« Text mining & Social Analysis Network in Twitter »

by

« Daniel Corralejo Alcuña»

A thesis submitted in conformity with the requirements
for the MSc in Economics, Finance and Computer Science

University of Huelva & International University of Andalusia

« Noviembre » « 2016 »

« Minería de texto & análisis de redes en Twitter »

« Daniel Corralejo Alcuña »

Máster en Economía, Finanzas y Computación

« Dr. Manuel Jesús Maña López »

Universidad de Huelva y Universidad Internacional de Andalucía

2016

Abstract

« Every year companies invest more of their own resources in online user behaviour analytics. This report shows the findings, optimally achieving a result from the analysis carried out over thousands of Tweets extracted between September & October 2016. Tweets were extracted from the account @Microsoft, with this in mind, a free open source software called KNIME was used for such analysis.

Text Mining techniques allowed us to determine which words are more popular among users. Sentiment analysis techniques was used to classify the polarity of a set of tweets resulting to almost 43% of the tweets were classified as a positive.

Social analysis networks helped to determine which users can be classified as influencers when a set of events, related with Big Data, were a trending topic on Twitter at that time. The network was also visually built allowing to see the relationship between users and the influential power of each user ».

Key words: Knime, Twitter, text mining, social network analysis, sentiment analysis, polarity analysis.

Resumen

« El análisis del comportamiento de los usuarios en redes sociales está cobrando gran importancia entre las empresas. Este trabajo es el resultado del análisis de miles de tweets extraídos de la cuenta @Microsoft entre Septiembre de 2016 y Octubre de 2016 utilizando como herramienta de análisis el software de código abierto KNIME.

El uso de técnicas de minería permiten la extracción de las palabras más usadas por los usuarios. Además gracias al análisis de sentimientos se descubre que en torno al 43% de los tweets analizados tienen una connotación positiva.

El uso de técnicas para el análisis de redes sociales permiten la identificación de los usuarios más influyentes en torno a unos eventos cuya temática está relacionada con el Big Data. También se reconstruye visualmente las conexiones entre usuarios y el poder de influencia que unos tienen sobre otros ».

Agradecimientos

« Me gustaría agradecer al profesor Dr. Manuel Jesús Maña López por colaborar en este TFM.

También me gustaría agradecer a mi familia y a mi novia Patricia el apoyo mostrado todo este tiempo ».

Tabla de contenidos

Parte I: Introducción

Este trabajo tiene como **objetivo** aportar información de utilidad a las áreas de Marketing y otra u otras áreas de tomas de decisiones. Del análisis de dichas redes sociales y de los resultados obtenidos estos pueden ser de ayuda para la toma de decisiones estratégicas en el futuro en el desarrollo de sus productos y servicios o mejora de la percepción de la marca.

La **minería de texto** se ha convertido recientemente en una de las técnicas más prometedoras en el campo del marketing [1]. Las técnicas de minería de texto y análisis del sentimiento se basan en la utilización de métodos computacionales aplicado a textos que se desean analizar. Debido al aumento de la utilización de redes sociales en Internet[1] tales como Twitter, Facebook, LinkedIn y muchas más, cada vez son más las empresas y organizaciones que utilizan este tipo de técnicas para comprender mejor tanto el lenguaje utilizado por los usuarios como su comportamiento en dichas redes sociales. El incremento exponencial del volumen de datos creados a través de estas redes sociales ha llevado a muchas organizaciones a invertir elevadas cantidades de dinero para entender o comprender mejor las necesidades o comportamientos de sus consumidores o grupos de individuos a los cuales van dirigido sus productos o servicios [2].

Para analizar el sentimiento de usuarios en Twitter se utiliza el método de **clasificación de polaridad**. En Twitter, a diferencia de otras plataformas tales como foros de opinión, la cantidad de texto contenida en un mensaje está limitada a 140 caracteres. La disponibilidad de sólo 140 caracteres dificulta la tarea de clasificar la polaridad de dicho texto pues se trabaja con menos información. En Twitter se suele hacer uso de vulgarismos de Internet, sacármos o alegorías a diferencia de los textos de opinión escritos en foros donde el lenguaje suele ser más claro, rico y extenso. Debido a lo mencionado anteriormente el análisis de sentimientos en Twitter se limita a la clasificación de los tweets según su polaridad en positivos ó negativos. En la clasificación de sentimientos de textos extraídos de foros de opinión el grado de clasificación de los mismos se realiza con más detalle y mayor extensión [3]. En este trabajo la clasificación de la polaridad de tweets hace uso de la técnica denominada como **clasificación de sentimientos a nivel de documento - document-level sentiment classification -**.

[1] Las redes sociales en Internet comienza su mayor periodo de expansión en el año 2002 con el lanzamiento de Friendster. En el año 2016 se calcula que hay total de 2,3 billones de usuarios activos en alguna de las muchas redes sociales disponibles.

En este trabajo también se verá cómo es posible aislar individualmente palabras contenidas en el conjunto de tweets y crear una **nube de palabras - tag cloud -**. La nube de palabras consiste en una representación gráfica y visual de aquellas palabras que más se repiten a lo largo del conjunto textos utilizados. Una de las ventajas que ofrece las nubes de palabras es la facilidad para detectar visualmente aquellas palabras relacionadas con una determinada temática, producto o noticias de interés [4]. También adicionalmente este tipo de análisis son útiles en lo que se denomina búsqueda de **palabras claves - keywords -** para llevar a cabo mejoras en SEO[2] o posicionamiento orgánico de páginas webs en los resultados de motores de búsqueda tales como Google, Yahoo o Bing [5].

El último tipo de análisis que se realiza en este trabajo tiene como objetivo la **detección de usuarios influyentes** en Twitter. El análisis de redes sociales o **Social Network Analysis (SNA)** será la metodología seguida. Se identificarán líderes de opinión o usuarios con mayor tasa de influencia sobre el resto de la comunidad. Se construirá una red que determine qué tipo de relaciones mantienen los usuarios en la red social Twitter y cuál fuerte estas relaciones son. Comprender el comportamiento de estas redes de usuarios puede resultar de gran utilidad para comprender mejor el comportamiento de los usuarios cuando estos hacen uso de las redes sociales [6].

La herramienta que será utilizada para la realización de todos los análisis y cálculos se llama **KNIME.** KNIME es una plataforma de código abierto de análisis de datos, generación de informes y plataforma integral con varias tipologías de software. KNIME integra varios componentes de **machine learning** y **data mining** a través de datos modulares. Una interfaz gráfica permite la conexión de llamados nodos disponibles en las librerías. Los nodos son herramientas o extensiones programados para realizar funciones concretas como leer tablas, filtrar columnas, introducir códigos java, *etcétera*. Todo trabajo realizado en KNIME se guarda en lo que se denomina **workflow** que es el entorno gráfico parecido a un mapa donde se muestra los nodos conectados entre sí, así como comentarios del autor para comprender la función que realiza dicho workflow.

En la figura 1 se muestra un ejemplo de workflow.

[2] Search Engine Optimization (SEO) es un conjunto de métodos estratégicos, técnicos y tácticos usados para atraer más tráfico a una página web mediante la mejora de su posicionamiento en la página de resultado en los motores de búsqueda.

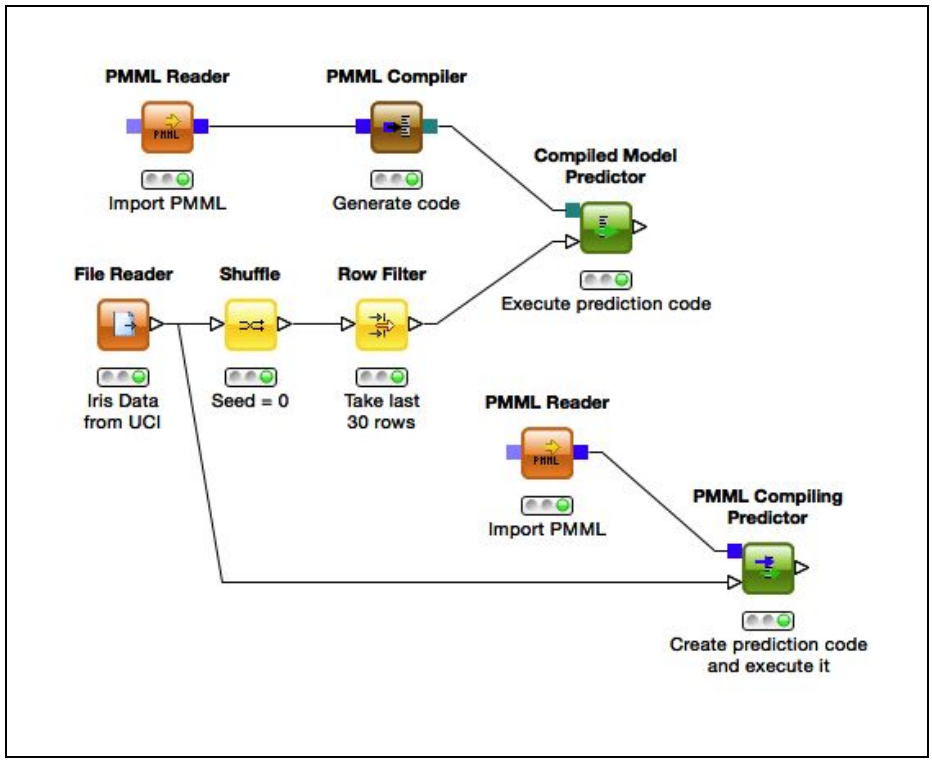

Figura 1. Ejemplo workflow en KNIME

En resumen a lo anteriormente y como forma de síntesis en este trabajo se realizan los siguientes tipos de análisis en la red social Twitter:

- Nube de palabras a partir de tweets
- Análisis y clasificación de la polaridad de tweets
- Construcción de red de usuarios influyentes

El los siguientes capítulos y sus correspondiente secciones tienen como objetivo: primero, dar detalles sobre el conjunto de datos utilizados y la metodología usada para su extracción y segundo, explicar la metodología seguida así como las técnicas usadas en la realización de los análisis.

Parte II: Datos y metodología

En esta parte se explica y se da detalles al lector acerca de los datos utilizados y la metodología tanto conceptual como técnica seguida en la realización de los distintos tipos de análisis.

Los resultados finales son mostrados y explicados en la parte tercera de este trabajo.

2.1. Dataset

Se trabaja con dos bases de datos, ambas extraídas de Twitter.

La **primera** base de datos contiene en total unos **11.443 tweets** extraídos de la cuenta de @Microsoft y menciones a la misma. Los tweets están fechados entre 1 de Septiembre 2016 y 10 de Octubre de 2016. Dicha base de datos es utilizada para los análisis de creación de **nube de palabras** y clasificación de la **polaridad de los tweets.**

La base de datos contiene las siguientes variables:

- **User:** Nombre del usuario
- **Tweet:** Tweet & retweets[3] publicados en el muro de usuario
- **Time:** Hora y fecha de la publicación del tweet

La **segunda** base de datos cuenta con un total de **8.271 tweets** extraídos entre 25 de Octubre de 2016 y el 1 de Noviembre de 2016 de la cuenta de @Microsoft y menciones a la misma. Durante tales fechas una serie de eventos relacionados con el Big Data, análisis de datos y desarrollos de nuevas tecnologías fueron muy comentados en Twitter. Esta base de datos es utilizada para la construcción de la **red de usuarios influyentes**.

