y negativas sin discriminación por categoría con la configuración de colores de la figura 36 y el tag cloud de la figura 37.

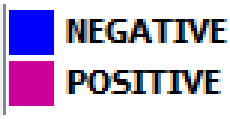


Figura 36. Configuración de colores

Figura 37. Tag Cloud: Sentimientos de todas las categorías

La raíz que predomina es *final*, sin embargo en general se muestra un equilibrio entre los sentimientos y positivos, por ello es necesario analizar cada categoría para encontrar mayores diferencias.

En las figuras 38 y 39 se muestra la configuración de los colores para cada categoría y la nube que las incluye.

Figura 38. Configuración de colores por categoría

Figura 39. Tag Cloud: Todas las categorías

La figura 39 logra dar mayor percepción de la orientación que toma cada categoría, donde la palabra *final* tiene mayor importancia y por su color rojo se puede concluir que es una opinión que se ubica entre los políticos y los guerrilleros, es decir, como un consenso de que para ambos lo más importante es el *final* del conflicto. Por otro lado, haciendo un match con la figura 37 se observa que los generadores de opinión son totalmente escépticos con la situación que se va presentar en el posconflicto, puesto que todas las palabras que los representan son negativas aunque no tienen mayor influencia sobre la generalidad de las palabras porque son muy pocas. Finalmente, quienes tienen mayores opiniones son los medios de comunicación aunque la baja tonalidad del color verde demuestra que no tienen mayor importancia, además tienen opiniones mixtas de las que no se logra diferenciar alguna tendencia hacia lo positivo o negativo.

Para el análisis del perfil de los medios de comunicación, el tag cloud de la figura 40 muestra una mayoría de palabras negativas concentradas alrededor de la palabra *final,* lo que indica que esta categoría está orientada hacia una expectativa de que de termine la guerra y la violencia, sin olvidar la reparación de las víctimas.

Figura 40.Tag Cloud: Medios de comunicación

En cuanto al tag cloud de los políticos que se muestra en la figura 41, se establece que aunque su centro e importancia es la *guerra*, son escépticos porque le temen al fracaso de los acuerdos de paz en el posconflicto por el odio, el terror, la maldad y la violencia que prevalecen entre la sociedad colombiana.

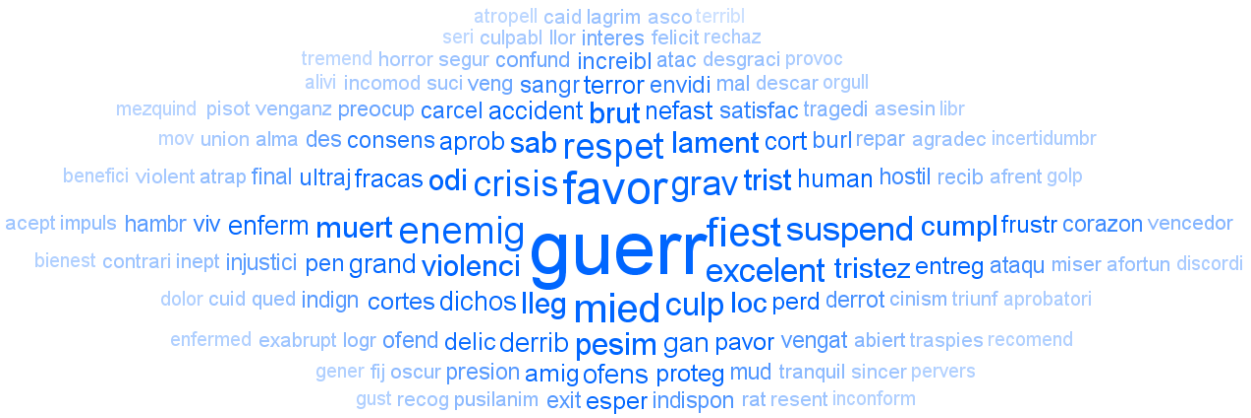


Figura 41. Tag Cloud: políticos

Luego, el tag cloud de los generadores de opinión de la figura 42 muestra solo dos palabras que se destacan: *guerra* y *miedo.* De acuerdo a lo que se observa, esta categoría no genera registros de información de mayor volumen por lo que se dificulta un análisis más profundo y, en segunda medida el 90% de las palabras son negativas con lo cual se concluye que para esta categoría el posconflicto va estar alrededor de sentimientos de culpa, miedo y reanudación de la guerra.

