UNIVERSIDAD DEL AZUAY

FACULTAD DE CIENCIAS DE LA ADMINISTRACIÓN

ESCUELA DE INGENIERÍA DE SISTEMAS Y TELEMÁTICA

MINERÍA DE TEXTO EN MEDIOS SOCIALES.
CASO DE ESTUDIO DEL PROYECTO TRANVÍA DE CUENCA.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SISTEMAS Y TELEMÁTICA

AUTOR: MARÍA BELÉN ARIAS ZHAÑAY

DIRECTOR: ING. MARCOS ORELLANA CORDERO

CUENCA, ECUADOR

2016

Dedicatoria

Dedico este trabajo a mis padres, Armando y Teresita, quienes son el soporte fundamental de mi vida.

A mis hermanas, Camila y Michelle, por ser el impulso diario.

A mi abuelita, Beatriz, por ser la inspiración.

A Javier, que desde el cielo, festeja este título de Ingeniero de Sistemas y Telemática.

Y al apoyo incondicional y desinteresado de Marcos Calle.

Agradecimientos

En el presente trabajo de tesis quiero agradecer principalmente a Dios y María Auxiliadora, por las bendiciones constantes en mi vida que han facilitado el cumplimiento de esta meta.

A mis padres por brindarme todo el apoyo, tanto económico como motivacional, para culminar con éxito mi carrera profesional.

A mi director, Ing. Marcos Orellana, por su apertura, disponibilidad, apoyo, tiempo e interés en el desarrollo de este trabajo.

Así también, quiero agradecer a mi asesor metodológico, Ing. Francisco Salgado, por su gran apoyo constante y paciencia a lo largo del proceso.

Índice

Índice de imágenes

Índice de Tablas

Resumen

Se investiga la minería de textos en datos no estructurados. Se tomó como caso de estudio el proyecto de Tranvía Cuatro Ríos de Cuenca, en concreto los datos no estructurados provenientes de las opiniones que las personas emitían a través del medio social *Twitter*. Para la extracción, almacenamiento, representación, aplicación de los algoritmos de minería de textos y visualización de resultados se utilizó el lenguaje *R*. Se aplicó el algoritmo de modelo vectorial para encontrar palabras frecuentes y algoritmo de asociación a-priori con el fin de encontrar patrones en la colección de datos y el sentido de positividad/negatividad predominante en los *tweets* emitidos por los usuarios.

ABSTRACT

Unstructured data text mining is the topic of this research. The tram project of Cuenca was taken as a case study, in particular the unstructured data from the opinions that people posted through Twitter social network. The R language was used for the extraction, storage, representation, application of text mining algorithms and the visualization of results. The algorithm vector model was applied to find common words and priori - association algorithm in order to find patterns in data collection and the sense of positivity / negativity prevalent in tweets issued by users.

UNIVERSIDAD DEL AZUAY
Dpto. Idiomas

Translated by,
Lic. Lourdes Crespo

Capítulo I: Introducción

Introducción

El avance de la tecnología ha incorporado nuevos medios para la comunicación entre las personas; estos son los medios sociales en donde un individuo interactúa con otros e intercambian criterios. A partir de esto aparecen técnicas que permiten analizar su opinión respecto a un tema u otro.

La minería de textos es una de estas nuevas técnicas que nos permite encontrar patrones ocultos en datos no estructurados, tales como los textos expresados en un lenguaje natural (ej.: castellano, inglés).

El propósito de esta investigación es demostrar la eficacia de dichos métodos para obtener conocimiento a partir de la información expresada por usuarios de *Twitter* con respecto al proyecto Tranvía Cuatro Ríos de Cuenca.

1.1 Objetivos

1.1.2 Objetivos específicos

- Sistematizar información sobre conceptos y métodos de minería de textos.
- Obtener y depurar información no estructurada (*tweets* en lenguaje natural) de la red social Twitter de los ciudadanos.
- Aplicar algoritmos de minería de textos para analizar los *tweets* a través del software *R.*
- Representar objetivamente los resultados importantes utilizando materiales ilustrativos.

1.1.1 Objetivo general

- Aplicar técnicas de minería de datos para el análisis de sentimiento de los usuarios del medio social Twitter en el caso “Tranvía Cuatro Ríos de Cuenca”.

1.2 Justificación

La implementación del proyecto Tranvía Cuatro Ríos de Cuenca genera comentarios y posiciones por parte de los ciudadanos. Estos comentarios demuestran diferentes puntos de vista sobre el proyecto y su análisis brindará información a las autoridades para la toma de decisiones.

El análisis de estas opiniones permitirá cumplir con el objetivo principal de la investigación, ya que a través del uso de minería de texto se busca obtener indicadores que permitan evaluar la posición que tienen los ciudadanos en cuanto al proyecto. Se opta por este tipo de análisis ya que se podrá analizar datos no estructurados provenientes de los *tweets* generados por los usuarios en el medio social *Twitter*.

1.3 Alcance y Limitaciones

El alcance de este trabajo es obtener el análisis de sentimiento en los usuarios de la red social *Twitter* que se refieran al caso de estudio Tranvía. Mediante el análisis de gráficos que permitan obtener indicadores que se refieran al criterio de los usuarios.

Capitulo II: Marco Teórico

Introducción

En este capítulo se describirá el concepto de minería de datos, sus algoritmos y aplicaciones. Se explicará el concepto de minería de textos, minería web y sus tipos. Se ahondará en el concepto medios sociales, sus clases y estadísticas dentro de nuestro país. Posteriormente se brindará una breve descripción del medio social *Twitter* que será el utilizado en el presente proyecto. Se explicará la calidad de datos y los algoritmos de minería de textos. Finalmente, se detallará el concepto de análisis de sentimiento que será utilizado para la presentación de resultados.

2.1 Minería de Datos

En la actualidad, la información generada por el ser humano crece en cantidad, velocidad y en sus temáticas. Por años, esta información ha estado relegada y sin brindarle la verdadera importancia. Esta información contiene grandes cantidades de conocimiento que ha permanecido oculto. Existe una pirámide de la información conformada por cinco niveles en donde se muestra el proceso de la transformación de los datos en sabiduría. Estos niveles son: (i) ruido, (ii) datos, (iii) información, (iv) conocimiento y (v) sabiduría.

(i) Ruido: es la base de la pirámide formada por la cadena de caracteres sin sentido, significado ni orden, se desconoce su naturaleza.
(ii) Datos: en este nivel la cadena de caracteres adquiere un significado por ejemplo puede ser un nombre, sexo, una dirección, etc.
(iii) Información: en este nivel los datos adquieren relaciones entre sí y un mayor significado, por ejemplo: Vinicio vive en Cuenca.
(iv) Conocimiento: en este nivel la información se transforma en conocimiento y permite la toma de decisiones.
(v) Sabiduría: es la cima de la pirámide representa el metaconocimiento que es el conjunto de información, hechos y reglas sobre un dominio específico.

(Hernandez, 2013)

Con el avance de la tecnología y de la constante evolución del ser humano han empezado a surgir nuevas técnicas que utilizan estos datos para crear información útil y relevante. Una de estas técnicas es la minería de datos, que se encuentra en sus primeros inicios en nuestro

país, pero que vale la pena conocer sus conceptos y sobretodo su gran utilidad. Es una metodología poderosa para el análisis de datos y para el descubrimiento de conocimiento.

A la minería de datos se la define como "conjunto de técnicas para la representación, análisis, manejo y descubrimiento de conocimiento a partir de diversas fuentes de datos (bases de datos, web, archivos, etc.)" (McDonell & Otros, 2012). Involucra aspectos de la estadística y de la inteligencia artificial y la información extraída se emplea para la toma de las decisiones.

Estas técnicas han surgido debido al incremento exponencial de la capacidad de almacenamiento y de procesamiento de las computadoras en la actualidad. El objetivo de la minería de datos es extraer información importante de las grandes bases de datos, a través del uso de algoritmos, descartando información irrelevante mediante la detección de patrones, relaciones significativas, efectos de interacción, y otras características que están ocultas a simple vista. (Castro & Otros, 2012)

Actualmente el área de minería de datos ha cobrado relevante connotación, ya que se encarga de encontrar la información que a simple vista está oculta pero que es muy valiosa para generar nuevas estrategias. En los países del primer mundo es común que las organizaciones o las personas utilicen estas técnicas para analizar sus grandes bases de datos con el fin de basar sus decisiones en la información fruto de la aplicación de técnicas y métodos de la minería de datos. Por ejemplo, se puede obtener las preferencias de sus consumidores en un local de comida rápida.

La minería de datos realiza su análisis a través de dos actividades:

(i) Describir en detalle a los generadores de datos: en esta actividad se realiza una revisión íntegra de toda la información disponible, permitiendo encontrar los generadores de datos en cada momento.

(ii) Predecir su comportamiento: identificando futuras situaciones deseadas o no deseadas permitiendo la toma de decisiones en las mismas.

2.1.1 Técnicas de minería de datos

Las técnicas de minería de datos sirven para predecir comportamientos y tendencias futuras, lo que permite, por ejemplo, aumentar ventas o disminuir pérdidas; esto requiere un

conocimiento profundo del negocio tanto de sus bases de datos, de sus procesos, de sus empleados, etc. Si se conoce a fondo el objetivo de análisis, las variables, la calidad, la redundancia y los objetivos de la organización, aumentan las posibilidades de que la información extraída sea eficiente y de interés.