La base de datos contiene las siguientes variables:

- **User:** Nombre del usuario
- **Tweet:** Tweet & retweets publicados en el muro de usuario

[3] Se dice que un usuario realiza un retweet cuando este publicó en su muro un tweet que fue originariamente creado por otro usuario.

- **Time:** Hora y fecha de la publicación del tweet
- **Favorited:** Número total de veces que el tweet ha sido marcado como favorito
- **Retweeted:** Número total de veces en que el tweet extraído ha sido retweet
- **Retweet from:** Cuenta de procedencia del tweet cuando este que ha sido republicado por otro usuario (retweeted)

En el siguiente capítulo se explica brevemente cuál es la metodología seguida en la extracción de los tweets.

2.2. Extracción de Tweets

KNIME ofrece la posibilidad de trabajar con la **API[4] oficial de Twitter**, lo que agiliza el proceso de extracción de tweets, ya que esto permite conectarse directamente a la base de datos oficial de Twitter. A continuación se detalla cómo se realiza la extracción de los datos que serán utilizados en el desarrollo de este trabajo.

2.2.1. Creación de base de datos

Los datos son guardados en una base de datos tipo **SQLite.** Para la utilización de la base de datos es imprescindible crear una tabla vacía compuesta de tantas columnas como variables se extraen de Twitter.

Esta base de datos ofrece la posibilidad de actualizar la misma con nuevos datos evitando la redundancia de los mismos pues permite descartar tweets que se encuentren duplicados, esto es, tweets que han sido realizados por un mismo usuario a la misma hora y cuyo contenido son exactamente iguales. Este tipo de opciones son útiles si la extracción de tweets se lleva a cabo en distintos días durante un periodo de tiempo determinado utilizando para ello el mismo término de búsqueda.

2.2.2. Twitter API & Twitter Search para la extracción de datos

Para la utilización de la API oficial de Twitter es imprescindible previamente crearse una cuenta para desarrolladores en la página oficial de Twitter. Una vez se ha creado la cuenta el

[4] Application programming interface (API) es un conjunto de rutinas, protocolos y herramientas para la creación de software y aplicaciones.

nodo **"Twitter Search"** es configurado con aquellos parámetros o variables que se deseen extraer, esto es, nombre de usuario, hora de publicación del tweet, número ID del usuario, *etcétera*.

2.2.3. Guardado de datos en la base de datos

Los datos son guardados en la tabla vacía que fue previamente prediseñada.

El tiempo ejecución en la extracción de datos siempre dependerá de la cantidad requerida y las fechas de publicación de los tweets.

Una vez ya se dispone de todos lo datos que se requieren para los análisis es posible comenzar con los mismos.

2.3. Creación de la nube de palabras

Uno de los objetivos de este trabajo consiste en generar una nube de palabras a partir del conjunto de tweets extraídos. La nube de palabras es un conjunto de palabras representadas de forma visual. En el lenguaje propio de KNIME, y en la parte teórica de la minería de texto, las palabras a título individual contenidas en un texto son referidas o llamadas como **términos**.

En la nube de palabras se distinguen dos atributos del conjunto de datos. Un primer atributo es representado por el **tamaño fuente** de cada término, el cual indica el **peso relativo** que cada término tiene respecto al conjunto total. El segundo atributo es representado mediante el **color** asignado a cada término, el cual indica la **categoría gramatical** a la que pertenece dicho término. En este trabajo se centra la atención en adjetivos, nombres y verbos pues suelen ser las palabras usadas para descripciones, lugares y acciones.

Una nube de palabras no deja de ser la representación visual de una **tabla** compuesta de dos columnas, una columna tipo cadena llamada **"columna término"** donde están almacenados los términos a representar y otra columna tipo numérico llamada **"columna valor"** donde se almacena el peso numérico o frecuencia relativa de cada término.

A continuación se explica todos los pasos a seguir para generar la nube de palabras.

2.3.1. Lectura de datos en la base de datos

La lectura de datos se realiza mediante un query en formato **SQL** a la base de datos SQLite.

Llamamos a la base de datos mediante la siguiente línea de comandos:

*SELECT * FROM microsoft,* donde *microsoft* es el nombre de la tabla SQLite donde se encuentran almacenados los datos.

2.3.2. Filtrado y preprocesado de términos

En esta parte es necesario filtrar y preprocesar los tweets de donde se realiza la extracción de términos. Del filtrado y preprocesado de tweets dependerá la **calidad** de la nube de palabras. Se debe asegurar que sólo términos que puedan aportar información relevante sean mostrados en el resultado final.

El preprocesado y filtrado de tweets se caracteriza por la **eliminación o filtración** de ciertos atributos gramaticales, palabras con un determinado número de caracteres, cualquier tipo simbología, números y otro tipo de datos que en cierta medida pueden ser prescindibles.

Otra de las razones por la cual la parte de filtrado y preprocesado de tweets es importante es porque así se asegura la exactitud del cálculo de frecuencia relativa de cada término.

A continuación se explica cómo se realiza y lleva a cabo el preprocesado y filtrado de tweets el cual estará dividido en varias etapas secuenciales:

Eliminación de links, hashtags y menciones

La eliminación de links, hashtags[5] y menciones[6] tiene como objetivo limpiar los tweets de todo tipo de enlaces externos, propiedades y comandos que son usados frecuentemente en Twitter con otras finalidades.

Filtrado de tweets

En esta fase se lleva a cabo una serie de filtrados: Primero, se eliminan **puntuaciones** tales como puntos, comas u otro tipo de simbología. Segundo, se eliminan términos compuestos

[5] Una palabra o frase precedida por el símbolo hash # es usado en Twitter para identificar mensajes de una temática específica.
[6] Una mención en Twitter se realiza cuando se incluye el símbolo @ más el nombre del usuario. Las respuestas a un usuario son consideradas como menciones.

por hasta un máximo de **2 caracteres** de forma que se descartan pronombres, artículos y otro tipo de categorías gramaticales con poco valor informativo. Tercero, se eliminan **números y símbolos** matemáticos. Cuarto, se transforman todas las palabras **minúsculas** para evitar problemas de sensibilidad a capitalización en el uso posteriormente de otros nodos. Quinto, se eliminan los **espacios** entre palabras para facilitar la extracción de términos en pasos posteriores.

Etiquetado y creación de la bolsa de palabras

Los términos son etiquetados según su categoría gramatical y luego estos son guardados individualmente en una tabla, es decir, los tweets son "despedazados".

El etiquetado de términos se realiza mediante la utilización del nodo **"Stanford tagger"**. Stanford tagger es un diccionario que asigna a cada término una parte del discurso - **part of the speech (POS)** -. El proceso y utilización de este diccionario se conoce como **enriquecimiento del lenguaje.** Este proceso está basado en el uso de un tipo de algoritmo desarrollado por la Universidad de Stanford que mediante la utilización de modelos computacionales consigue identificar categorías gramaticales dentro de una cadena de texto [7]. Etiquetar los términos con todas las posibles categorías gramaticales permitirá posteriormente filtrar sólo aquellos términos que han sido etiquetados como nombres, adjetivos y verbos.

Una vez filtradas las categorías gramaticales se procede a la creación de lo que se denomina **bolsa de palabras**. La bolsa de palabras permite la creación de una tabla con dos columnas: una primera columna donde aparece cada término (debidamente etiquetado según su categoría gramatical) y una segunda columna adyacente que contiene el tweet al que el término pertenece.

En total **63.703 términos** contenidos en **11.453 tweets** han sido extraídos mediante el proceso de creación de la bolsa de palabras.

Row ID	Term	Document
Row1	name[VB(POS)]	"name women inventors girls love science question"
Row2	women[NNS(POS)]	"name women inventors girls love science question"
Row3	inventors[NNS(POS)]	"name women inventors girls love science question"
Row4	girls[NNS(POS)]	"name women inventors girls love science question"
Row5	love[VBP(POS)]	"name women inventors girls love science question"
Row6	science[NN(POS)]	"name women inventors girls love science question"
Row7	question[NN(POS)]	"name women inventors girls love science question"

Tabla 1. Extracto de la tabla contenida en la bolsa de palabras.

Cálculo de frecuencias relativas

En este paso se calcula la frecuencia relativa de cada término. El cálculo de la frecuencia relativa se realiza sobre un total de 73.703 términos contenidos en 11.453 tweets.

El cálculo matemático viene dado por la división de la frecuencia absoluta de un término en base al tweet por el número total términos encontrados en el tweet.

2.3.3. Representación gráfica de la nube de palabras

El último paso en la creación de la nube de palabras consiste en realizar la representación gráfica o visual de los términos.

La representación gráfica de los términos se lleva a cabo en dos etapas: asignación de colores a términos según categoría gramatical (adjetivos, verbos y nombres) y creación / configuración de la nube de palabras.

Asignación de colores según categoría gramatical

La asignación de colores según la categoría gramatical tiene como objetivo identificar de forma visual e inequívoca si los términos son nombres, adjetivos o verbos. Se enriquece de esta forma la información disponible en la nube de palabras. Los términos cuya categoría gramatical es **nombre** tienen asignado el color **rojo**, los términos cuya categoría gramatical es **verbo** tienen asignado el color **azul** y por último los términos cuya categoría gramatical es **adjetivo** tienen asignado el color **amarillo.**

Representación de los términos en la nube de palabras

La representación de términos en una nube de palabras es el último paso a realizar. El nodo **"Tag Cloud"** permite la representación de la nube de palabras a partir de la tabla donde están almacenados los términos y tweets.

En la configuración de dicho nodo hay que indicar cuál es el columna **"Term column"** donde se encuentran los términos y la columna **"Value column"** donde se encuentra la frecuencia relativa de cada término.

Otras de las opciones llamada **"Ignore tags"** permite detectar cuando dos términos son exactamente iguales pero pertenecen a distintos tweets. Esta opción evita la duplicidad de términos en la nube de palabras y el valor numérico asignado a cada término será igual a la suma de todos los términos cuando estos son iguales.

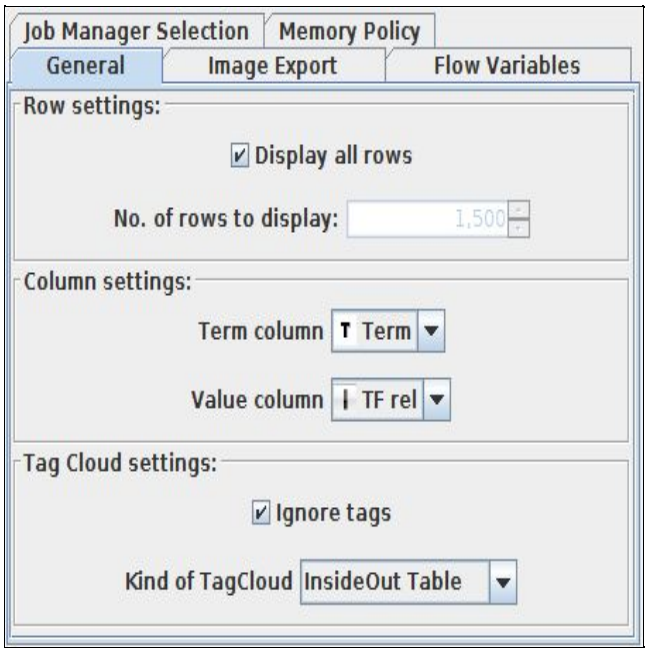

Figura 2. Configuración nodo "Tag Cloud".

De la configuración del nodo **"Tag Cloud"** dependerá el aspecto visual de la nube de palabras.