Figura 42. Tag Cloud: Generadores de opinión

Por último, el análisis de sentimientos de los guerrilleros se representa en el tag cloud de la figura 43 y de igual forma que la categoría de los generadores de opinión, la poca cantidad de informacion representada no permite extraer un análisis más profundo, pero da unas bases de lo que en principio son las opiniones de esta categoría. En esta nube de etiquetas sólo hay una palabra positiva y las palabras restantes son negativa, las cuales van encaminadas a que los guerrilleros tienen especial cuidado al futuro relacionado con quiénes deben estar en la cárcel y la re inserción a la sociedad de su ejército, teniendo en cuenta el rechazo y odio que pueden recibir de la sociedad civil.

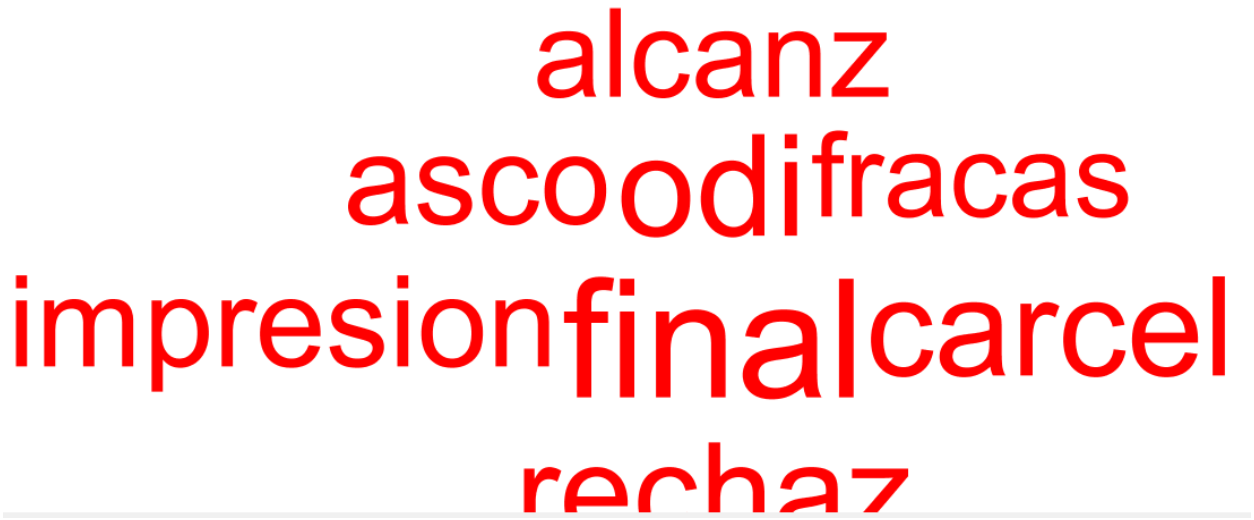
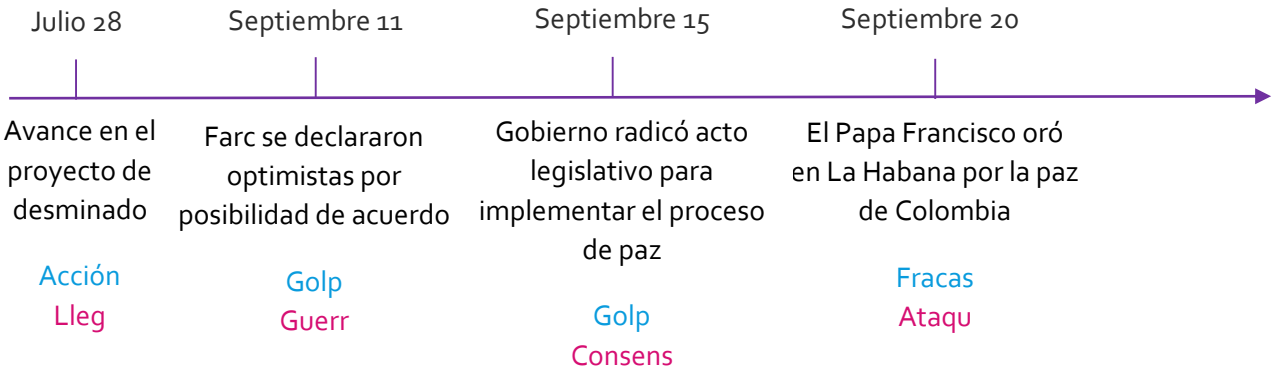