Algunos algoritmos de minería de datos son:

(i) Umbrales: este algoritmo se encarga de observar un registro conforme pasan los días y cuando detecta un valor más allá de los umbrales fijados indica de lo sucedido. Por ejemplo: controlar las ventas de un producto donde se fija límites de existencia máxima y mínima.

(ii) Tendencias: este algoritmo controla si de un periodo a otro existe un crecimiento o decrecimiento considerable de una variable de estudio. Por ejemplo: control de las ventas semanles de un producto en una zona geográfica que va estableciendo su línea base para que luego se convierta en una tendencia.

(iii) Franja de normalidad: este algoritmo controla que la variable de interés no sobrepase una "franja de normalidad" comparándola con su comportamiento en un período anterior de análisis. Por ejemplo: en época de frío se vende menos helado que en la de calor.

(iv) Comportamiento errático: este algoritmo se encarga de informar cuando el comportamiento de una variable registre tumbos: suba o baje.

(v) Máximos y mínimos: este algoritmo barre los registros e informa los valores más altos o bajos como se presentan. Por ejemplo: en un cierto caso se podría detectar qué productos son los que mayores ventas tienen; en un contexto de análisis epidemiológico, se podría encontrar una enfermedad que casi se ha erradicado.

(vi) Patrones frecuentes: este algoritmo se encarga de encontrar reglas, examinando todos los registros y encontrar los más frecuentes; una vez detectados se empieza por buscar pares de patrones, luego trio de patrones y asi sucesivamente. Por ejemplo: cada vez que alguien compra pan, compra leche, el patrón sería (leche, pan).

(vii) Reglas de asociación: este algoritmo toma los patrones frecuentes y se encarga de descubrir, por ejemplo, qué producto causa que se compren los otros. Por ejemplo: quien compra leche, también pan.

(viii) Cúmulos (clusters): este algoritmo utiliza técnicas de agrupación para clasificar los registros en categorías con propiedades parecidas entre sí, pero distintas a los pertenecientes a las otras categorías. Por ejemplo: clasificar a todos los pacientes por tipo de hospital (público, privado, beneficencia u otros).

Lo que diferencia a la minería de datos con respecto a la estadística, es que la estadística utiliza una muestra (una parte) de los datos, mientras que la minería analiza todos los datos. (Martinez Luna, 2011)

2.1.2 Clasificación de los algoritmos

Estos algoritmos pueden estar clasificados por su propósito general en:

Modelos predictivos

Se encargan de describir el modelo para los datos basados en un conocimiento previo, permiten predecir el valor de un atributo desconocido. Para aceptar como válido el proceso de minería de datos es necesario comprobar el modelo supuesto con los datos. Estos modelos deben pasar por cuatro fases:

(i) Identificación objetiva: el punto de partida son los datos a los cuales se les aplican reglas con el fin de identificar el mejor modelo que se ajuste a los datos.

(ii) Estimación: en esta fase se realiza el cálculo de los parámetros del modelo establecido en la fase anterior.

(iii) Diagnosis: se encarga de validar el modelo estimado.

(iv) Predicción: toma el modelo que ha pasado por las tres fases anteriores, es decir, el modelo identificado, estimado y validado para predecir valores futuros o prospectivos de las variables dependientes.

En algunos casos, el modelo se obtiene como mezcla del conocimiento obtenido antes y después de aplicar minería de datos, por ejemplo, las redes neuronales van descubriendo modelos complejos y mejorándolos a medida que se exploran los datos, debido a su

capacidad de aprendizaje, descubren relaciones complejas entre las variables sin intervención externa.

En los modelos predictivos se incluyen todos los tipos de regresión, series temporales, análisis de varianza y covarianza, análisis discriminante, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas.

Modelos descriptivos

Estos algoritmos no suponen la existencia de variables dependientes o independientes, no se crea un modelo previo a los datos. Estos modelos son creados a partir del reconocimiento de patrones.

En los modelos descriptivos se incluyen los algoritmos de clustering y segmentación, técnicas de asociación y dependencia, técnicas de análisis exploratorio de datos, las técnicas de la dimensión y de escalamiento multidimensional.

Modelos auxiliares

Algoritmos de apoyo específicos, basados en técnicas estadísticas descriptivas, consultas e informes cuyo objetivo es la verificación. (Pérez López & Otros, 2007)

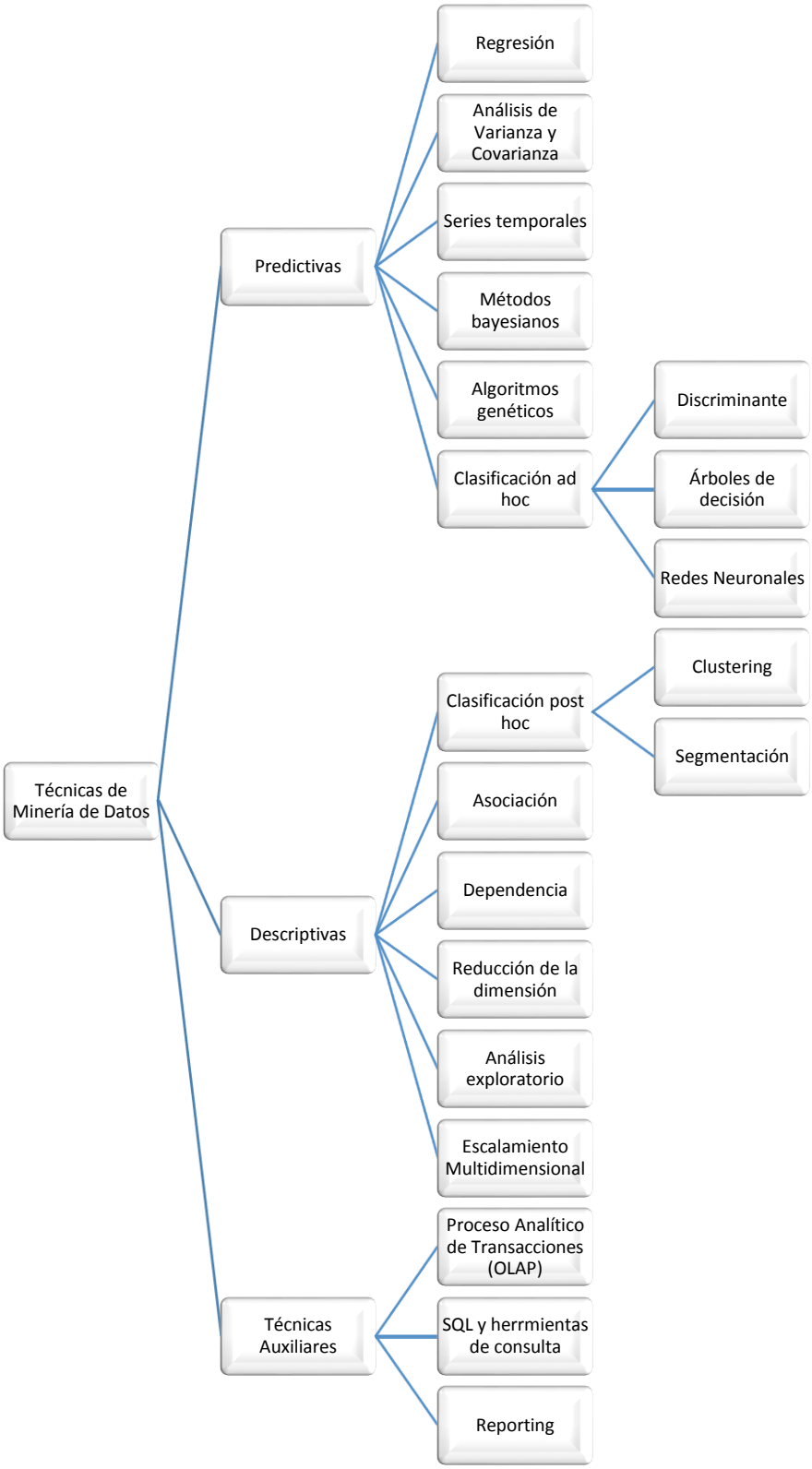


Ilustración 1 Clasificación Técnicas de Minería de Datos (Pérez López & Santín González, 2007)

En el gráfico anterior podemos observar que existen técnicas de clasificación que pertenecen sea al grupo de modelos predictivos o al de los descriptivos. Las técnicas de clasificación predictivas se denominan técnicas de clasificación ad hoc, debido a que estas clasifican individuos dentro de grupos previamente definidos. Las técnicas que clasifican sin especificación previa de los grupos se denominan técnicas de clasificación post hoc, y son del grupo de las descriptivas. (Pérez López & Santín González, 2007)

La minería de datos ha sido aplicada en diferentes ámbitos, por ejemplo, para el análisis de organizaciones, desde la adquisición de clientes hasta el diseño de estrategias para aumentar sus ingresos. Se han analizado campañas comerciales, proveedores, gestión de inventarios. Por ejemplo, los bancos la emplean para analizar el perfil de clientes en cuanto a créditos. (Ocerín, 2011)

Así también, se aplica en áreas como la climatológica para predecir fenómenos naturales; en la medicina, con el fin de encontrar la eficacia de un tratamiento; en la televisión, para identificar los niveles de audiencia; y en la industria, con el fin de predecir fallas; en la gestión de movilidad, para la predicción del tráfico en las grandes ciudades.