En el siguiente capítulo se describe y explica el método usado para la clasificación de la polaridad de los tweets.

2.4. Clasificación de la polaridad de Tweets

En este capítulo se clasifica la polaridad del conjunto de tweets extraídos de la cuenta de @Microsoft. Mediante el análisis de los tweets estos van a ser clasificados según su nivel de **polaridad** positivos o negativos. Los resultados presentados serán de tres tipos: positivos, negativos o neutrales en caso de que no se pueda determinar su polaridad. Con la clasificación de la polaridad se busca la extracción de información acerca de la **opinión o comentarios** de los usuarios en relación a los productos y servicios de Microsoft ó noticias relacionadas con la marca. La polaridad de tweets permitirá a los responsable de la empresa obtener información de forma sencilla acerca de sus consumidores / seguidores y la satisfacción de los mismos.

El análisis se divide en tres etapas: preprocesado de tweets, etiquetado de términos y cálculo clasificación de la polaridad.

2.4.1. Preprocesado y filtrado de tweets

Al igual que en la creación de la nube de palabras (*ver sección 2.3.2*) el preprocesado y filtrado de tweets es necesario en la clasificación de la polaridad de los tweets.

En el preprocesado y filtrado de tweets se eliminan todo tipo de links, símbolos o caracteres especiales, hashtags, puntuaciones, números y palabras compuestas por menos de uno o dos caracteres.

Una vez los términos han sido preprocesados y filtrados estos deben ser etiquetados.

2.4.2. Etiquetado de términos

El uso de diccionarios semánticos permitirá la clasificación de los tweets según polaridad. Un **diccionario semántico** se compone de un conjunto de palabras que han sido previamente etiquetadas como positivas, negativas o neutrales. El diccionario en particular que se va a utilizar en este análisis se llama **MPQA**. MPQA es un conjunto de recursos públicos con aplicaciones en el campo de la minería de texto.

En nuestro caso los diccionarios descargados se componen de un conjunto de aproximadamente **8000 palabras** más usadas en Twitter clasificadas según su polaridad.

Originariamente estos diccionarios fueron confeccionados como parte de un proyecto de investigación que se llevó a cabo en el año 2005 en la universidad de Pittsburgh. Los diccionarios han sido actualizados desde entonces con la inclusión de nuevas palabras.

El nodo que permite etiquetar de términos a partir de diccionarios se llama **"Dictionary Tagger"**. Este nodo etiqueta aquellos términos presente en los tweets cuando estos coinciden con alguna de palabras presentes en los diccionario. En este caso el etiquetado de polaridad se realiza a **nivel de término** y no a nivel tweet como es nuestro objetivo (*ver tabla 2)*.

Row ID	Term	Document	...	SENTI...
Row1	can[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row2	you[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row3	name[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row4	any[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row5	women[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row7	asked[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row8	girls[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row9	who[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row10	love[POSITIVE(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	POSITIVE
Row11	science[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row12	this[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL
Row13	question[NEUTRAL(SENTIMENT)]	"can you name any women inventors asked girls who love science this question"	...	NEUTRAL

Tabla 2. Extracto de tabla con términos clasificados según su polaridad.

En la siguiente sección se explica cómo es posible clasificar la polaridad de los tweets tomando como referencia las palabras que han sido etiquetadas como positivas, negativas o neutrales [8].

2.4.3. Cálculo del número de términos por tweet

A lo largo de esta sección se explica cómo se calcula el número de términos positivos, negativos y neutrales existen dentro de un mismo tweet.

Para realizar dicho cálculo se utiliza el nodo **"Pivoting"** que permite agrupar los tweets en filas y pivotar la tabla según la etiqueta de los términos. Como método de agregación de los valores a la tabla se cuenta el número de términos presentes por cada tweet.

La utilización de tablas tipo pivoting tienen gran utilidad cuando se realizan análisis como el llevado a cabo en este capítulo pues permite de forma sencilla la búsqueda de valores únicos dentro de una tabla de grandes dimensiones.

Cuando un término no pudo ser etiquetado en el proceso de etiquetación o este no existe se muestra el símbolo ? .

Row ID	Orig Document	NEUTRAL...	POSITIVE...	NEGATIVE...
Row5395	"Microsoft is getting pretty good at math! (72/73?) @Microsoft @windowsstore @GearsofWar https://t.co/vExy5XgCau"	3	2	?
Row2458	"@Microsoft ya know, I wish I could use an operating system that doesn't force my wifi off 4 times until I update my computer."	13	1	1
Row2406	"@Microsoft when the fuck was it a good idea to automatically turn on game dvr in and update your fucking mongoloids?"	11	1	1
Row6333	"RT @CitrixAMPartner: What the new @Citrix-@Microsoft partnership means for #VDI https://t.co/9Dfzyc1w0k https://t.co/6MoCg...	6	?	?
Row442	".@Microsoft & @versiumsocial bring predictive intelligence to @MSFTDynamics365, by @_JessNelson_ https://t.co/0EqlcRaQPo"	2	1	?
Row5571	"Oh for f**k's sake @Microsoft your forced #windows10 updates boned my system drivers AGAIN! Please stop this shit for the l...	12	2	?
Row4550	"Guess what else you "can't do on a Mac" @microsoft. Sit through this bollocks whilst on a deadline. 😠 #Windows10 #ad http...	9	?	?
Row9719	"RT @SeattleShepard: He's nerdy & he knows it. The ethical challenges of AI as seen by @POTUS, @Microsoft and @Facebook...	5	1	?
Row3506	"@navug @Microsoft need to fix structure of standard code base first - otherwise extensions will not work without recoding in ...	11	1	1
Row13346	"Thank you @Yoobigives @WorldVision @Microsoft for the wonderful supplies! @MidwayMustang are thankful! ♥ https://t.co/b9...	4	3	?
Row920	"@Coppertop004 @Microsoft @MicrosoftStore @microsoftband I will exchange non existent London store 🏬 I can even buy a ...	9	2	?
Row3426	"@microsoft am getting sick of this message. https://t.co/aru7aYrQMK"	3	?	1
Row2958	"@Xbox How do you claim your free Gears of Wars (1,2,3) as promised when I bought my Xbox One S Gears 1TB pack from @Mi...	15	1	?
Row2111	"@Microsoft hey, stop focusing on making shit look nice and make sure it runs smoothly. Is this really that hard of a concept t...	14	2	1
Row14016	"You can grow your business faster? download free eBook w @yourbrandmrktng, @Microsoft @CarolJSRoth & @PointA_PointB ...	8	?	?

Tabla 3. Muestra de la tabla pivotada donde se muestra el número de términos etiquetados por cada tweet.

Una vez se sabe el número de términos etiquetados por cada tweet se procede al cálculo de la polaridad a nivel de Tweet.

2.4.4. Cálculo y clasificación de la polaridad a nivel de tweet

Sabiendo cuántos términos neutrales, positivos o negativos contiene cada tweet se calcula si el tweet en su conjunto puede ser clasificado como positivo, negativo o neutral. El trabajo realizado por J. Ramos [9] servirá en nuestro caso como guía. En dicho trabajo se explica cómo es posible determinar y calcular mediante la metodología **term frequency–inverse document frequency (TF-IDF)** el peso que un término tiene dentro de un determinado documento. Matemáticamente sería:

$$TF * IDF_w = f_{w,d} * log\frac{D}{f_{w,d}} \quad (1)$$

Donde en $f_{w,d}$, equivale al número de veces que el término *w* aparece en un documento *d,* entiéndase por documento un tweet. *D* es el el tamaño del corpus o conjunto de datos.

Para poder calcular y clasificar la polaridad de cada tweet primero se asigna un valor numérico a cada término según estos fueron etiquetados en categorías gramaticales.

Los términos etiquetados como **negativos** toman el valor numérico **-1**, los términos **neutrales** toman el valor **0** y, por último, los términos **positivos** toman el valor **+1**.

Una vez los términos han sido asignados con un valor numérico se aplica el método *TF-IDF* en combinación con el valor numérico asignado a cada término.

Hay que calcular la puntuación total de un término teniendo en cuenta su peso dentro de un Tweet (*ver fórmula 1*).

Matemáticamente:

$$Total\ score_w = TF * IDF_w * S_{w,d} \quad (2)$$

Dónde $S_{w,d}$, es la puntuación de un término *w* que pertenece al documento - tweet *d.*

Knime dispone de dos nodos llamados **"TF"** y **"IDF"** respectivamente que facilita el cálculo $TF * IDF_w$

Row ID	Term	Orig Document	SENTIMENT	SENTI...	TF rel	IDF	TFIDF	Total score
Row6061	the[NEUTRAL(SENTIMENT)]	""##Intelligence is the new normal" #SaaS #Cloud @Link...	NEUTRAL	0	0.25	0.651	0.163	0
Row6062	new[NEUTRAL(SENTIMENT)]	""##Intelligence is the new normal" #SaaS #Cloud @Link...	NEUTRAL	0	0.25	1.338	0.335	0
Row6064	---[NEUTRAL(SENTIMENT)]	""##Intelligence is the new normal" #SaaS #Cloud @Link...	NEUTRAL	0	0.25	3.756	0.939	0
Row6063	normal[POSITIVE(SENTIMENT)]	""##Intelligence is the new normal" #SaaS #Cloud @Link...	POSITIVE	1	0.25	3.154	0.789	0.789

Tabla 4: Muestra de la tabla resultante con el cálculo de la fórmula 2 dada por la columna Total $score_w$.

Para calcular el valor total (**Total $Score_d$**) a **nivel de tweets** estos son agrupados en una tabla utilizando como método de agregación la suma de $Total\ Score_w$ tantas veces como términos se encuentran presente en un documento ó tweet.

$$Total\ Score_w = \sum_{n=0}^{f_{w,d}} TF * IDF_w * S_{w,d} \quad (3)$$

La clasificación de la polaridad de cada tweet se realiza en base al siguiente criterio:

- Total $score_d$ = 0, tweet clasificado como neutral
- Total $score_d$ > 0, tweet clasificado como positivo
- Total $score_d$ < 0, tweet clasificado como negativo

En la parte de resultados se muestra una extracción de la tabla con la clasificación de la polaridad a nivel de tweets.

2.5. Detección de influenciadores en Twitter

En este capítulo del trabajo se construye y analiza la red de usuarios existente entorno a la cuenta de @Microsoft. Para llevar a cabo esta parte, el trabajo se basa en lo que se denomina **Social Network Analysis (SNA)** y en concreto se utiliza una simplificación de la metodología denominada **weighted networks** [10].

SNA es el proceso de investigación de estructuras sociales mediante el uso de redes y teorías de gráficos [11]. Se caracteriza por estructuras de redes compuestas de **nodos - nodes -** (individuos, usuarios, gente, etc) y **aristas - edges -** que conectan los nodos según la relación existente entre estos. SNA tiene aplicaciones muy variadas entre las que se encuentran el análisis de redes sociales, relaciones sentimentales entre individuos, gráficos de colaboración, *etcétera* [12].

La metodología **weighted networks** asigna pesos a las aristas que conectan los nodos y por tanto se mide la fuerza de dicha conexión.