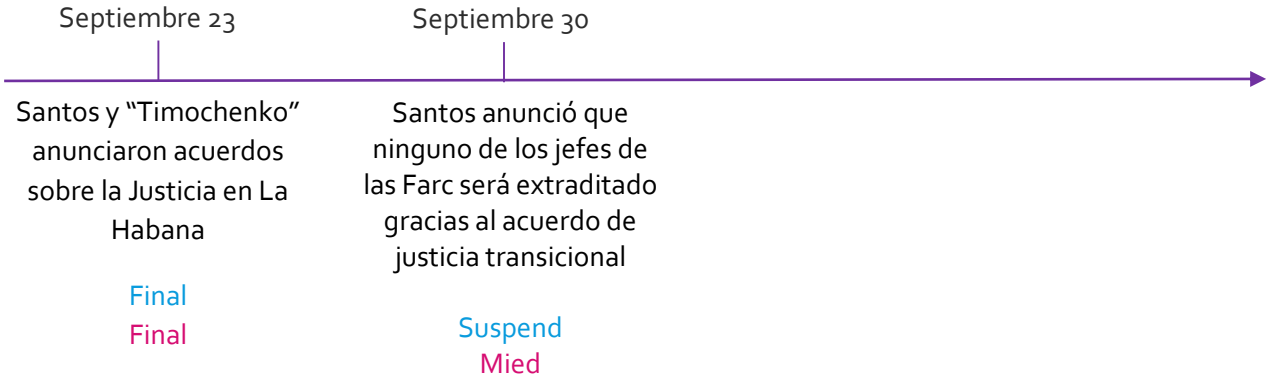
Figura 43. Tag Cloud: Guerrilleros

Después de analizar los perfiles de cada categoría, se seleccionaron hitos que se han presentado en el proceso de paz desde Julio hasta Octubre de 2015 según la infografía del periódico El Tiempo "Línea del tiempo de los diálogos de paz", para identificar cuáles sentimientos se producían entre las categorías y observar las diferencias según un acontecimiento. De esta forma se obtuvo lo siguiente:

Para hallar dichos sentimientos sólo se analizaron las fechas específicas según los hitos. Las categorías de las que se extrajo información de esas fechas se resaltan con los siguientes colores:

Medios de comunicación
Políticos
Generadores de opinión





En cada hito se observa que los sentimientos de los medios de comunicación tienden a expresar que cada acontecimiento es una decaída del proceso de paz sin importar que los hitos puedan significar acciones positivas por parte de la Guerrilla o del Gobierno. Además, dan un aire de poca esperanza en el éxito del proceso porque pareciera que cualquier hecho que ocurra va a ser algo negativo o de mucha incertidumbre.

Por otro lado, los políticos se muestran en intervalos de sentimientos negativos, muy negativos y sólo uno positivo. Existe indecisión e inseguridad con respecto al proceso de paz pero según la línea de tiempo, el sentimiento de menor incertidumbre es el consenso sobre el acto legislativo para implementar el proceso de paz, luego, el resto de los sentimientos se interpretan como si estuvieran a la defensiva de cualquier cosa que ocurre y quisieran proteger sus intereses.

Finalmente, es importante mencionar que el propósito de este trabajo es generar una herramienta que a través del tratamiento de información y haciendo uso del análisis de sentimientos de un aporte para profesiones como periodismo, comunicación social y ciencias políticas, de tal manera que la labor de generar una valoración sobre un tema específico tenga mayor soporte investigativo y tengan un mayor acercamiento al problema que están tratando.

6. TRABAJOS FUTUROS

A partir del conocimiento generado en este trabajo, se puede profundizar en mejorar la extracción masiva de informacion no estructurada desde redes sociales diferentes de Twitter, páginas web dinámicas, links, imágenes, audios y otras fuentes que no fueron contempladas. De igual forma llama la atención incluir en el procesamiento del lenguaje natural el uso de emoticones en los textos, que pueden ser manejados como "Stop words" o que ayuden a alimentar el diccionario semántico asignándoles sentimientos positivos o negativos.

También existe la posibilidad de crear la programación de la extracción de información de tal manera que el proceso sea diario, semanal o mensual de forma automática, teniendo en cuenta los límites de consultas de registros, lapsos de tiempo y permisos que establecen los buscadores.