2.2 Minería de Textos

Minería de texto son un conjunto de métodos que, mediante los datos aportados por los usuarios, obtiene información y encuentra los tópicos más hablados, que permiten categorizar a los usuarios según la información aportada. En la minería de texto cada opinión es tratada como un documento. (Lobos & Bartolome, 2012)

La minería de texto según García-García es aquella que ofrece con sistemas automáticos una solución, sustituyendo o completando el trabajo de personas, sin que importe la cantidad de texto. Encuentra información que antes era desconocida. Pueden ser patrones que de otro modo serían extremadamente difícil de descubrir. (García-García, & otros, 2014)

Así también, la minería de textos consiste en extraer la información que será útil de documentos no estructurados e identificar patrones interesantes de conocimiento (Rodriguez Aldape, 2013). La minería de textos es una extensión de la minería de datos ya

que muchas de las técnicas utilizadas en la minería de datos pueden ser aplicadas en la de textos. Ésta, busca obtener información a partir de grandes repositorios de datos.

En resumen, la minería de textos se encarga de examinar una colección de datos no estructurados con el fin de identificar patrones que generen información, sin importar que esta colección sea de gran tamaño, ya que sustituye muy eficientemente con sistemas automáticos el trabajo que difícilmente realizarían las personas.

2.2.1 Áreas de la minería de textos

Feldman y Sanger la dividen en cuatro áreas: tareas de pre-procesamiento, operaciones de minería, presentación y navegación, y técnicas de refinamiento.

Tareas de pre-procesamiento

Este proceso incluye rutinas, procesos y métodos requeridos para obtener y preparar los datos para las operaciones de descubrimiento de información, es decir, es el proceso en el los textos se convierten en una estructura que facilite su análisis. Estas tareas se centran en las actividades de preprocesamiento y categorización de los datos, se encargan de convertir los datos en un formato canónico, antes de aplicar diferentes tipos de extracción de características a estos documentos con el fin de representarlos por conceptos.

Operaciones de minería

Estas operaciones son el corazón de un sistema de minería de textos, se incluye la detección de patrones, análisis de tendencias y algoritmos de descubrimiento de conocimiento. Los patrones más comunes que se utilizan son: distribuciones y proposiciones, conceptos del cercano más frecuente y asociaciones. Así también, realiza comparaciones entre los niveles de intereses de esos patrones.

Presentación y navegación

Se incluye herramientas de visualización y navegación de los patrones orientados al usuario. La capa de presentación puede incluir caracteres o herramientas de gráficos para crear o modificar las agrupaciones conceptuales, así como para la creación de perfiles especificos para patrones o conceptos.

Técnicas de refinamiento

Se incluye los métodos que filtran información redundante, la ordenan, resumen, generalizan y agrupan. Estas técnicas pertenecen a la etapa de pos-procesamiento, en ella se realizan actividades como: suprimir, ordenar, poda, generalizar y agrupar con el fin de optimizar la información adquirida. (Feldman & Sanger, 2007)

Las técnicas de pre-procesamiento y de refinamiento son operaciones de la minería de texto más críticas, ya que en estas se preparan los datos para la aplicación de los algoritmos y se depuran los datos luego de la aplicación de los mismos.

Estas técnicas se aplicarán dentro del proyecto, ya que se tomará en cuenta operaciones de minería de textos, entre ellos podemos mencionar por ejemplo que los *tweets* generados son parte de los usuarios e inherente a su lenguaje natural, es decir, son datos no estructurados, que, gracias a la minería de textos, podrán ser analizados para describir la opinión de los usuarios en cuanto al proyecto Tranvía Cuatro Ríos de Cuenca.

2.3 Minería Web

La minería web consiste en aplicar las técnicas de minería de datos a documentos y servicios de la web. Consiste en extraer la información, imágenes, textos, audio, video, documentos y multimedia de un sitio web. La minería web se diferencia de la minería de datos en que su procesamiento se realiza en línea mientras que la otra fuera de línea. (Sharma & Otros, 2011)

Cuando un usuario utiliza la web deja registros digitales que pueden ser las direcciones IP, navegador, sitios visitados, videos e imágenes descargadas que son almacenados en los *logs* del servidor. Son estos *logs* los que utiliza la minería web para analizarlos y obtener conocimiento.

Según los objetivos de análisis se puede dividir en tres tipos: (i) minería del contenido de la web, (ii) minería de la estructura de la web, (iii) minería de los registros de navegación en la web.

(i) Minería del contenido web: esta se encarga de extraer información o conocimiento del contenido de una página web. Se puede descubrir

descripciones de productos, fotografías publicadas, opiniones de clientes y sentimientos de consumidores.

(ii) Minería de la estructura de la web: se encarga de extraer información de los hipervínculos respecto a la estructura de la página que pudiera ser importante. Se puede descubrir comunidades de usuarios con intereses en común. Puede ser útil para clasificar o agrupar documentos.

(iii) Minería de los registros de navegación en la web: se encarga de extraer información de la relación que existe entre documentos en la web por búsquedas anteriores, registrando búsquedas y accesos de los usuarios a los mismos, es decir, extrae información de los hábitos y preferencias de los usuarios.
(Guevara, 2011)

Es oportuno mencionar que la minería de contenido web, se especializa en la localización de patrones en el texto de los documentos, de modo que en este tipo se basará el proyecto ya que el análisis de la opinión de los usuarios se encuentra en el contenido del medio social *Twitter*, por ello esta área es la indicada para el proyecto.

2.4 Medios Sociales

Los medios sociales surgen a finales del siglo XX, alrededor de 1997. No todos han tenido la suerte de nacer y de continuar en vigencia, ya que algunos tienen más usuarios y fama debido a los servicios que proporcionan, a pesar de esto existen para todos los gustos.

Los medios sociales son la reunión de personas que interactúan entre sí mediante medios digitales donde pueden conocerse o no, pero existe retroalimentación entre ellos. Es ahí en la retroalimentación, donde surge la opinión de los usuarios en uno u otro tema. (Domínguez, 2010)

Los medios sociales toman lo que cada usuario publica y esto no se queda ahí, sino empieza un diálogo virtual entre los mismos, tomando sus referencias y opiniones, generando así una comunicación más realista, debido a que un usuario se siente mucho más cómodo de dar su idea a través de un medio social, que decirla personalmente. En realidad, un medio social refleja la personalidad de un individuo, a tal punto que para Aciar y colaboradores (2015) los usuarios expresan sus opiniones en formato de texto libre usando su lenguaje

natural. A través de los medios sociales, los usuarios se conectan, comparten y discuten cualquier tema, en cualquier lugar y en cualquier momento (Del Fresno & Otros, 2015).

El concepto «medios sociales» (*social media*) describe la confluencia de distintas aplicaciones de la llamada web 2.0, que permiten de un modo sencillo, la creación, la clasificación y el intercambio de contenidos generados por los propios usuarios de la red.

Los medios sociales se basan en la teoría de los seis grados de separación, según la cual, cualquier persona en el planeta puede estar conectada a cualquier otra a través de una cadena de conocidos, que tiene, en promedio, seis personas de separación. Si dichas conexiones están ocultas no sería un problema ya que internet se encarga de hacerlas visibles.

2.5 Clasificación de medios sociales

Los medios sociales pueden clasificarse en: (i) medios sociales directos y (ii) medios sociales indirectos.

2.5.1 Medios sociales directos

Son aquellos en los que los usuarios pueden controlar la información que comparten utilizando servicios de internet entre grupos de personas que comparten intereses en común y en igualdad de condiciones. Los usuarios crean perfiles en los que pueden manejar su información personal y la relación con otros usuarios, controlan la privacidad y el acceso a su información.

Estos medios directos pueden clasificarse según el enfoque empleado:

(i) Según la finalidad: analiza el objetivo que persigue el usuario del medio social.

   a. Según finalidad de ocio: el usuario busca principalmente entretenimiento, mejorar sus relaciones con otros usuarios mediante el intercambio de información en forma de texto, imágenes o videos a través de comentarios, y publicaciones.
   b. Según finalidad de uso profesional: el usuario busca informar sobre su nivel profesional, conocimientos o conocer colegas en su área.

(ii) Según el modo de funcionamiento: analiza el conjunto de procesos que estructuran los medios sociales.

   a. Según modo de funcionamiento de contenidos: el usuario crea y distribuye contenido en forma de texto, imágenes o videos a través del medio social con otros usuarios. La diferencia con el resto de medios es que el contenido está disponible sin necesidad de crear un perfil de usuario.
   b. Según modo de funcionamiento de perfiles tanto profesionales como personales: conformado por fichas en donde los usuarios llenan con su información personal o profesional, es obligatorio la creación de un perfil para emplear las funciones del medio social.
   c. Según modo de funcionamiento de *microblogging*: diseñadas para compartir y comentar información que se mide en caracteres que puede emitirse desde dispositivos fijos o móviles.

(iii) Según grado de apertura: analiza la capacidad de acceso por parte de cualquier usuario.

   a. Según grado de apertura pública: utilizadas por cualquier tipo de usuario sin necesidad de pertenecer a un grupo u organización.
   b. Según grado de apertura privada: el acceso es solo para usuarios que pertenezcan a una organización privada que se hace cargo del costo de la misma.