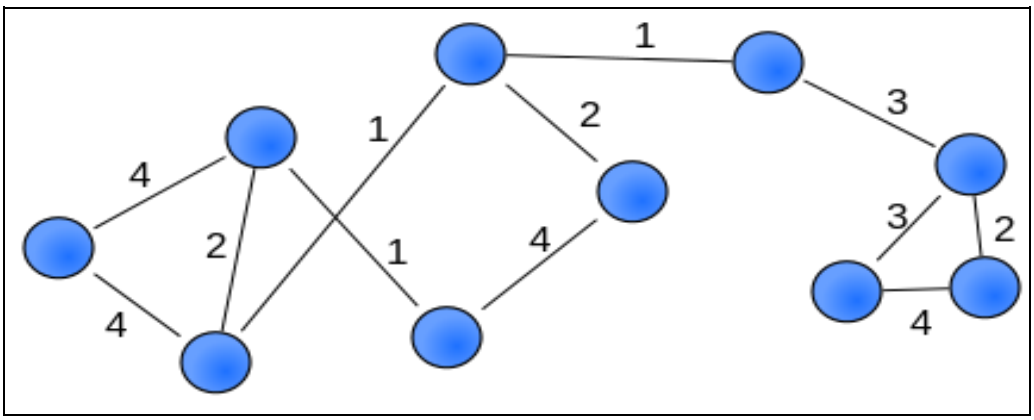

Figura 3. Ejemplo ilustrativo weighted network.

En este trabajo se hace uso de las herramientas de análisis de redes - Network Analysis - disponibles en la librerías de KNIME. En la red que se construye los **nodos - nodes -** representan **usuarios** y las **aristas - edges -** conectan los nodos entre sí utilizando como criterio matemático el número de veces que los usuarios realizan **retweets** entre ellos en un

determinado periodo de tiempo. La red que se construye se hace en torno a los **hashtags** más populares utilizados por los usuarios. De esta forma se pone énfasis en estudiar aquellos eventos que han sido etiquetados con la utilización de hashtags.

También, y con el objetivo de añadir más información al trabajo, se presenta una **tabla con información a nivel usuario.** La tabla incluye aquellos **20 usuarios** que más tweets tienen publicados en sus cuentas en el momento de la extracción de los datos.

Por último, una **nube de palabras** que contienen los **hashtags** más utilizados por los usuarios es generada con el objetivo de identificar aquellos **temas o discusiones** que rodean al conjunto de tweets extraídos y la red que se construye en torno a los mismos.

La primera sección de este capítulo describe la base de datos con las que se trabaja pues esta difiere un poco de la anteriormente usada.

2.5.1. Dataset utilizado

A modo de recordatorio para el lector, el dataset utilizado en este capítulo difiere del utilizado en los análisis anteriores *(ver capítulo 2.1)*.

Los siguiente análisis serán descritos y explicados en las correspondientes secciones:

- Nube de palabras a partir de hashtags
- Tabla estadística a nivel de usuario
- Construcción de red de influyentes

Los resultados de los análisis se muestran en la parte de resultados dentro de este trabajo.

2.5.2. Nube de palabras de hashtags

Para generar una nube de palabras que contengan los hashtags encontrados en el dataset se sigue un proceso similar al utilizado anteriormente *(ver capítulo 2.3)*. El proceso se divide en dos partes: una primera parte donde se extraen los hashtags y una segunda parte donde se calcula la frecuencia relativa de cada hashtags respecto al total, lo que servirá para representar los hashtags en la nube de palabras.

Extracción de hashtags

Para la extracción de hashtags se utiliza una expresión regular tipo regex a través del nodo **"Wildcard tagger".** Dicho nodo permite recorrer y etiquetar toda la base de datos buscando aquellos términos que coinciden con la **expresión regular**. Después únicamente queda filtrar aquellos términos que han sido etiquetados como hashtags.

Cálculo de frecuencias y representación de la nube de palabras

Una vez han sido identificados los hashtags se calcula la **frecuencia relativa** de cada hashtag respecto al conjunto total. Solo aquellos hashtags que cuentan con un mínimo de **20 ocurrencias** en total son utilizados en la representación de la nube de palabras.

2.5.3. Tabla con información a nivel de usuario

La extracción de una tabla con datos **nivel de usuario** permite conocer mejor a aquellos usuarios que pertenecen a la base de datos utilizada en este trabajo.

Debido al gran número de usuarios, **258**, en la tabla se mostrarán sólo aquellos **20 primeros usuarios** que más tweets tienen publicados en sus cuentas. Los tweets publicados incluyen tanto tweets propios como retweets.

Para realizar dicha operación se utiliza el nodo **"Twitter user"** el cual permite la extracción de información a nivel de usuarios tal como el número de followers, número de usuarios a los que siguen, número de tweets marcados como favoritos, foto de perfil, entre otros.

2.5.4. Construcción de red de influyentes

La construcción y análisis de la red social (SNA) es un proceso que debe hacerse con cuidado para realizar los cálculos necesarios. En el caso de este trabajo se sigue la metodología llamada **weighted network**. El análisis se realiza a nivel de **retweets**, es decir, se construye la red de forma que los usuarios que están conectados entre sí son, por una parte, los usuarios que publican tweets originales (cuenta origen) y, por otra parte, los usuarios que realizan retweets, es decir, realizan una republicación de un tweet (proveniente de la cuenta origen). Recordemos que la red que se construye sólo incluye los **hashtags** más populares, es decir,

aquellos que se encuentran representados en la nube de palabras de hashtags. Así se centra la atención en los eventos más llamativos, o al menos aquellos que más actividad han generado.

Esta sección está dividida en **tres partes** diferentes: una **primera parte** en la cual se **identifican y se conectan** aquellos los usuarios que han realizado un retweet con la cuenta del tweet donde originariamente fué publicado. Por ejemplo: un usuario *x* hace un retweet del tweet "New #Azure update available from today at Microsoft Play Store" que fue originariamente publicado por el usuario *y,* el cúal es buen conocedor de las tecnologías. Se trata entonces de poner en relación el usuario *x* e *y*.

En la **segunda parte** se identifica aquellos usuarios que más retweets realizan. Una vez identificados los usuarios que más retweets realizan la tabla resultante es filtrada de modo que sólo se mantiene a los 50 primeros usuarios que más retweets generan. Mediante el filtrado de **tweeters**[7] se consigue que la red no quede sobredimensionada y la interpretación de los resultados resulte más fácil y sencilla.

Una vez las dos primeras partes se han llevado a cabo, en la **tercera parte** se construye la red a partir de las tablas que contienen la información de las dos partes anteriores. Dicho proceso se realiza mediante la utilización de la librería para el análisis de redes disponible en KNIME.

La metodología a seguir es la siguiente:

Parte primera: Identificar a los usuarios y sus conexiones

En el dataset con el que se trabaja existe la columna "**Retweet from**" en la cual se identifica la cuenta original o procedencia de un tweet, si este había sido republicado en forma de retweet por otro usuario. Previamente los datos han sido filtrados de forma que sólo se trabaja con usuarios que han realizado algún retweet.

Con la utilización de la metodología weighted network la estructura de la red con *N* nodos es representada por la matriz binaria NxN $A=\{a_{ij}\}$, conocida como matriz adyacente, cuyo elementos a_{ij} equivalen a 1, cuando existe una conexión o relación entre el nodo *i* y el nodo *j*.

[7] Persona que publica en la red social Twitter.

En la tabla con los datos con los que se trabaja los elementos *i* son aquellos elementos incluido en la columna "User", en cambio, los elementos *j* son aquellos elementos incluidos en la columna "Retweet from".

El grado k_i de un nodo *i* es definido como el número de su vecindad o el número de conexiones incidentes al nodo *i*:

$$k_i = \sum_{j \in \Pi(i)} a_{ij} \tag{4}$$

Donde a_{ij} son los elementos adyacente a la matriz A y $\Pi(i)$ la vecindad del nodo *i*.

El **peso - weight -** de las aristas entre nodos son descritos como una matriz NxN $W=\{w_{ij}\}$.

El peso w_{ij} es 0 si los nodos *i* y *j* no están conectados entre sí. En nuestro caso consideramos el caso de **pesos simétricos positivos** ($w_{ij} = w_{ij} \geqslant 0$).

La fuerza s_i de las conexiones existente entre los nodos viene representada por:

$$s_i = \sum_{j \in \Pi(i)} w_{ij} * a_{ij} \tag{5}$$

La fuerza s_i tiene en cuenta tanto la conectividad a_{ij} como el peso de estas conexiones w_{ij}

La fuerza s_i es calculada **agrupando** en una tabla los usuarios - User - y los usuarios que publican retweets - Retweet from -. El método de **agregación** utilizado se realiza en base a contar el número de veces que un mismo usuario ha realizado un retweet desde una única cuenta donde el Tweet fue originalmente publicado entre el periodo en cual fueron extraídos los tweets.

Row ID	User	Retweet from	Count(Time)
Row26	HokstadConsult	SiliconArmada	6
Row98	jeremy_chauvet	SA_infra	6
Row52	SAPonTheCloud	InsideSAP	4
Row100	jeremy_chauvet	SiliconArmada	4
Row47	Raimon7	AkuaroWorld	3
Row90	jeremy_chauvet	Dreamcareers_TJ	3
Row96	jeremy_chauvet	MamboLook	3
Row3	Adriana_Rday	LichtensMichael	2
Row6	Andrei_Rday	LichtensMichael	2
Row11	CphSW	CphSW	2
Row17	EdwardTufte	EconAndrew	2
Row20	EraPasi	bobehayes	2
Row25	HerrDoktorFunk	EconAndrew	2
Row29	JuliannaLMD	EconAndrew	2
Row32	MatteoDavanzo1	Sciarring	2
Row33	McAldowney	EconAndrew	2
Row34	MrDinoSossi	EconAndrew	2
Row39	NickolayV11	bobehayes	2

Tabla 5. Muestra de la tabla donde se encuentra los elementos necesarios para la construcción de la red.

En la tabla 5 la columna "Count(time)" representa la fuerza s_i, la columna "User" son los elementos *i,* la columna "Retweer from" son los elementos *j*.

La interpretación de la tabla se hace de la siguiente forma: En la fila 26 el usuario "HokstadConsult" ha realizado hasta 6 retweets de tweets publicados de la cuenta "SiliconArmada".

Parte segunda: Identificación de usuarios que más retweets populares publican

En esta parte se agrupa la tabla a nivel de usuarios utilizando como método de agregación la suma del número de veces en total que un retweet ha sido publicado. Este método de agrupación y agregación de datos tiene como objetivo identificar aquellos tweets que son más **populares**, es decir, que han sido publicados por otros usuarios (retweeted) un mayor número de veces en total.

La tabla 5, que contiene los datos de la red, es filtrada tomando como **referencia** la tabla 6 que contiene el top 50 tweeters. De esta forma se consigue reducir la dimensión de la red pero se mantiene a aquellos usuarios - users- más populares o con más actividad dentro de la red.

Al quedar reducida la dimensión de la red la interpretación de la misma se convierte en un proceso más sencillo.

Row ID	S User	I Sum(Retweeted)
Row89	NickolayV11	18
Row83	NathanDSweeney	19
Row79	MrDinoSossi	51
Row78	McAldowney	51
Row66	JuliannaLMD	51
Row54	HerrDoktorFunk	51
Row44	EraPasi	18
Row40	EdwardTufte	51
Row39	EconAndrew	51
Row256	yaroslav_f	51
Row255	xqa	51
Row249	topiclybigdata	51
Row243	themeistro	51

Tabla 6. Extracto de la tabla que contiene el top 50 tweeters.

Después de realizar los pasos anteriores ya se puede construir de forma visual la red de influyentes.

Parte tercera: Construcción de la red influenciadores

Es necesario describir brevemente los nodos que se utilizan en la construcción de la red dada su importancia y función.