En cuanto a la próxima selección de un diccionario semántico de sentimientos se puede complementar tomando como referencia un diccionario especializado en inglés que tenga especificado un rango de

valores para cada palabra y traducirlo al español, así se construiría un diccionario enriquecido complementando el que existe ahora.

Y por último, es interesante incluir el manejo del sarcasmo en el procesamiento del lenguaje natural porque aunque en este trabajo se hizo una aproximación sobre este aspecto fue muy básico y no se profundizo, así se obtendría un aprendizaje supervisado más acertado.

7. BIBLIOGRAFÍA

Associates, L. E. (2013). *Linguistic Inquiry and Word Count*. Obtenido de http://liwc.net/liwcespanol/index.php

Baccianella, S., Esuli, A., & Sebastiani, F. (s.f.). *SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.* Istituto di Scienza e Tecnologie dell'Informazione, Pisa.

Barrero, J. S. (31 de Mayo de 2015). *Carreta Digital*. Obtenido de Carreta Digital: http://carretadigital.com/index.php/destacado/item/292-los-8-periodistas-colombianos-mas-influyentes-en-twitter

Cardoso García, Y., & Pérez Amarillo, A. M. (s.f.). *Herramientas de minería de datos*. Recuperado el 04 de Octubre de 2015, de http://www.monografias.com/trabajos92/herramientas-mineria-datos/herramientas-mineria-datos.shtml

Cervantes, C. v. (2015). *Centro virtual cervantes*. Recuperado el 03 de 11 de 2015, de http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/linguisticacorpus.htm

Chapman, P., Clinton, J., & Kerber, R. (2000). *CRISP-DM 1.0 Step by step data mining guide.* Recuperado el 22 de Septiembre de 2015, de https://the-modeling-agency.com/crisp-dm.pdf

Córdoba Fallas, L. (2011). *Minería de datos - WEKA.* Costa Rica. Recuperado el 13 de Septiembre de 2015, de http://cor-mineriadedatos.blogspot.com.co/2011/06/weka.html

Cortizo, J. C. (13 de Mayo de 2011). *Baquía*. Obtenido de Baquía: http://www.baquia.com/tecnologia-y-negocios/entry/emprendedores/2011-05-13-mineria-de-opiniones-o-analisis-del-sentimiento

Cubero, J. C., & Berzal, F. (s.f.). *Sistemas Inteligentes de Gestión - Guía de prácticas de la minería de datos.* Granada, España. Recuperado el 13 de Septiembre de 2015, de http://elvex.ugr.es/decsai/intelligent/workbook/D1%20KNIME.pdf

Dataprix. (s.f.). *El modelo de referencia CRISP-DM.* Recuperado el 13 de Septiembre de 2015, de http://www.dataprix.com/es/el-modelo-referencia-crisp-dm

Dinero. (04 de Abril de 2014). *Dinero*. Obtenido de Dinero: http://www.dinero.com/edicion-impresa/pais/articulo/los-personajes-mas-influyentes-colombia/194261

Escobar-Pérez, J., & Angela, C.-M. (15 de Febrero de 2016). *Validez de contenido y juicio de expertos: una aproximación a su utilización* . Obtenido de Validez de contenido y juicio de expertos: una aproximación a su utilización:

http://www.humanas.unal.edu.co/psicometria/files/7113/8574/5708/Articulo3_Juicio_de_expertos_27-36.pdf

Española, R. A. (2015). *Sentimiento.* España. Recuperado el 17 de Septiembre de 2015, de http://lema.rae.es/drae/srv/search?key=sentimiento

Española, R. A. (2015). *Significado de subjetivo.* Recuperado el 17 de Septiembre de 2015, de http://www.significados.com/subjetivo/

Esquivel Gámez, I. (2009). *Eumend.net*. Recuperado el 12 de Octubre de 2015, de http://www.eumed.net/tesis-doctorales/2009/ieg/index.htm

Fernández, J., Boldrini, E., Gómez, J. M., & Martínez-Barco, P. (Septiembre de 2011). Análisis de Sentimientos y Minería de Opiniones: el corpus EmotiBlog. *Procesamiento del Lenguaje Natural, Revista nº 47*, 179-187.