(iv) Según nivel de integración: analiza el nivel de afinidad, interés del usuario.

   a. Según nivel de integración vertical: utilizadas por usuarios que tengan el mismo interés profesional.
   b. Según nivel de integración horizontal: utilizadas por usuarios que no tengan un interés concreto.

Facebook, Twitter, YouTube, Wikipedia, Hi5, Meetic, LinkedIn, Xing, MySpace, Fotolog son ejemplos de medios sociales directos. A continuación, se presenta una tabla que describe la situación de algunos medios sociales según la clasificación descrita anteriormente.

	Según finalidad		Según modo de funcionamiento			Según grado de apertura		Según nivel de integración	
	De ocio	De uso profesional	De contenidos	perfiles personales/	Microblogging	Públicas	Privadas	De integración vertical	De integración horizontal
Facebook	**X**	**X**		**X**		**X**			
You Tube	**X**		**X**	**X**		**X**			**X**
Twitter	**X**	**X**		**X**	**X**	**X**			**X**
LinkedIn		**X**		**X**		**X**			**X**
Yammer		**X**		**X**			**X**		
Dir&Ge		**X**		**X**				**X**	

Tabla 1 Tabla de clasificación de medios sociales (Ureña, & Otros, 2011)

La finalidad de cada medio social se clasifica de variadas formas que como se muestran en el cuadro, pueden ir desde medios de ocio hasta medios de uso profesional. En el caso de la red *LinkedIn* por ejemplo, su tendencia es de uso profesional, pero si observamos al medio social *Twitter* según su funcionalidad puede ser de ocio y de uso profesional, puede ser según su modo de funcionamiento personal y profesional, según grado de apertura es pública y según nivel de integración es horizontal.

2.5.2 Medios sociales indirectos

Son aquellos en los que los usuarios no suelen tener un perfil visible para todos siendo controlados por un individuo o un grupo que controla y dirige la información en torno a un tema concreto. Se clasifican en:

(i) Foros: son servicios a través de internet por parte de expertos dentro de un área de conocimiento específico, se realizan intercambios de información y opiniones. La relación es bidireccional permitiendo responder preguntas planteadas.

(ii) Blogs: son servicios a través de internet con elevado grado de actualización, recopilación cronológica de uno o varios autores. (Ureña & Otros, 2011)

En nuestro país, los medios sociales han tenido un crecimiento exponencial en los últimos años, ya que de un momento a otro se expandieron en todos los niveles sociales del país. Es notorio que la gran mayoría de los ecuatorianos cuentan con un medio social, tal es así que según datos del INEC (Instituto Nacional de Estadísticas y Censos), *Facebook* es el medio social más utilizado en el Ecuador; cerca del 98% de las personas cuya franja de edad está sobre los 12 años posee una cuenta de Facebook. En cambio, *Twitter* cuenta con un promedio de 2'000.000 de usuarios (datos obtenidos a enero de 2015). (El Comercio, 2015)

Como se comentó en la sección anterior, *LinkedIn* es el principal medio social para la búsqueda de empleo, contactos profesionales, grupos de discusión de temas empresariales, negocios e industriales en el Ecuador. Este medio tiene 1'251.148 usuarios registrados.

Twitter Marketing Conference 2015 es un informe generado por Formación Gerencial, identifica los medios sociales más visitados en Ecuador. Los datos obtenidos de esta fuente están descritos a continuación:

Puesto	Medio Social
1	Facebook
2	Youtube
3	Twitter
4	Ask
5	Instagram
6	Slideshare
7	LinkedIn
8	Scribd

9	Pinterest
10	Badoo
11	Tumblr
12	Twoo
13	Hi5
14	Tagged
15	Flickr

Tabla 2 Formación Gerencial Internacional y Ranking Alexa 4 de Enero de 2015 (Del Alcanzar Ponce, 2015)

En conclusión, los medios sociales se han convertido en vías de relación entre las personas utilizando medios digitales, estas personas pueden ser conocidas o no, e intercambian opiniones entre ellas utilizando texto, imágenes, y videos, dando a conocer su punto de vista frente a un tema u otro.

Estos medios abarcan diferentes temáticas de relación, por ejemplo, puede involucrar el área personal o profesional, se buscan temas de interés e interactuar criterios entre sí. La comunicación es bidireccional, es decir, existe retroalimentación entre usuarios.

Los medios sociales se encuentran en auge en nuestro país, es decir, se encuentran con una popularidad muy alta y las personas lo utilizan para socializar y compartir información. Así también, los negocios utilizan estos medios para aumentar la popularidad entre sus clientes o como medio de publicidad entre los mismos.

Un medio social se ha convertido en el canal de comunicación de los ecuatorianos, ya que es común informarse de noticias a través de estos incluso mucho más rápido que por la televisión o por la radio. Es normal observar a una persona con su dispositivo móvil, leyendo noticias que previamente fueron expuestas en un medio social.

En este punto se puede argumentar que es conveniente utilizar un medio social como *Twitter* para analizar la opinión de los ciudadanos en la implementación del proyecto

Tranvía Cuatro Ríos de Cuenca, debido a que se encontrarán los criterios de los usuarios y se podrá obtener juicios positivos, negativos o neutros que serán analizados con minería de textos, facilitando la toma de decisiones.

2.6 Twitter

Twitter es un medio social de microblogging que surgió en el año de 2006 y que se caracteriza por permitir a sus usuarios enviar y publicar mensajes de 140 caracteres denominados *tweets* (trinos), representa una fuente de opinión de millones de usuarios en temas de actualidad. El concepto de este medio social se basa en tener una cuenta y amigos conectados a los cuales se les informa lo que va sucediendo a través del uso de los *tweets*. (Sánchez Arteaga, 2011)

Un usuario de *Twitter* contará con su perfil y para interactuar con los otros usuarios utilizará la opción de "seguir" a otros usuarios de esta manera logrará la comunicación entre los mismos.

Twitter se ha convertido en una herramienta de comunicación que crece de manera exponencial, se extiende en temas sociales, culturales, políticos y económicos de un país o de una región, esto se debe a que el medio permite enterarse de los seguidores de políticos, artistas, cantantes y de lo que cada usuario opina en diferentes áreas.

Es una poderosa herramienta de popularización por ejemplo el presidente de Estados Unidos Barack Obama utilizó este medio para aumentar su popularidad cuando se encontraba en campaña ya que cuando la inició no era muy conocido. (Fainholc, 2011)

Twitter es un medio social que se caracteriza por analizar los temas de mayor impacto mediáticos a través de los "*hashtags*" y "*trending topics*", en un momento dado, lo que permite obtener conclusiones sobre los temas que le preocupan a un ciudadano o a un medio, es por esto que este medio nos permitirá el análisis de la opinión de los ciudadanos de Cuenca en cuanto a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca. (Fernández, 2012)

La principal característica de este medio social es que establece nuevas tendencias de la información global, por lo tanto se convierte en un medio de comunicación donde se comparte hechos y noticias, esto lo diferencia de *Facebook* en donde el aspecto es el

personal. Así también, los usuarios de *Twitter* son aquellos que tienen edades comprendidas entre los 20 y 45 años, es decir, es un público homogéneo. En contraste con *Facebook* que sus usuarios no están en un rango de edad definido siendo usuarios heterogéneos.

2.7 Calidad de Datos

Se puede definir calidad de datos como un concepto multifacético que se encarga de la exactitud, completitud, consistencia y actualidad de los datos. Si los datos son de mala calidad afectan de manera muy significante y profunda en la efectividad y eficiencia de las organizaciones, llevando en algunos casos a pérdidas millonarias. (Bianchi Gallo & Otros, 2009)

Para lograr una calidad de datos se presentan las siguientes áreas:

(i) Dimensiones: mediciones sobre el nivel de calidad de los datos que se aplican a las dimensiones de interés.
(ii) Metodologías: proveen guías de acción.
(iii) Modelos: representan las dimensiones y otros aspectos de la calidad de datos.
(iv) Técnicas: brindan soluciones a problemas de calidad de datos.
(v) Herramientas: facilitan de manera efectiva que se realicen las metodologías y técnicas.

2.7.1 Dimensiones de la calidad de datos.

Una dimensión es la encargada de reflejar un aspecto distinto de la calidad de datos, es decir, las dimensiones reflejan las características de los datos, estas dimensiones pueden referirse a su valor o su esquema y definen un factor de calidad como un aspecto particular de una dimensión, de manera similar, una métrica define la forma de medir un factor de calidad y el proceso que implementa una métrica se denomina método de medición.

Exactitud y unicidad

La correcta y precisa asociación entre los objetos del mundo real y los estados del sistema de información se denomina exactitud. Existen tres factores de exactitud:

(i) Exactitud sintáctica: se refiere a la cercanía entre un valor *v* y los elementos de un dominio *D*, es decir, si para *v* le corresponde un valor válido de *D* sin importar si este valor pertenece a uno del mundo real.

(ii) Exactitud semántica: se refiere a la cercanía que existe entre un valor *v* y un valor real *v'*, es decir, mide qué tan bien se encuentran representados los estados del mundo real y se mide con valores booleanos.

(iii) Precisión: se refiere al nivel de detalle de los datos.

La unicidad es la dimensión encargada de verificar que las tuplas de información que representan el mismo objeto del mundo real tengan diferentes claves. La unicidad presenta dos factores:

(i) Duplicación: cuando una entidad aparece repetida de manera exacta.