Los nodos utilizados son: **"Object inserter"** y **"Network viewer"**.

El nodo "Object inserter" permite la creación de una red a partir de las tablas anteriores. Este nodo permite la creación de una red a partir de una tabla donde cada columna representa los nodos (nodes) y las filas las aristas (Edges). Un ejemplo sería tal como sigue en la siguiente tabla.

Node1 ID	Node2 ID
node1	node2
node2	node4
node3	node4

Tabla 7: Tabla ejemplo nodo "Object Inserter".

De una correcta configuración de este nodo dependerá que la red de influyente se muestre correctamente.

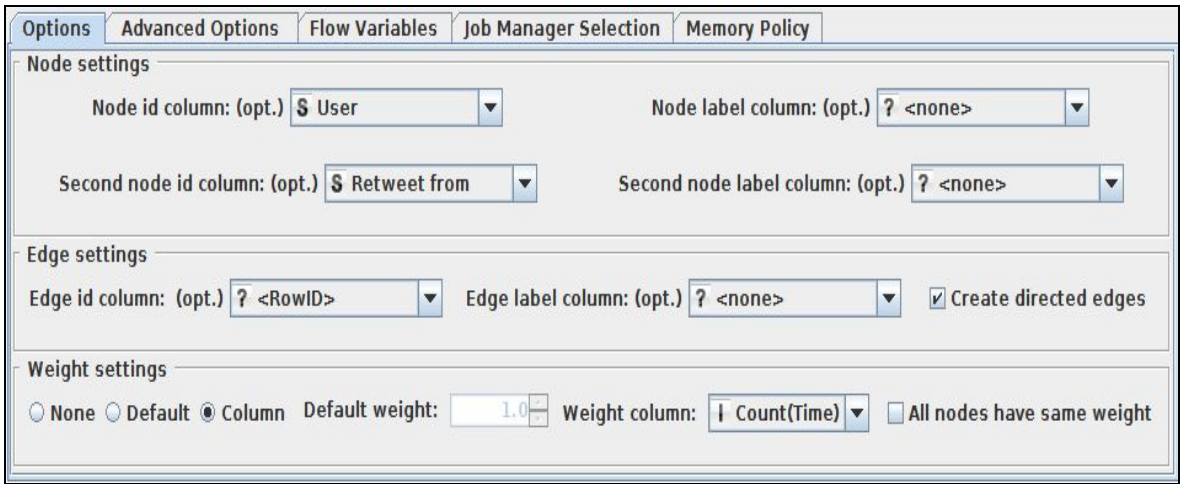

Figura 4. Configuración final del nodo "Object Inserter".

La red final contiene en total **67 nodos** y **65 aristas**. Como se mostrará en la parte de resultados, existen usuarios dentro de la red que no necesariamente están directamente conectados entre sí pero publican retweets de un mismo usuario que funciona como enlace de los usuarios anteriores.

Por otra parte el nodo **"Network viewer"** permite visualizar la red construida anteriormente mediante la utilización. Entre las opciones de configuración del nodo se encuentran opciones de aspectos visuales y resolución de la imagen.

Una vez la metodología seguida en los distintos análisis ha sido convenientemente explicada, en la siguiente parte de este trabajo se muestra los resultados y la interpretación que se le puede dar a los mismos.

Parte III: Resultados e interpretaciones

En esta parte del trabajo se muestran los resultados obtenidos de los análisis anteriores así como las interpretaciones que se le puede dar a los mismos. Primero se muestra la nube de palabras generada, segundo se muestra la polaridad del conjunto de tweets analizados y por último se muestra la red de influenciadores.

3.1. Nube de palabras

Un total de **11.443 tweets** fueron extraídos de la cuenta de @Microsoft o menciones a la misma entre el 1 de Septiembre 2016 y 10 de Octubre de 2016. Dichos tweets extraídos fueron utilizados para la generación de la nube de palabras (*ver capítulo 2.3 para la explicación metodológica y técnica*).

La representación gráfica de aquellos términos que más se repiten en el conjunto de tweets analizados permite a la persona que lo visualiza rápidamente captar de forma genérica aquellas palabras o temas (topics) que han sido más comentado por la comunidad de Twitter en torno a la cuenta de @Microsoft o menciones a la misma.

En la nube de palabras los **colores** representan la **categoría gramatical** de cada palabra, así el color rojo representa los nombres, el color azul los verbos y el color amarillo a los adjetivos.

El tamaño de la fuente de las palabras representa el peso que cada palabra tiene en relación al total de palabras encontradas en todo el conjunto de tweets.

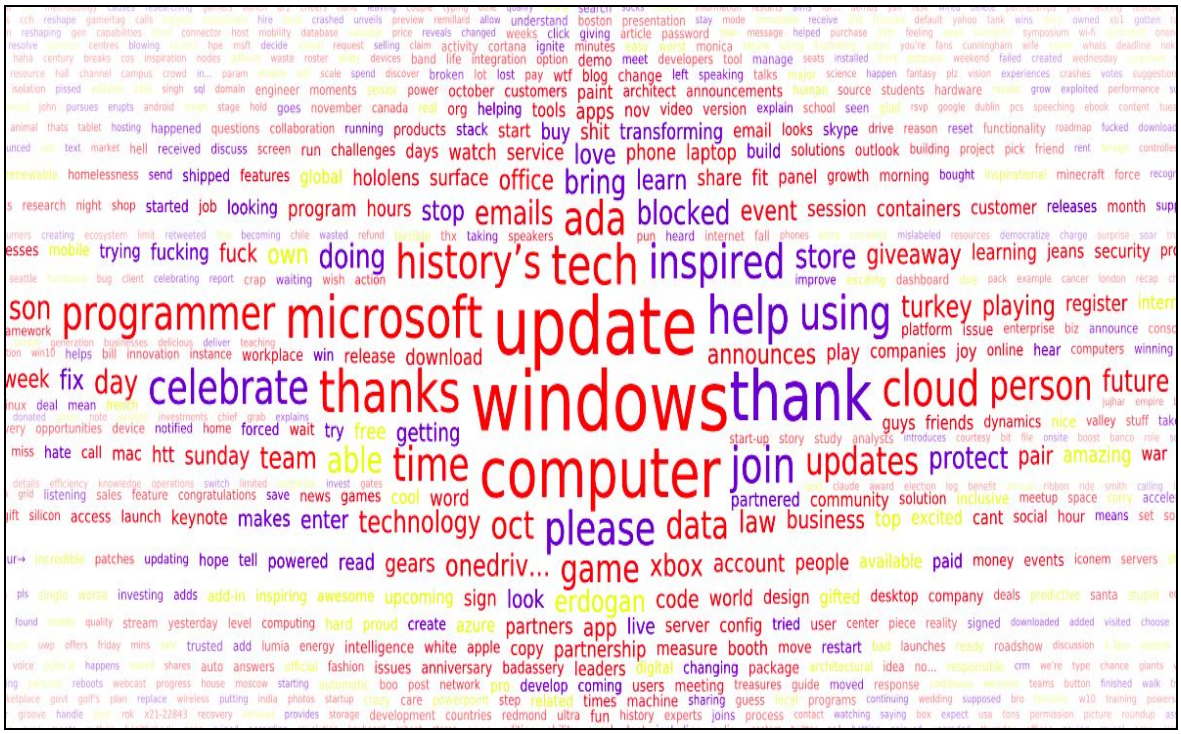

Figura 5. Nube de palabras correspondiente a la cuenta @Microsoft.

Las interpretación que se realice de la nube de palabras siempre estará sujeta al criterio de la que persona que realiza dicho análisis pues dependerá de dónde centre su atención y que partido quiera obtener del análisis. En este caso podemos centrar la atención en analizar 4 palabras que pueden ser consideradas más destacables. Las palabras en concreto son:

- **Windows:** Esta palabra se ha mencionado en muchas ocasiones a consecuencia del lanzamiento de un paquete de actualizaciones y mejoras para Windows 10 en el periodo en el cual los tweets fueron extraídos.

- **Updates:** Algunas personas o usuarios de Twitter han tenido algunos problemas a la hora de actualizar su sistema operativo Windows 10. Leyendo algunos de los tweets se puede puede concluir que los problemas o quejas en concreto han sido: actualizaciones forzosas, horas inapropiadas para realizar actualizaciones de forma automáticamente y ordenadores que se quedan bloqueados.

- **Thanks:** Un evento, el cual Microsoft ha sido unos de los patrocinadores, se ha llevado a cabo para celebrar el Día del programador el 12 de Septiembre. Se han

realizados numeroso sorteos entre la comunidad de Microsoft o sus followers donde los participantes agradecen el entrar en el sorteo.

- **Turkey:** Comentarios o noticias relacionados la filtración de emails confidenciales del gobierno de Turquía en verano. Como consecuencia y medidas de precaución muchos servicios en Internet fueron bloqueados cautelarmente en el mes Octubre. Entre dichos servicios bloqueados se encuentran el servicio de mensajería Outlook y One Drive de la compañía Microsoft.

3.2. Polaridad del conjunto de tweets

Del total de 11.409 tweets analizados **4.896 tweets** han sido clasificados como **positivos**, **4.001**clasificados como **neutrales** y **2.511** clasificados como **negativos** (*ver capítulo 2.4 para la explicación metodológica y técnica*)

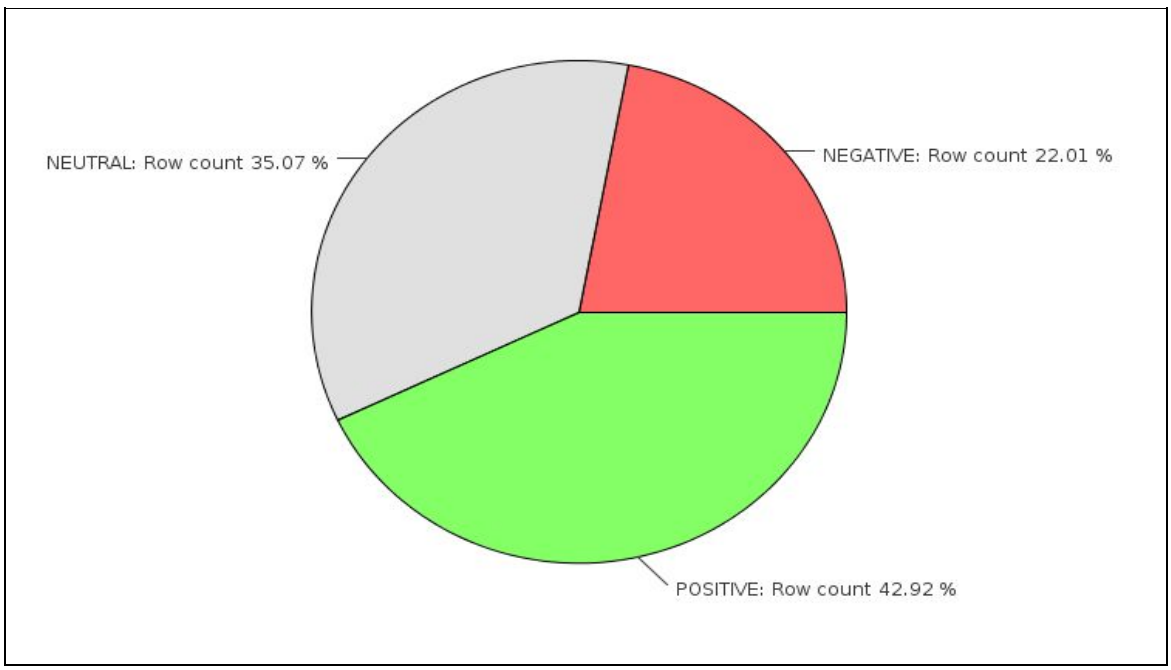

Figura 6. Gráfico circular con la proporción de tweets clasificados según polaridad.