Fernández, J., Miranda, N., Guerrero, R., & Piccoli, F. (2014). *Datos no estructurados no textuales: Desarrollo de nuevas tecnologías.* San Luis, Argentina. Recuperado el 17 de Septiembre de 2015, de http://sedici.unlp.edu.ar/bitstream/handle/10915/19506/Documento_completo.pdf?sequence=1

Gálvez-Pérez, J. R., Gómez-Torrero, B., Ramírez-Chávez, R. I., Sánchez-Sandoval, K. M., Castellanos-Cerda, V., García-Madrid, R., . . . Villatoro-Tello, E. (2015). *Sistema automático para la clasificación de la opinión pública generada en Twitter.* Recuperado el 15 de Agosto de 2015, de Research in Computing Science 95: http://www.micai.org/rcs/2015_95/Sistema%20automatico%20para%20la%20clasificacion%20de%20la%20opinion%20publica%20generada%20en%20Twitter.pdf

Garrote, P. R., & Rojas, M. d. (06 de Marzo de 2016). *Revista Nebrija*. Obtenido de Revista Nebrija: http://www.nebrija.com/revista-linguistica/la-validacion-por-juicio-de-expertos-dos-investigaciones-cualitativas-en-linguistica-aplicada

Gómez Sandoval, S. M., & Castillo Blanco, O. R. (Diciembre de 2006). *Clementine 9.0*. Recuperado el 04 de Octubre de 2015, de http://www.fce.unal.edu.co/uifce/pdf/MANUAL%20CLEMENTINE%209.0.pdf

Guerrera Velasco, L. P. (2008). *Primeros pasos con Knime.* Recuperado el 13 de Septiembre de 2015, de http://laurel.datsi.fi.upm.es/_media/docencia/cursos/inap/ejemplodm.pdf

Guio Fonseca, O. I. (Diciembre de 2014). *Herramienta para el análisis de sentimiento en el proceso de paz colombiano.* Recuperado el 17 de Agosto de 2015, de http://pegasus.javeriana.edu.co/~PA133-07-SentimProcPaz/recursos.html

Hearst, M. (13 de Octubre de 2003). Obtenido de http://www.ischool.berkeley.edu/~hearst/text-mining.html

Hernández G., C. L., & Rodríguez R., J. E. (20 de Abril de 2008). Preprocesamiento de datos estructurados. Structured Data Preprocessing. *Revista vínculos*, 27-48. Recuperado el 18 de Octubre de 2015, de http://revistavinculos.udistrital.edu.co/files/2012/12/Preprocesamiento.pdf

III, E. U. (s.f.). *Metodología de la minería de textos.* Madrid, Leganés, España. Recuperado el 20 de Septiembre de 2015, de http://textmining.es/contacto.html

Infórmese. (2015). *IBM SPSS MODELER.* Medellín, Antioquia, Colombia. Recuperado el 13 de Septiembre de 2015, de http://www.informese.co/mineria-de-datos/

Internacional, C. d. (18 de Febrero de 2013). *Centro de Prensa Internacional*. Recuperado el 30 de Agosto de 2015, de Centro de Prensa Internacional: http://wsp.presidencia.gov.co/cepri/medios-colombia/Paginas/default.aspx

Iribarra, F. (Junio de 2013). *Minería de Datos. Universidad Tecnológica Metropolitana*. Recuperado el 18 de Octubre de 2015, de http://mineriadatos1.blogspot.com.co/2013/06/etapas-de-la-mineria-de-texto.html

Jasso-Hernández, M., Pinto, D., Vilariño, D., & Lucero, C. (2014). Análisis de sentimientos en Twitter: impacto de las características morfológicas. *Research in Computing Science 72*.

KienYKe. (15 de febrero de 2012). *Kien y Ke*. Recuperado el 30 de Agosto de 2015, de Kien y Ke: www.kienyke.com

Leyva Abreu, L. (22 de Octubre de 2012). *Ecured. Conocimiento con todos y para todos*. Recuperado el 12 de Octubre de 2015, de http://www.ecured.cu/index.php/Clustering

Minería de textos. Sistemas avanzados de recuperación de la información. (2010). Recuperado el 2015 de Septiembre de 17, de http://mineriadetextos.tripod.com/

Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., . . . Gordon, J. (2012). *Empirical Study of Opinion Mining in Spanish Tweets.* Recuperado el 16 de Agosto de 2015, de http://www.cic.ipn.mx/~sidorov/

Ohsawa, Y., Benson, N. E., & Yachida, M. (s.f.). *KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor.* Tesis, Dept. Systems and Human Science, Graduate School of Engineering Science Osaka University,, Toyonaka. Recuperado el Octubre de 2015, de 06

Ollero Fernández, I. (2015). *Mineria de Textos o Text Mining*. Recuperado el 12 de Octubre de 2015, de http://textmining.galeon.com/

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Hanover, MA, USA.