(ii) Contradicción: cuando una entidad aparece repetida con contradicciones.

Completitud

Se define como la medida en que los datos son de suficiente alcance y profundidad. Se presentan dos factores de completitud:

(i) Cobertura: se refiere a la porción de datos de la realidad que se encuentran en el sistema de información, así también, compara la información del sistema con el mundo real.

(ii) Densidad: describe la cantidad de información contenida y la que falta acerca de las entidades del sistema de información.

Dimensiones relacionadas con el tiempo

La actualización de los datos es importante para la calidad de datos ya que un dato que no se encuentre actualizado es de mala calidad y puede llegar a ocasionar grandes problemas. Existen las siguientes dimensiones relacionadas con el tiempo:

(i) Actualidad: se refiere a la actualización de los datos y su vigencia, se mide según la información de la última actualización.

(ii) Volatilidad: trata la frecuencia con la que los datos cambian en el tiempo, para medir esta dimensión se utiliza la cantidad de tiempo que los datos permanecen válidos.

(iii) Edad: define que tan actuales o antiguos son los datos o los eventos en cuestión, se considera una métrica de actualidad y se comprueba que los datos estén dentro del límite establecido.

Consistencia

Esta dimensión verifica el cumplimiento de las reglas semánticas de los datos, es decir, comprobar que no exista más de un estado del sistema asociado al mismo objeto de la realidad, por ejemplo, si en una tabla se almacenan datos de una persona como la fecha de nacimiento y la edad pero la edad no coincide con la fecha almacenada existe una inconsistencia.

Relaciones entre las dimensiones

Las dimensiones se interrelacionan de manera estrecha, es decir, no son independientes entre sí. Al momento de mejorar las dimensiones se debe cuidar esta, ya que podría afectar negativamente a otra, por lo que es conveniente considerar a las dimensiones de mayor valor y dar prioridad según la magnitud en la que afectan al sistema.

2.7.2 Técnicas y Actividades de Calidad de Datos

Con el objetivo de mejorar la calidad se realizan actividades que utilizan técnicas. Algunas actividades son:

(i) Obtención de nueva información: se encarga de actualizar la información almacenada con datos de mayor calidad

(ii) Estandarización: se encarga de almacenar los datos con un formato en común.

(iii) Identificación de objetos: se identifica registros que hacen referencia al mismo objeto de la realidad.

(iv) Integración de datos: se encarga de resolver problemas de redundancias, consistencia y duplicación.

(v) Confiabilidad de las fuentes: verifica las diferentes fuentes de información según la calidad de los datos que poseen.

(vi) Composición de calidad: calcula la composición o agregación de las dimensiones de calidad de datos.

(vii) Detección de errores: detectar que registros no cumplen con las reglas.

(viii) Corrección de errores: se encarga de rectificar errores con el fin de que todas las reglas se respeten.

(ix) Optimización de costos: mejorar la relación costo-beneficio.

2.7.3 Limpieza de Datos

Con el fin de mejorar la calidad de los datos se intenta resolver la problemática de la detección y corrección de errores e inconsistencias que ocurren en los datos, esto se conoce como limpieza de datos.

Fases de limpieza

Un proceso de limpieza de datos consta de las siguientes fases:

(i) Análisis de datos: se determinan los errores e inconsistencias que se deben eliminar.
   a. *Data profiling*: se analizan los datos de una base de datos con el fin de obtener propiedades de la misma.
   b. *Data Mining*: identifica patrones en conjunto de datos.

(ii) Definición de transformaciones de datos y reglas de mapeo: consiste en realizar transformaciones a nivel de esquema e instancias.

Detección y corrección de errores

En esta etapa se realizan tareas de verificación de los datos, es decir, después de identificar los datos se debe verificar si estos son correctos o no y después se debe corregirlos.

Prevención de errores

Consiste en evitar que las inconsistencias se produzcan en un futuro. (Bianchi Gallo & Otros, 2009)

2.7.4 Calidad de datos en redes sociales

Los medios sociales se han convertido en grandes repositorios de información sobre las personas, esta información si es tratada con algoritmos de minería de textos genera conocimiento que permite tomar decisiones. Sin embargo, se debe comprobar la calidad de la información previo al uso de los algoritmos, debido a que podría ser inexacta, incompleta o desactualizada.

Los datos que se generan a través de un medio social reflejan la opinión de un usuario en uno o en otro tema, por ejemplo, un usuario de *Twitter* puede expresar su opinión sobre un político o sobre una nueva ley. Es por esto, que se ha elegido este medio para analizar la

opinión de los usuarios sobre la implementación del proyecto Tranvía Cuatro Ríos de Cuenca.

A pesar de estas ventajas, es necesario comprobar que la información que se va a utilizar sea verídica, ya que las cuentas de usuarios podrían ser generadas por robots. Una cuenta robot también se conoce como usuario fantasma, cuya finalidad es crear una falsa comunidad con un alto número de seguidores en muy poco tiempo. Se caracteriza por publicar enlaces repetidos, por alto índice de usuarios que lo han bloqueado o reportado y alto índice de duplicación de avatar. Un avatar es una representación gráfica de la forma humana para su identificación. Algunos *bots*, como también se los conoce, publican *tweets* automáticamente con enlaces a algún sitio, dan *retweet* o siguen automáticamente a un usuario que mencione alguna palabra o *hashtag*. Sin embargo, existen páginas web que permiten la detección de este tipo de usuarios lo que facilitará la eliminación de estos usuarios en el análisis. También, la información podría contener datos irrelevantes, repetidos, vacíos o equivocados, por lo que es conveniente realizar una limpieza de datos. (Ponce, 2012)

2.8 Algoritmos

2.8.1 Categorización

Este algoritmo se encarga de clasificar un dato dado en un conjunto predefinido de categorías, a esto se lo conoce como categorización de texto o por sus iniciales CT, el proceso consiste en que a partir de un conjunto de categorías y una colección de documentos de texto, se encuentre la categoría a la que pertenece dicho documento.

La categorización surgió en la década de 1960, donde su principal función era ayudar en la indexación de la literatura científica, por medio de vocabulario controlado. Para 1990, se desarrolló completamente, ya que se analizó un mayor número de documentos de texto en formato digital, y se lo organizó eficientemente para facilitar su uso.

La CT en nuestros días se aplica en una variedad de contextos: transacciones online, filtrado de correos no deseados, páginas web, generación de metadatos, entre otros. Los dos enfoques principales encargados del estudio de la clasificación de textos son: la Ingeniería del Conocimiento y el Aprendizaje Automático.

La Ingeniería del Conocimiento es un área enfocada a generar nuevo conocimiento, que antes no existía, basándose en la información que se encuentra en bases de datos documentales y a través del cruce del contenido de los documentos (Bailón-Moreno, 2013), en tanto el aprendizaje automático se refiere a crear algoritmos que sean capaces de reconocer patrones y generalizar comportamientos, es decir, es un proceso de inducción del conocimiento ya que permite obtener un enunciado general a partir de casos particulares. (Caparrrini, 2015)

Las aplicaciones más comunes de la clasificación de texto son:

(i) Indexación automática: en esta aplicación la CT asigna a cada documento, palabras o frases que lo describen, para ello utiliza un conjunto controlado de términos y frases claves llamado diccionario, el cual es creado por una persona.
La indexación de texto es utilizada frecuentemente en los sistemas booleanos de recuperación de información, estos sistemas se encargan de revisar documentos y generar índices a partir de su contenido como lo hacen los buscadores, por ejemplo: *Google*, *Yahoo*, entre otros.

(ii) Filtrado de texto: Esta técnica se refiere a la actividad de clasificar un flujo de documentos por un filtro especificado por el usuario, ya que uno de los grandes problemas en categorización de textos es definir la categoría a la que pertenece cada documento. En esta técnica se realiza solo en dos contenedores: "irrelevantes" y "relevantes", por ejemplo, un usuario puede clasificar en "correo deseado" y "correo no deseado".

(iii) Categorización Jerárquica de páginas web: se realiza la clasificación de páginas web bajo un catálogo, que permite la navegación directa y restringir la búsqueda basado en consultas a páginas que pertenecen a un tema en particular; como lo hacen los motores de búsqueda. En las aplicaciones anteriores se limitaba el número de categorías, en esta se limita el número de documentos que puedan pertenecer a una categoría, por lo que el sistema de clasificación debe ser compatible y eliminar los obsoletos. (Feldman & Sanger, 2007)

Maron y Kuhns (1960) presentaron un categorizador basado en el modelo *naive* Bayes, que utiliza el supuesto de independencia estadística en la co-ocurrencia de términos en un

mismo documento. En este categorizador, cada documento se representa por el conjunto de términos que constituyen su texto en la que cada término se representa en un grafo, el cual está dirigido a través de un nodo. Para representar las categorías se utiliza un nodo adicional; entre el nodo que representa el término y el que representa una categoría se tiende un arco y se etiqueta cada arco con el coeficiente que representa la probabilidad, de que dicho término pertenezca a una determinada categoría. El documento es asignado a la categoría que obtiene la máxima probabilidad de pertenencia.

Para categorizar documentos, se asigna un valor booleano a cada par <d,c> ε D x C, donde D representa a la colección documental, C representa un grupo de clases, d es un documento que pertenece a D y c es una clase que pertenece a C. El valor Booleano asignado al par <d,c> es 1 si se ha tomado la decisión de categorizar a d en c, 0 en caso contrario. Esta categorización se conoce como dura.