La mayoría de los usuarios realizan comentarios positivos acerca de los productos o servicios de Microsoft o al menos cuando la marca aparece mencionada en algún tweet utilizando para ello @Microsoft.

La tabla con los resultado muestra los tweets originales y su clasificación según la polaridad.

Row ID	Tweet	Total score	Polaridad
Row4136	"Dear @Microsoft , is your @MicrosoftHelps account so badly staffed our did they decide not to bother with my problem any more?"	-0.452	NEGATIVE
Row627	".@Microsoft, even though I've completely disabled your update auto-restart it still restarts, making me lose my work. You're despicable!"	-0.584	NEGATIVE
Row8738	"RT @_CaRtOoNz: Want me to giveaway 2 Recore codes? Fine. Retweet and I'll pick some winners in a bit! Thanks @Microsoft 😁"	0.195	POSITIVE
Row4992	"If @Microsoft ever bought @Twitter, in addition to it acquiring LinkedIn, boy...would that be some monster."	-0.358	NEGATIVE
Row6653	"RT @H2ODelirious: Got this awesome cool thing today, Thanks @Microsoft Gears of war!!!!!!!! https://t.co/7NtLrCV7BE"	0.012	POSITIVE
Row6487	"RT @DocLogix: DocLogix @DirectionsEMEA started the day 2! Keynote by Paul White @Microsoft Market trends and threats - get ready for th...	0.099	POSITIVE
Row5501	"Nothing makes me want to try Linux more than ads in @microsoft @windows 10 for @google pixel"	-0.057	NEGATIVE
Row2576	"@NickvVuren @DarrylvdPeijl @Microsoft @MSDN uhhh weird. I just tweeted the same dialog with correct list a few days ago. Have url to iso?"	0.066	POSITIVE
Row10637	"That's why lol ur companies compete...@microsoft can't even monopolize one f***ong market...pathetic @BillGates"	-0.109	NEGATIVE
Row3427	"@microsoft won't last these cockroaches compete at everything they fight on a level playing field on everything"	-0.262	NEGATIVE

Tabla 8: Extracto de la tabla con tweets clasificados según polaridad.

En un análisis más profundo de los resultados cabe la posibilidad de la creación de una nube de palabras de aquellos tweets considerados positivos, realizar un seguimiento de los links incluidos en los tweets, *etcétera*.

En general se denota del análisis que los usuarios en twitter expresan un sentimiento positivo hacia la marca Microsoft cuando estos publican tweets, retweets o respuestas a tweets que la misma marca publica.

La información mostrada en la tabla final puede resultar de utilidad para el departamento de Marketing y atención al cliente de Microsoft pues con un análisis de la polaridad se consigue detectar cual es el contenido de los tweets cuando estos tiene una connotación positivos y por otra parte cuál es el contenido de los mismo cuando estos tienen una connotación negativa.

3.3. Red de usuarios influyentes

Los resultados presentados en este capítulo incluyen resultados de la nube de palabras de hashtags, tabla con información a nivel de usuario y finalmente la red de influenciadores construida (*ver capítulo 2.5 para la explicación metodológica y técnica*).

Nube de palabras de hashtags

La nube de palabras contiene en total de **15 hashtags** distintos donde cada uno de los hashtags tiene una ocurrencia de 20 repeticiones identificadas dentro de la base de datos

utilizada. Los hashtags son tomados como referencia a la hora de construir la de red de usuarios influyentes.

Figura 7. Nube de palabras generada a partir de los hashtags más populares.

El tamaño de fuente de las palabras determina la frecuencia relativa de cada una de ellas en relación al conjunto total de hashtags analizados.

De los resultados de esta nube de palabras cabe destacar cómo algunos de los productos de Microsoft tales como Azure, Xbox, Windows 10 y Surface son consideramos como tema de discusión entre los usuarios de Twitter. También se observa como aquellos temas relacionados con el Big data y diferentes aplicaciones se encuentran entre los hashtags más populares.

Sin duda una de las características que cabría destacar es la presencia que tiene la marca Microsoft en el campo del big data y sus aplicaciones.

Tabla con información a nivel de usuario

En la tabla con información a nivel usuario se muestran los **20 primeros usuarios** que más tweets se encuentran publicados en sus cuentas. Al ser tweets publicados en la cuenta de un determinado usuario se incluyen tanto tweets propios como retweets.

La columna “tweet” en la tabla 9 sólo recoge aquellos tweets y retweets diferentes unos de otros en la base de datos y que estos pueden ser atribuidos a la cuenta de un usuario mediante el recuento de los mismos.

El resto de columnas de la tabla recoge cifras totales con datos extraídos a nivel de cuenta de usuarios.

Row ID	User (0)	Count(Tweet) (0)	User - Followers (0)	User - Friends (0)	User - Favourites (0)
Row0	jeremy_chauvet	36	2208	197	646
Row1	SA_infra	20	555	272	67
Row2	SiliconArmada	20	35166	34599	3473
Row3	Dreamcareers_TJ	11	422	776	402
Row4	TechNewsbit	9	209	241	0
Row5	El_Ross_M	8	305	922	90
Row6	Hestisdabest	8	1529	2712	1796
Row7	SoftwareTopNews	8	105	67	0
Row8	_VRr0Om_	8	494	215	118
Row9	edgfuentes13	8	309	400	25690
Row10	roomeezon	8	6647	104	285789
Row11	HokstadConsult	7	2608	1429	40
Row12	MamboLook	7	291	672	1
Row13	AkuaroWorld	6	771	3702	1117
Row14	Zyaldar	6	1508	1487	43829
Row15	bobehayes	6	13853	6814	15189
Row16	Atrion	4	1574	1326	326
Row17	CherryRasulka	4	676	2205	2597
Row18	DeadlightRising	4	18	69	60
Row19	Follower419	4	25	133	93

Tabla 9. Tabla con el top 20 usuarios con más tweets publicados dentro del conjunto de datos.

En la tabla se observa que el usuario ***@jeremy_chauvet*** es sin duda al usuario que más tweets y retweets publicados han sido contados dentro del dataset con un total de **36 tweets** no repetidos. Este usuario según su cuenta de Twitter es profesional en el campo del Big Data y análisis de datos.

Sin embargo esto lo anterior no significa que un mayor número de tweets atribuidos con un usuario equivale a ser más activo en la red Twitter pues sin embargo el usuario ***@roomeezon*** cuenta con un total de 285.789 tweets marcados como favorito[8] lo que significa que este usuario es muy activo aunque su tasa de publicación sea menor que la de otros usuarios dentro de la red . Este usuario es un profesional del campo de las tecnologías según se desprende de su perfil de Twitter.

Por otro lado tenemos el usuario ***@SiliconArmada*** que es el usuario que cuenta con un mayor número de followers. Se trata de una cuenta corporativa de una empresa de reclutamiento en el mundo de las tecnologías.

[8] Los tweets marcados como favoritos son aquellos marcados con una estrella. De esta forma se hace saber al autor del tweet que a alguien le gusta ese tweet. Esta opción no requiere necesariamente publicar el Tweet por parte de otros usuarios (retweet).

Red de influenciadores

La red de influenciadores ha sido construida tomando como punto de referencia los **15 hashtags** más populares (*ver sección 2.5.2*) y los **50 usuarios** que más retweets populares han publicados en sus cuentas (*ver sección 2.5.4*). En dicha red se observa cómo los usuarios están conectados entre sí cuando estos han publicado algún retweets en sus cuentas. La serie de eventos cuya temática está relacionada con el big data y técnica de análisis de datos (*ver figura 7*) ha generado en la red social Twitter una serie de intercambio de información entre usuarios. Los tweets que se han analizado tienen, además, la peculiaridad que todos ellos realizan alguna mención a la cuenta oficial de Microsoft en Twitter @Microsoft.

En la red (*ver figura 8*) los **nodos** en de color rojo representa cada uno de los **usuarios** que han sido analizados.

Las **aristas** representan las **conexiones** entre sí, donde tanto el grosor de las mismas como los números representan la fuerza de los pesos de las aristas S_i (*Ver fórmula 2, sección 2.5.4*). La dirección de las aristas indica la relación que existe entre dos usuarios.

Las aristas **entrantes** a un nodo indica que ese nodo es la **fuente de procedencia** de retweets publicados por el resto de usuario en sus cuentas.

En cambio las aristas **salientes** de un nodo indica que ese usuario ha realizado al menos un **retweet** procedente del nodo al que este conecta.

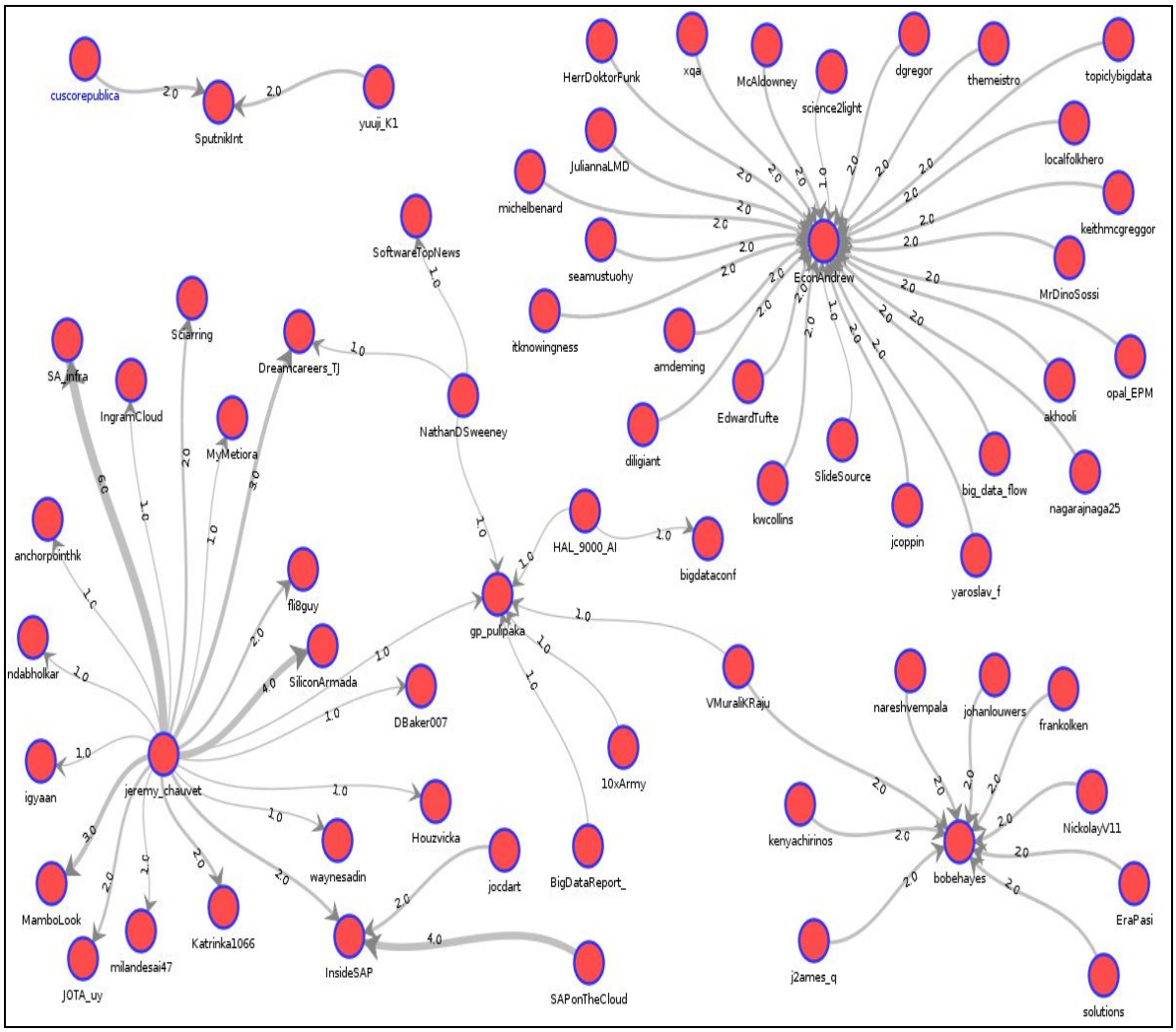

Figura 8. Red de influenciadores construida a partir de los tweets extraídos.