Peñalver Martínez, I. (2015). *Minería de opiniones basada en características guiada por ontología.* Universidad de Murcia. Departamento de Informática y Sistemas, Murcia. Recuperado el 01 de Noviembre de 2015, de http://hdl.handle.net/10803/307061

Pérez Cárcamo, C. A. (2007). *Asignaturas DIICC, UdeC*. Recuperado el 16 de Noviembre de 2015, de Asignaturas DIICC, UdeC: http://asignaturas.inf.udec.cl/mt/public_html/Propuestas/2006-1/Propuestas%202006-1/claudioperez.pdf

R. García, C. (08 de 12 de 2012). *Category Management Inc. – Retail Management*. Recuperado el 16 de Noviembre de 2015, de Category Management Inc. – Retail Management: http://blogcategorymanagement.com/

Radio, C. (08 de Julio de 2012). *Caracol Radio*. Obtenido de Caracol Radio: http://caracol.com.co/radio/2012/08/07/nacional/1344321600_738497.html

Rangel, I. D., Sidorov, G., & Suárez-Guerra, S. (2014). *Creación y evaluación de un diccionario marcado con emociones y ponderado para el español.* Recuperado el 15 de Agosto de 2015, de http://www.cic.ipn.mx/~sidorov/sel_onomazein_2014_29.pdf

Rodríguez Rojas, O. (s.f.). *Metodología para el desarrollo de proyectos en minería de datos CRISP-DM* . Recuperado el 04 de Octubre de 2015, de Oldemar Rodríguez, Consultor en minería de datos: http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

Salas-Zárate, M. d., Rodríguez-García, M. Á., Almela, Á., & Valencia-García, R. (2011). *Estudio de las categorías LIWC para el análisis de sentimientos en español.* Murcia. Recuperado el 01 de Junio de 2015

Semana. (31 de Agosto de 2013). *Semana*. Recuperado el 30 de Agosto de 2015, de Semana: http://www.semana.com/nacion/articulo/20-mejores-lideres-colombia/355903-3

Semana. (20 de Julio de 2015). *Semana*. Obtenido de Semana: http://www.semana.com/tecnologia/articulo/quienes-son-los-tuiteros-mas-influyentes-de-colombia/435580-3

Sidorov, G. (07 de Diciembre de 2012). *Resource available: Spanish Emotion Lexicon*. Recuperado el 16 de Agosto de 2015, de Resource available: Spanish Emotion Lexicon: http://permalink.gmane.org/gmane.science.linguistics.corpora/16865

Suarez, W. (14 de Marzo de 2013). *Prospectador. Vigía estratégica de Redes Sociales*. Recuperado el 30 de Agosto de 2015, de Prospectador. Vigía estratégica de Redes Sociales: http://prospectador.co/21-personajes-de-la-politica-colombiana-con-mas-de-50-000-followers/

Tecnósfera, R. (21 de Julio de 2015). *El Tiempo*. Recuperado el 30 de Agosto de 2015, de El Tiempo: http://www.eltiempo.com/tecnosfera/novedades-tecnologia/tuiteros-mas-influyentes-de-colombia/16123675

Uribe, I. A., & Jiménez Ramírez, C. (09 de Junio de 2009). Hacia una metodología para la selección de técnicas de depuración de datos. (U. N. Colombia, Ed.) *Avance en sistemas e informática, Vol.6 No.1*, 185-190. Obtenido de www.redalyc.org

Vallez, M., & Pedraza-Jimenez, R. (2007). *Anuario académico sobre documentación digital y comunicación interactiva*. Recuperado el 12 de Octubre de 2015, de http://www.upf.edu/hipertextnet/numero-5/pln.html

Vargas Govea, B. (13 de Marzo de 2014). *Introducción a R con fundamentos en minería de datos*. Recuperado el 04 de Octubre de 2015, de http://es.slideshare.net/blancavg/introduccin-a-r-con-minera-de-datos

Villalta, P. A. (2015). *RapidMiner, Software Business Intelligence.* El Salvador. Recuperado el 13 de Septiembre de 2015, de http://ingenieria-en-sistemas-informaticos.blogspot.com.co/2015/03/rapid-miner-software-business-intelligence.html