Existe una categorización blanda en donde se asigna un puntaje a cada par<d,c>, permitiendo que eventualmente un documento pueda ser clasificado en más de una clase. El modelo tiene la siguiente forma:

$$p(d) = \sum_{j=1}^{|c|} p(cj)p\ (d\backslash cj)$$

donde d ε D y cj ε C. Para obtener la probabilidad de que d genere a c usamos la regla de Bayes:

$$p(c|d) = \frac{p(c)p(d|c)}{p(d)}$$

Al clasificar un documento, se toma a la clase con mayor probabilidad a posteriori, escenario en el cual p(d) se comporta como una constante por lo que el discriminante se reduce al producto p(c)p(d|c). Las probabilidades p(c) corresponden a probabilidades a-priori, las cuales pueden ser estimadas contando el número de documentos de ejemplo que pertenecen a c.

Las probabilidades p(d|c) pueden ser estimadas usando los términos que componen el contenido de d. Existen varios modelos Bayesianos que hacen diferentes supuestos acerca

de cómo los documentos son formados a partir de los términos que los componen. (Mendoza, Ortiz, & Rojas, 2011)

2.8.3 Modelo vectorial

En esta técnica cada documento se modela como un vector de dimensión n, y se representa usando la siguiente fórmula:

$$Di = (d1, d2, \dots dn)$$

Donde cada palabra del documento se representa en cada posición del vector. El éxito o fracaso de estos algoritmos se basa en la ponderación o peso de los términos. Mediante este modelo se pueden explotar las relaciones geométricas entre dos vectores documento y términos a fin de expresar similitudes y diferencias entre términos. (Guevara, 2011)

La colección de n documentos indexados por m términos se representa a través de una matriz de dimensión *n* x *m*, donde cada elemento *aij* se define por una frecuencia ponderada del término *i* en el documento *j* con el fin de mejorar el rendimiento en la recuperación de la información.

A continuación, se presenta una matriz en la que se representa cada término de la colección en las columnas, y cada fila representa un documento y cada elemento de la matriz es la ocurrencia del término en el documento.

Se observa en el término 1 se encuentra en el documento 1 y 3 pero no en los otros.

	Término 1	Término 2	Término 3
Documento 1	1	0	0
Documento 2	0	0	1
Documento 3	1	1	1
Documento 4	0	1	0

Ilustración 2 Matriz nxm del modelo vectorial (Guevara, 2011)

Cada elemento de la matriz se representa con la siguiente fórmula:

$$aij = lij * gi * dj^{-1}$$

Donde:

- *lij* es el peso local del término i en el documento j,

- *gi* es el peso global del término i en la colección de documentos, y

- *dj* es el factor de normalización para j-ésimo documento.

El peso local se encarga de medir la importancia del término i en el documento j, sólo depende de las frecuencias en el documento. Existen algunas fórmulas para calcular el peso local, entre ellas:

(i) Binaria: el número de veces que aparece la palabra en el documento es irrelevante, debido a que a cada palabra del mismo documento le da la misma importancia. La fórmula para calcular es:

$$x(t) = \begin{cases} 1 \; si \; t > 0 \\ 0 \; si \; t = 0 \end{cases}$$

(ii) Frecuencia del término: esta fórmula va contando los términos en el documento, mientras más veces se presente un término t en el documento j es más probable que sea relevante para el documento. Sin embargo, si una palabra aparece 10 veces no quiere decir que sea más importante que una q aparezca 1 vez.

$$f_{ij}$$

(iii) Frecuencia aumentada de términos normalizados: en esta fórmula se toma en cuenta el número de veces que aparece un término en un documento y la frecuencia con la que aparecen. Su fórmula es:

$$k * x(fij) + (1 - k) * \frac{fij}{max_k \; (f_{kj})}$$

(iv) Logaritmo y logaritmo alternativo: estas fórmulas garantizan minimizar el efecto de la frecuencia, son los más utilizados.

$$\log(f_{ij} + 1)$$

$$x(f_{ij}) * (\log(f_{ij}) + 1)$$

2.8.4 Reglas de Asociación

Las reglas de asociación tratan de buscar posibles relaciones entre elementos, por ejemplo, se buscan relaciones entre los artículos que compra una persona: "Si compra pan, compra leche", esta relación se representa mediante la notación pan →leche. A la parte de la izquierda se le conoce como antecedente y a la de la derecha como consecuente. . (Concha & Alarcón, 2005)

El modelo de Agrawal está representado por la siguiente forma:

$$X \rightarrow Y, \text{ donde: } X, Y \subset I \text{ y } X \cap Y = 0$$

Entendiendo con ello que toda transacción que cumple en X también cumple a Y. (Agrawal & Otros, 1993)

Para encontrar reglas de asociación se utiliza el algoritmo a-priori en el cual se considera cada posible combinación de pares atributo-valor. A cada par atributo-valor se denomina *ítem* y el conjunto de *ítems* se denomina *ítem-sets.* Para realizar este algoritmo se utiliza las siguientes fórmulas:

$$soporte(A \Rightarrow B) = P(A \cap B)$$

$$confianza(A \Rightarrow \mathrm{B}) = p(B|a) = \frac{p(A \cap B)}{p(A)}$$

Cada *ítem-set* debe cumplir con un máximo soporte es por esto que se comienza con *ítem-sets* con un solo ítem. Los *ítem-sets* cuyo valor sea inferior al mínimo establecido se eliminan, a los restantes se le añade un nuevo *ítem* formando *ítem-sets* con tres *ítems*. Se continúa el proceso hasta que el número de *ítem-sets* sea menor en 1 al número de ítems.

Al acabar de formar los *ítem-sets*, se empieza a genera reglas que cumplan con la condición de confianza. Un *ítem-set* puede dar más de una regla de asociación así como puede dar ninguna.

Un ejemplo clásico de reglas de asociación es analizar las compras que realiza una persona, en el siguiente gráfico podemos observar la obtención de estas reglas:

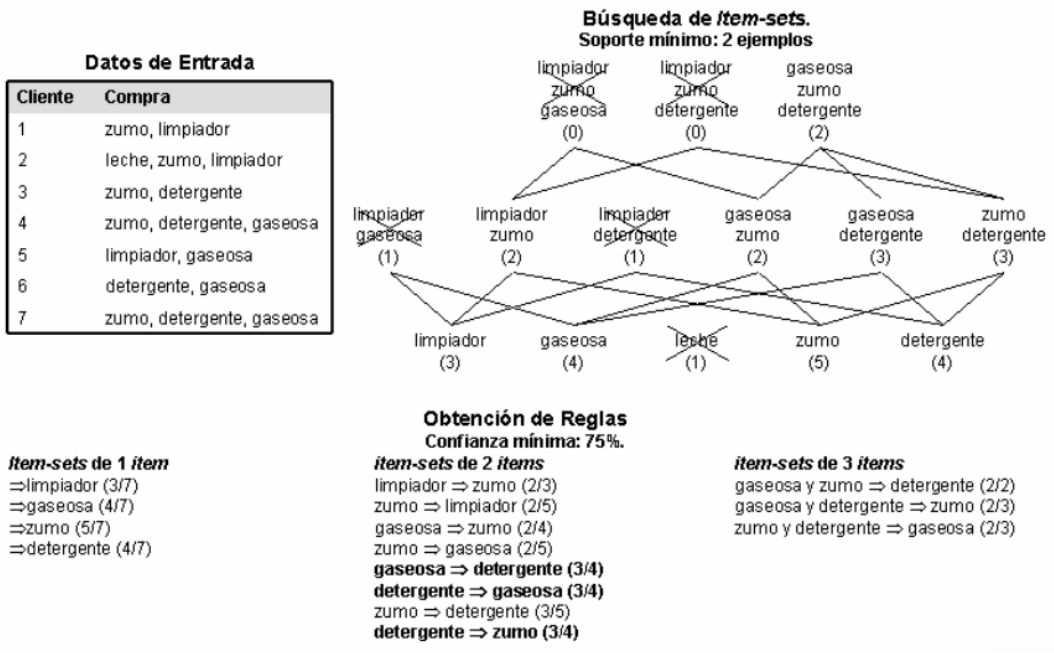


Ilustración 3 Algoritmo a-priori.

Fuente: (Molina López & Otros, 2006)

Se observa el proceso por el cual pasan los datos hasta encontrar las reglas de asociación. En un principio, se extrae la información de los productos que compran los clientes de una empresa. Posteriormente, se forman los *ítem-sets* de 1 solo *ítem*, aquí se toma la frecuencia con la que aparecen los datos en la colección, por ejemplo: el *ítem* zumo aparece 5 veces, como son 7 los clientes analizados su frecuencia será $\frac{5}{7}$. A continuación, se forman los *ítem-sets* de 2 *ítems*, en donde se toma la frecuencia en la que aparecen este par de *ítems*, por ejemplo: el *ítem-set* formado por limpiador y zumo aparecen 2 veces de las 3 en las que limpiador es el antecedente. Después de esto, se toman los *ítems-sets* con mayor frecuencia y se forman *ítem-sets* de 3 *ítems*, aquí se verifica la frecuencia de los mismos. Este proceso se repite hasta que el número de *ítems* que forman los *ítem-sets* sea menor al número de *ítems*. Después de esto, se forman las reglas de asociación:

- Cada vez que un cliente compra gaseosa y zumo, compra detergente.
- Cada vez que un cliente compra gaseosa y detergente, compra zumo.
- **Cada vez que un cliente compra zumo y detergente, compra gaseosa.**

2.9 Análisis de Sentimiento

El Análisis de Sentimiento denominado también minería de opinión se encarga de clasificar los documentos según la polaridad de la opinión que expresa una persona, es decir, determina si una opinión indicada por un usuario tiene un criterio positivo, negativo o neutro. (Cámara & Otros, 2011).