El usuario **@EconAndrew** es un usuario que puede ser clasificado como el que más **poder de influencia** tiene dentro de esta red puesto que unos **25 usuarios** han realizado una media de 2 retweets de tweets publicados originalmente por @EconAndrew. Este usuario publica tweets que son republicados por sus seguidores pero raramente el usuario @EconAndrew realiza un retweets de otros usuarios.

En el lado opuesto al usuario @EconAndrew tenemos al usuario **@jeremy_chauvet** el cual es usuario muy activo puesto que ha realizado una gran cantidad de retweets de cerca de **19 usuarios** distintos.

De la red se observa además como existen al menos tres grupos de nodos que aparentemente no tienen conexión entre sí. Esto puede deberse a que la temática de los tweets no sea

exactamente la misma en todos ellos y de ahí la utilización de diferentes hashtags. Aunque bien si es cierto que todos estos usuarios tienen en común que en todos se realizan menciones al usuario @Microsoft.

Para completar la interpretación de los resultados se puede extraer información a nivel de usuarios de aquellos que se encuentran presentes en esta red y que son aquellos que más han influido en el resto.

Row ID	User (0)	Description (0)	Location (0)	Statuses...	Followers...	Friends...	Favourites...
Row0	jeremy_chauvet	Étudiant en M2 @ITESCIA / Alternant @CapgeminiLabs #Nantes. Intéressé pa...	Cergy / Nantes	126939	2844	198	651
Row1	EconAndrew	Data Scientist, Development Data Group, @WorldBank. Previously at @nesta...	Washington, DC	2585	760	623	520
Row2	gp_pulipaka	Ganapathi Pulipaka | Founder and CEO @deepsingularity | Bestselling Author...	Los Angeles, CA	17491	12115	8463	113
Row3	bobehayes	B.O.B. is Chief Research Officer @AppuriCorp. PhD in industrial-organizational ...	Seattle, WA	40388	15051	7750	19884
Row4	NathanDSweeney	Technical Account Manager (TAM) @ cogecopeer1	Portsmouth, UK	2581	148	104	1
Row5	InsideSAP	Online magazine for #SAP professionals in Australia and New Zealand. Subsc...	Sydney, Australia	3874	3338	2535	165
Row6	SputnikInt	Sputnik is a global wire, radio and digital news service. We exist to tell the st...		156997	153512	266	11

Tabla 10. Tabla con información de los usuarios más influyentes en la red.

En la tabla 8 se observa cómo el usuario **@EconAndrew** es de los que con menos cuenta follower pero si embargo es es aquel que más actividad generada ha influido al resto de usuarios en la red. Hay que tener en cuenta que esta red analiza a una serie de eventos que fueron utilizados con la utilización de hashtags. Por tanto el usuario @EconAndrew, quien según su propio perfil, es **científico de datos** pudo actuar como altavoz de los eventos acontecidos que están relacionados con Big Data y análisis de datos principalmente.

El usuario @SputnikInt en realidad es una cuenta corporativa que pertenece a un **medio de comunicación** de noticias relacionadas en el mundo digital. Esta cuenta es la que cuenta con mayor número de publicaciones y followers.

El usuario @babehayes tiene también bastante influencia dentro de la red. Según la descripción de su perfil sus **intereses** están relacionados con el análisis de datos, big data y machine learning. A su vez este usuario es propietario de una página web que está especializada en big data y análisis de datos.

Del análisis anterior se denota como las abundantes menciones a la marca @Microsoft muestra que esta está bastante involucrada en el desarrollo de las nuevas tecnología o desarrollo de técnicas de análisis de datos pues los usuarios anteriormente analizados tienen

en común que sus intereses en el Big Data, análisis de datos y desarrollos de tecnología con aplicaciones en el mismo.

Parte IV: Conclusiones

Con el desarrollo y presentación de los resultados en el presente trabajo se pone en énfasis la importancia que puede tener para organizaciones y empresas el análisis de redes sociales en combinación con la minería de texto.

Se ha demostrado cómo es posible extraer una serie de tweets relacionados con la cuenta de @Microsoft en un periodo determinado para realizar un análisis de los mismo con posterioridad. El resultado de dicho análisis y conclusiones del mismo puede ser usado por los directivos de la empresa o responsable de departamento de Marketing para desarrollar su comunicación con los usuarios y consumidores.

Como resumen se puede extraer las siguientes conclusiones:

- Cerca del 42% de los tweets analizados tienen una connotación positiva
- Los usuario identificados en la red de influenciadores tienen en común que o bien son profesionales en el campo del big data, análisis de datos & desarrollo de técnicas o bien muestran fuertes intereses en estos campos
- La marca de Microsoft genera un gran volumen de tráfico incluso en periodos relativamente cortos en el que fueron extraídos los datos. No es para menos lo anterior pues es una cuenta con cerca 8 millones de seguidores y una marca líder en el desarrollo de las tecnologías

Referencias

[1] Bo Pang and Lillian Lee (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 92.

[2] http://knime.org

[3] Pang, Lee and Vaithyanathan, 2002; Turney, 2002.

[4] James Sinclair and Michael Cardew-Hall, The folksonomy tag cloud: when is it useful?; Journal of Information Science 2008.

[5] https://moz.com/blog/how-to-improve-your-rankings-with-semantic-keyword-research

[6] Avnit, A. 2009. The Million Followers Fallacy, Internet Draft, Pravda Media. http://tinyurl.com/nshcjg.

[7] Kristina Toutanova and Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[8] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, University of Pittsburgh, Pennsylvania, 20.

[9] J. Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries”, Rutgers University, New Jersey, 2012.

[10] M. E. J. Newman (2004), Analysis of weighted networks, Department of Physics and Center for the Study of Complex Systems, University of Michigan.

[11] Otte, Evelien; Rousseau, Ronald (2002). "Social network analysis: a powerful strategy, also for the information sciences". Journal of Information Science.

[12] Grandjean, Martin (2016). "A social network analysis of Twitter: Mapping the digital humanities community". Cogent Arts & Humanities. 3 (1): 1171458.

Lista de figuras: Workflows KNIME

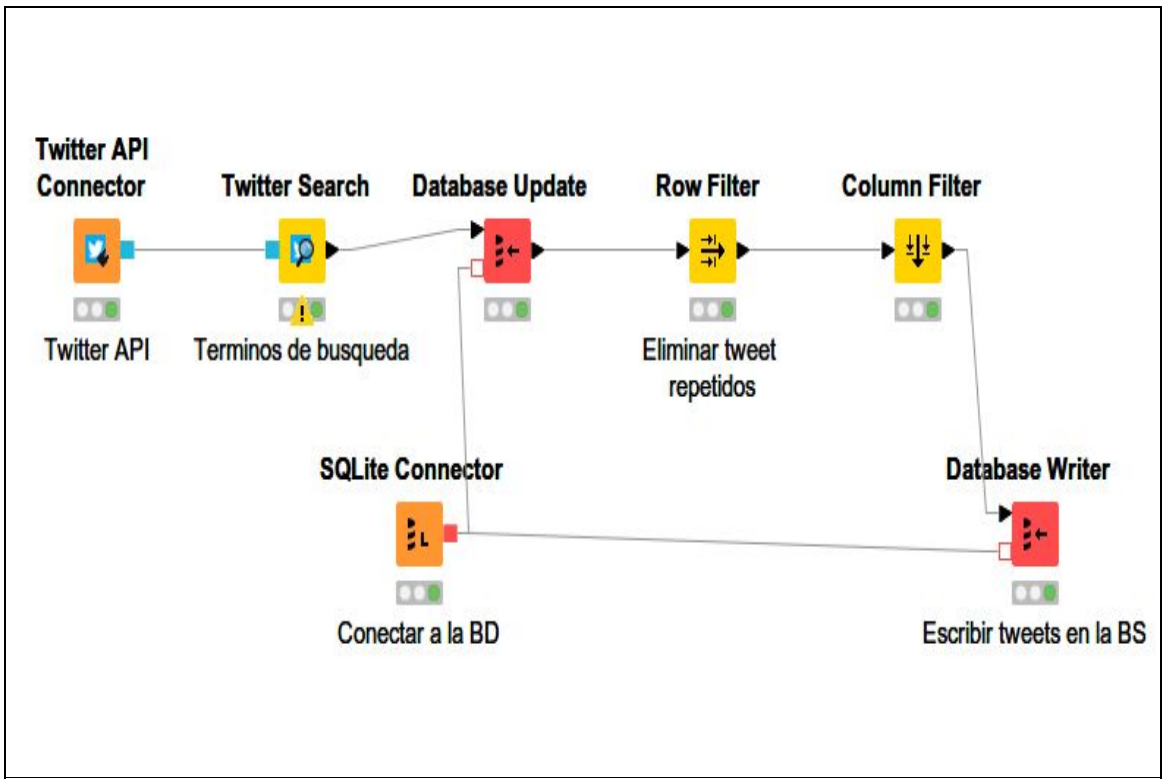

Figura 9 . Workflow para la extracción de datos de la red social Twitter (*ver capítulo 2.1*).