Las opiniones influyen sobre el comportamiento de las personas, así también, sobre la toma de decisiones, por ejemplo, una compañía al conocer la opinión que tienen sus clientes sobre sus productos puede mejorar la calidad de los mismos con el fin de satisfacer la necesidad de su cliente. Un análisis de sentimiento puede ser aplicado en diferentes ámbitos, por ejemplo, conocer la opinión sobre un político, sobre una campaña de marketing, sobre la popularidad de un cantante, entre otros.

Mediante este método se puede obtener resultados que demuestran el margen de aceptación que se tiene frente a los usuarios en un determinado tema. Las opiniones expresadas por los usuarios son datos no estructurados. Un dato no estructurado es aquel que no está almacenado en una base de datos tradicional (estructura relacional, jerárquica o de red). Más bien es aquel dato que se toma de la web, de las cámaras móviles y videos, medios sociales, sensores de ciudades y edificios, etc. Es un dato que proviene de un texto normal, y se caracterizan por la variedad de su origen así como con la rapidez con la que incrementan su volumen.

El auge de los medios sociales involucra usuarios de todas las edades, poco a poco el rango de edad que estaba muy marcado en un principio, se está rompiendo por lo que es un buen medio para extraer la opinión de los usuarios. Debido a esto, aplicar un análisis de sentimiento en los *tweets* obtenidos referentes a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca nos brindará el punto de vista sobre los cuencanos respecto a este tema.

Conclusiones

El soporte fundamental de la minería de texto, y en general de la minería de datos, son los algoritmos aplicados en cada caso. Tomando en cuenta que la investigación se sustenta en el análisis textual, los algoritmos que se identifican son categorización de textos, modelo vectorial, en especial el algoritmo de asociación, ya que este permite encontrar las posibles

relaciones entre elementos. Otro que tiene amplia incidencia es el modelo vectorial, porque este identifica la frecuencia de los términos en un documento con lo que permite determinar cuáles son las palabras que más usa un usuario en un determinado tema. No se podría obtener información veraz si antes de procesar la información, no se pasa por un proceso de calidad de datos, en el cual se limpia los datos antes de usar las técnicas de minería de textos; por ejemplo, se eliminan los signos de puntuación o los conectores del lenguaje como por ejemplo: y, con, etc. Así también, es conveniente eliminar enlaces, números y espacios vacíos.

Capítulo III: Experimentación

Introducción

En este capítulo se describirá el proceso para obtención de los datos de la red social *Twitter*, cuyo contenido sea referente al proyecto Tranvía Cuatro Ríos de Cuenca. Así también, se detallará las sentencias utilizadas para realizar el procesamiento de los mismos mediante técnicas de minería de textos que ofrece el lenguaje de programación *R*. De igual manera se detallarán las sentencias que fueron utilizadas para realizar la depuración de los datos.

Mediante la minería de texto se encontrarán los tópicos más hablados con el fin de categorizar a los usuarios según los datos aportados. Este método permite sustituir o completar el trabajo de las personas independientemente de la cantidad de texto a procesar. Es por esto, que se aplicará a continuación los algoritmos con el fin de encontrar patrones que son extremadamente difíciles de descubrir.

3.1 Obtención de datos

Para la obtención de los *tweets* se utilizará la *API* de *Twitter* y el lenguaje estadístico *R* para la extracción de datos y la generación de gráficas.

Una *API* es un conjunto de comandos, funciones y protocolos informáticos con los cuales los desarrolladores crean programas reutilizando el código de la misma y permitiendo la conexión entre diferentes lenguajes o sistemas operativos. (Álvarez Díaz, 2015)

Como primer paso, se realiza la creación de una aplicación (*App*) que servirá para conectar *R* y *Twitter*. Esta creación se realiza mediante un usuario de *Twitter* en la *API* para desarrolladores.

Se ingresa al sitio web de desarrolladores de *Twitter* en el siguiente enlace:

https://apps.twitter.com/

A continuación se presenta una imagen del sitio web:

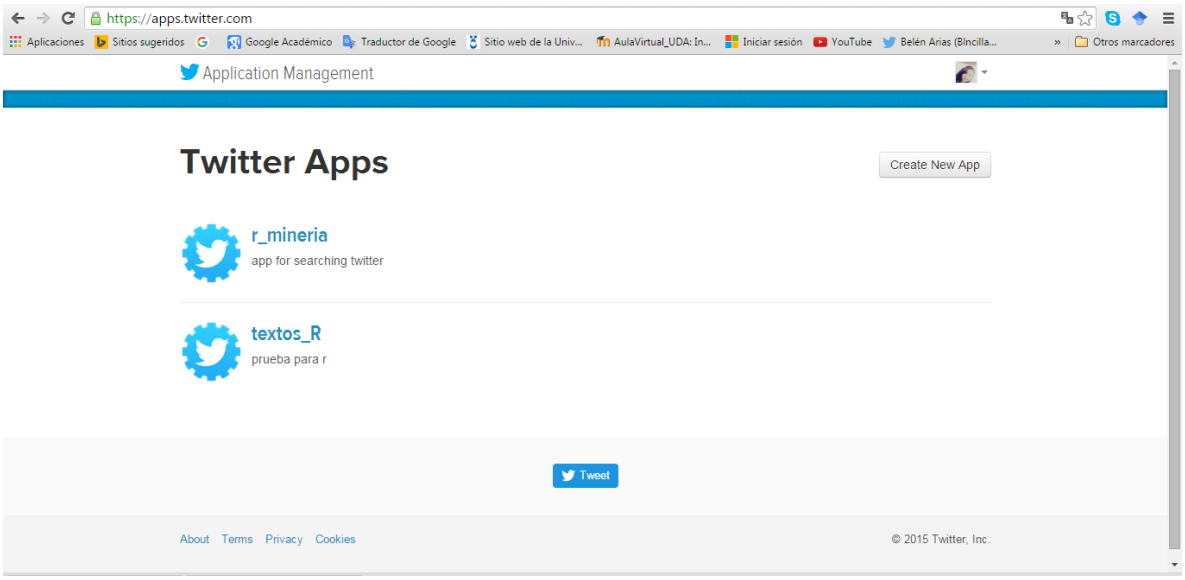


Ilustración 4 Captura de Pantalla Twitter Apps

Se crea una aplicación nueva, en donde se llenan los siguientes campos:

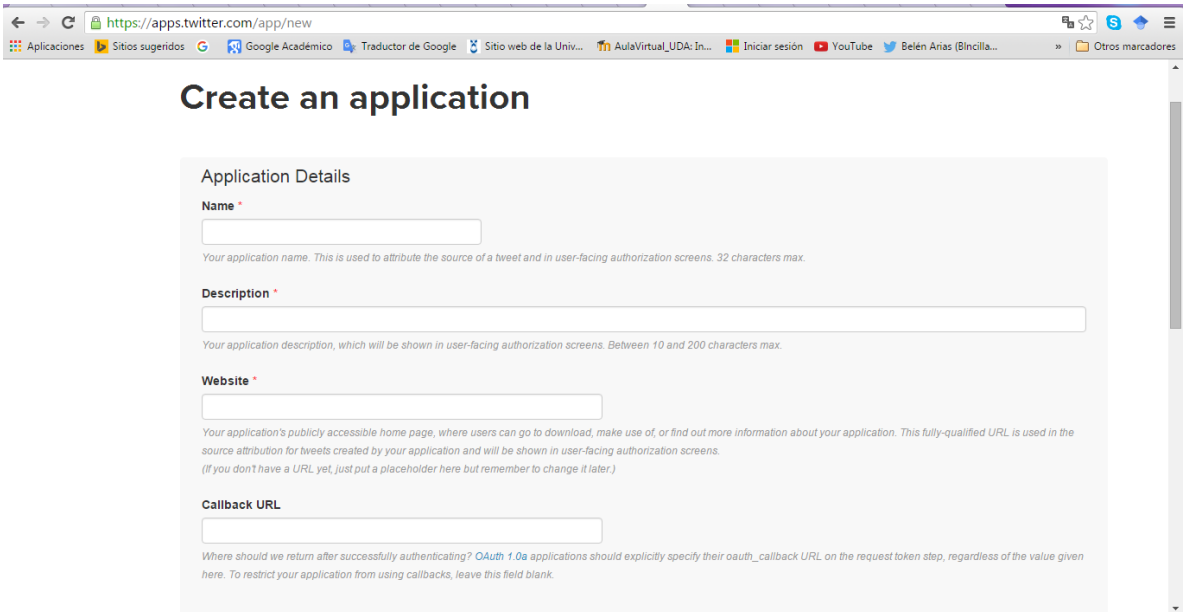


Ilustración 5 Captura de Pantalla para crear una aplicación en Twitter Apps

- *Name*: para el nombre de la aplicación
- *Description*: una breve descripción de la aplicación
- *WebSite*: se debe colocar la *URL* del sitio web que va a usar la *API*, en este caso se utiliza la de pruebas que proporciona *Twitter:* http://test.dev/

- *Callback URL*: en este campo se debe colocar la dirección a la que se debe ir luego de realizar la conexión, en este caso la dejamos en blanco porque solo se debe establecer la conexión no direccionar a un sitio web.