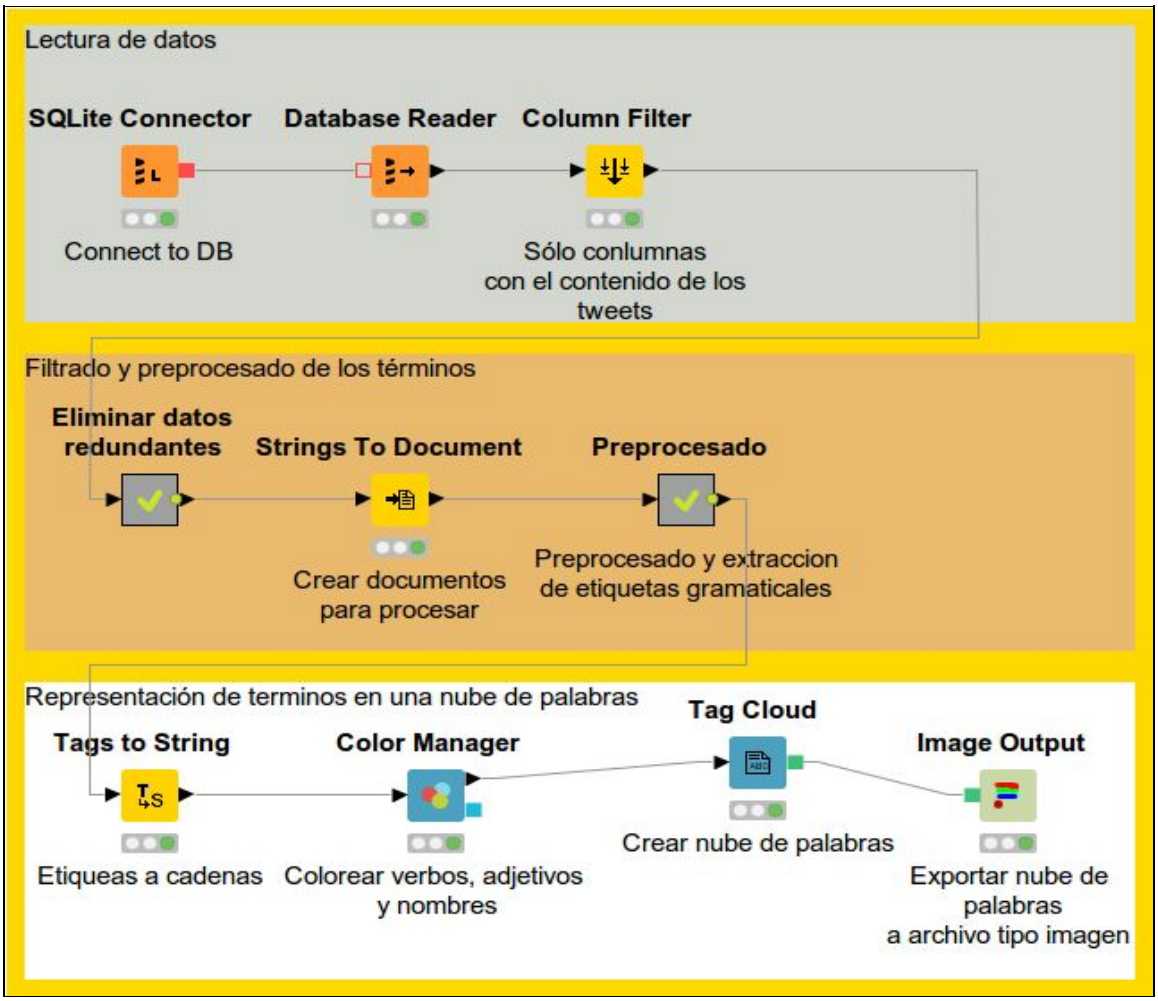

Figura 10. Workflow que permite la creación de la nube de palabras (*ver capítulo 2.3*).

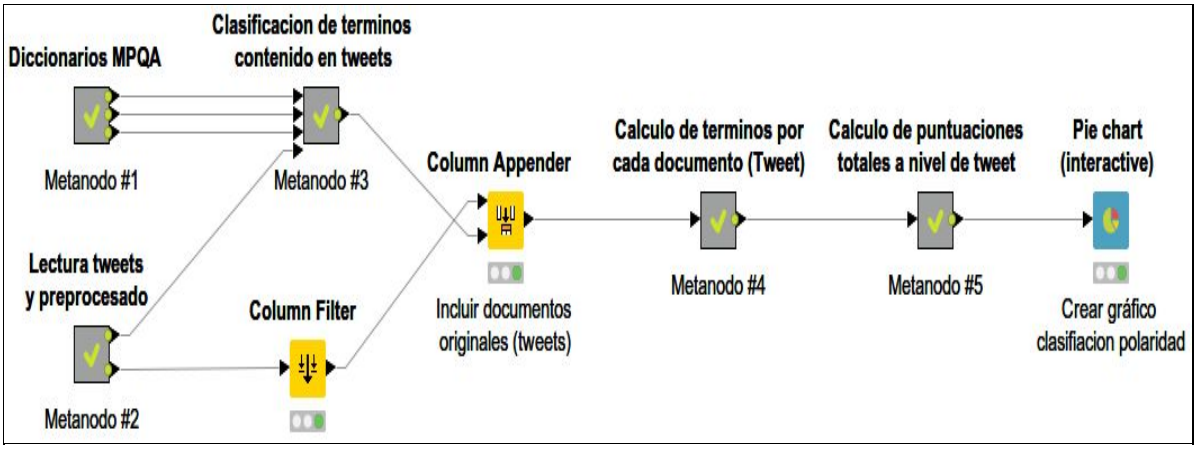

Figura 11. Workflow correspondiente a la clasificación de la polaridad de tweets (*ver capítulo 2.4)*.

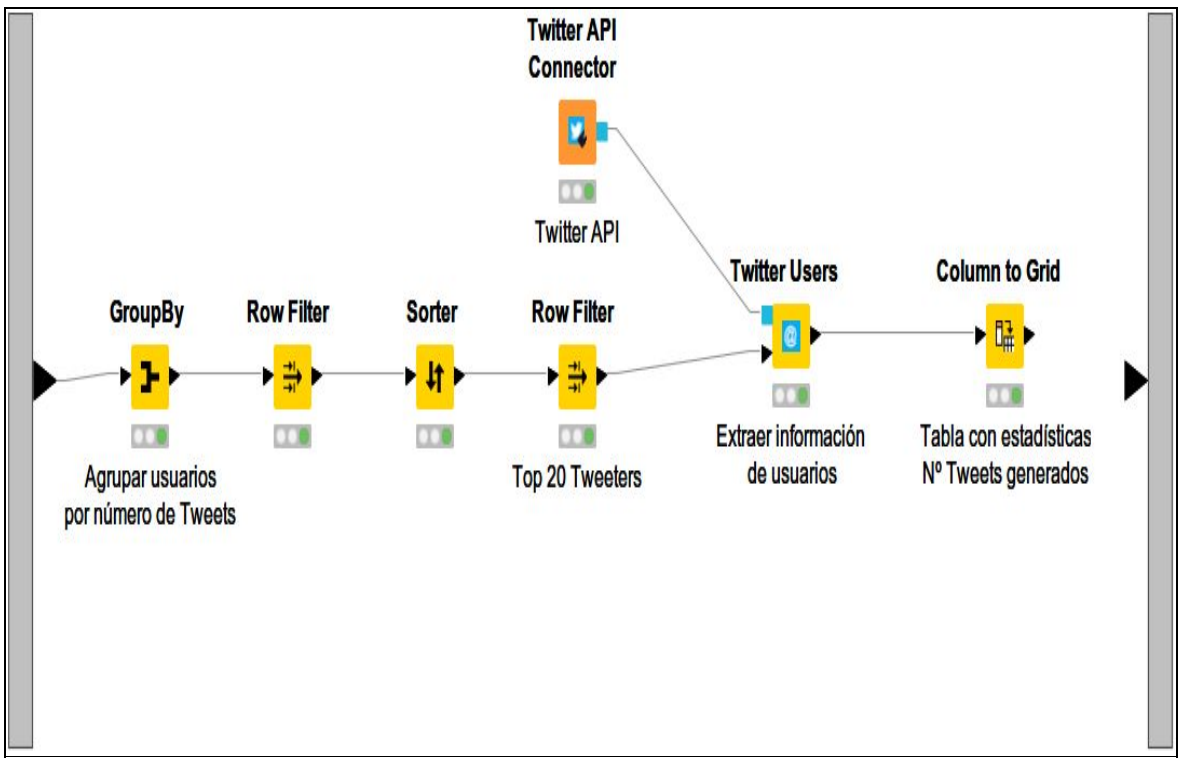

Figura 12. Workflow extracción información a nivel usuario (*ver sección 2.5.3*).

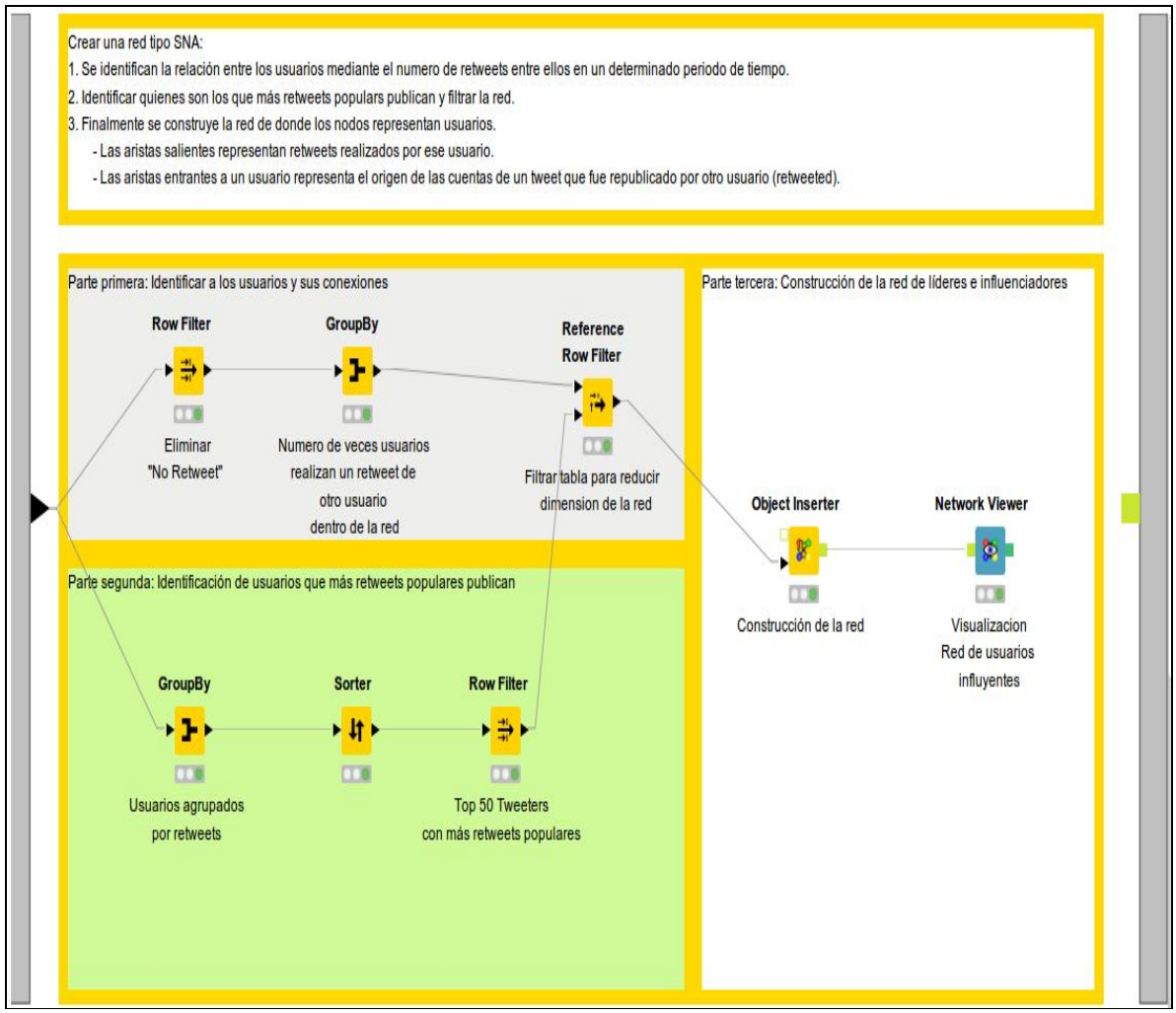

Figura 13. Workflow que permite la construcción de la red de influenciadores (ver sección 2.4.4).