Al completar este formulario, se aceptan los términos y condiciones de la *API* de *Twitter* y se procede a la creación de la aplicación.

Una vez creada la aplicación podemos obtener: *consumer Key*, *consumer secret*, *Access token y Access secret.* Estos datos son necesarios para la conexión con *R* mediante el uso de la librería "`twitteR`".

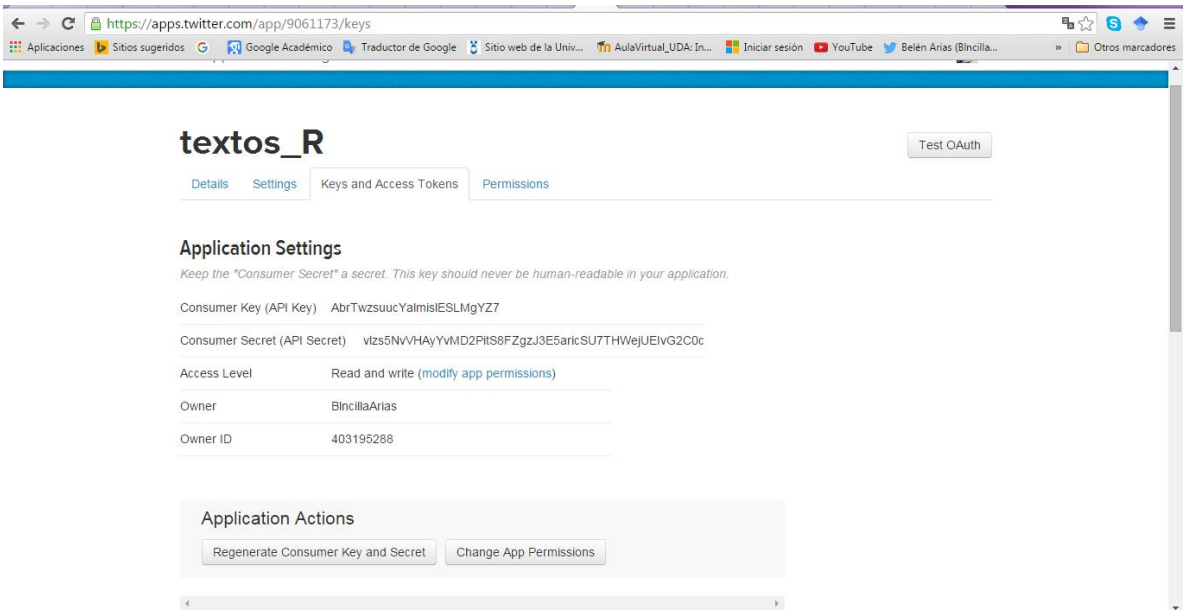


Ilustración 6 Captura de pantalla de configuración de una app de Twitter

Como segundo paso será extraer los datos desde *R* utilizando las siguientes sentencias:

```
library(twitteR)
#iniciar sesion en twitter para extraer información
consumer_key <-'zovysySilMJyfCyWFENVs4Um1'
consumer_secret <-'GQA6Dy9SJjCExELMGmNDG2JRFFTeObqhuFnPX6hiIjbeVfJIZP'
access_token <- '403195288-xWEMPxgQ61DX1JAi7oW3ArHoQ5ivQuhmenSpv2cr'
access_secret <-'j2pIc92QP416QMDlCWgrFOnQ8GVlab8bWaFqkvWfxvBjY'
setup_twitter_oauth(consumer_key, consumer_secret, access_token,
access_secret)
# recolecta tweets con el hashtag #tranviacuenca
```

```
tweets = searchTwitter("#tranviacuenca", n=300)
# transforma la información de los tweets a una tabla
df = twListToDF(tweets)
# obtiene el texto de los tweets
txt = df$text
```

En el código anteriormente descrito, se realiza la conexión entre *R* y *Twitter*, a través de la función *setup_twitter_oauth* se establece las credenciales de *OAuth* para iniciar una sesión de *Twitter* y con esto extraer los *tweets* en tiempo real. Mediante *searchTwitter* obtenemos los tweets que contengan el *hashtag* #tranviacuenca. Para realizar el procesamiento del texto se transforma los datos obtenidos en una colección de datos y finalmente se obtiene solo el texto de los mismos.

3.2 Depuración de datos

Los datos que se obtuvieron de *Twitter* necesitan pasar por un proceso de limpieza para garantizar la calidad de los mismos. Se necesita la librería `gsub` de R para realizar este proceso. Se describen las siguientes sentencias:

```
##### inicio limpieza de datos #####
library(gsub)
# limpia retweets
txtclean = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", txt)
# limpia @usuarios
txtclean = gsub("@\\w+", "", txtclean)
# limpia símbolos de puntuación y caracteres especiales
txtclean = gsub("[[:punct:]]", "", txtclean)
# limpia números
txtclean = gsub("[[:digit:]]", "", txtclean)
# limpia enlaces
txtclean = gsub("http\\w+", "", txtclean)
##### fin limpieza de datos #####
```

Para realizar la depuración de los datos se utilizan expresiones regulares que ofrece el lenguaje *R*. Una expresión regular es un conjunto de caracteres que forman un patrón con el fin de comparar con otra expresión y encontrar si existen coincidencias o no. Mediante el

uso de estas expresiones se irá eliminando los datos que coincidan para obtener solo la información pura y analizarla con las técnicas de minería de textos.

Al ejecutar el código anterior estamos eliminando los nombres de usuario, signos de puntuación, caracteres especiales, números y enlaces con el fin de obtener solo el texto de los *tweets* que sean de relevancia para el procesamiento.

A continuación, se presenta una tabla que contiene las expresiones regulares de la librería `gsubfn` con su respectiva descripción que permite *R*:

Sintáxis	**Descripción**
`\\d`	Números de 0-9
`\\D`	Sin números
`\\s`	Espacios
`\\S`	Sin espacios
`\\w`	Palabras
`\\W`	Sin Palabras
`\\t`	Tabulador
`\\n`	Enter
`^`	Comienzo de una palabra
`$`	Final de una palabra
`\`	Escape special characters, e.g. \\ is "\", \+ is "+"
`|`	Patron alternative
•	Cualquier caracter, excepto enter o fin de línea
`[ab]`	a o b
`[^ab]`	Cualquier caracter excepto a o b
`[0-9]`	Todos los números
`[A-Z]`	Todas las letras en mayúscula de A-Z
`[a-z]`	Todas las letras en minúscula de a-z
`[A-z]`	Todas las letras mayúsculas o minúsculas de a – Z
`i+`	Al menos una vez
`i*`	Cero o más veces

i?	Cero o una vez
i{n}	Ocurre n veces en secuencia
i{n1,n2}	Ocurre n1, n2 veces en secuencia
i{n1,n2}?	Comparación entre cadenas
i{n,}	Ocurre >=n veces
[:alnum:]	Caracteres alfanuméricos
[:alpha:]	Caracteres alfabéticos mayúsculas o minúsculas
[:blank:]	Caracteres en blanco: tabulación o espacio
[:cntrl:]	Caracteres de control
[:digit:]	Números: 0, 1,2 ,3,4,5,6,7,8,9
[:graph:]	Caracteres gráficos
[:lower:]	Caracteres minúsculas
[:print:]	Caracteres imprimibles: alfanuméricos, puntuación y espacio.
[:punct:]	Caracteres de puntuación: ! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [\ ] ^ _ ` { | } ~
[:space:]	Caracteres de espacio: tabulación, enter, tabulación vertical, retorno de carro, espacio
[:upper:]	Caracteres mayúsculas
[:xdigit:]	Dígitos hexadecimales: 0 1 2 3 4 5 6 7 8 9 A B C D E F a b c d e f

Tabla 3 Sintáxis de Expresiones Regulares

Fuente: https://cran.r-project.org/web/packages/gsubfn/gsubfn.pdf

3.3 Procesamiento de datos

Después de la etapa de limpieza, los datos están listos para ser minados. Se utilizarán dos algoritmos encargados de encontrar la frecuencia de las palabras y de la polaridad de las mismas para obtener un análisis de sentimiento.

Modelo vectorial

Se pretende encontrar las palabras con mayor frecuencia y representarlas en una nube de palabras, las librerías necesarias son `tm` y `wordcloud`. Para lograr esto se usaron las siguientes sentencias:

```
Library(tm)
Library(wordcloud)
# construye un corpus
corpus = Corpus(VectorSource(txtclean))
# convierte a minúsculas
corpus = tm_map(corpus, tolower)
# remueve palabras vacías (stopwords) en español
corpus = tm_map(corpus, removeWords, c(stopwords("spanish"),
"cuenca","tranvía","tranvia","tranviaCuenca"))
corpus = tm_map(corpus, removeWords, sw)
# limpia espacios en blanco extras
corpus = tm_map(corpus, stripWhitespace)
corpus <- tm_map(corpus, PlainTextDocument)
# crea una matriz de términos
tdm <- TermDocumentMatrix(corpus)
# convierte a una matriz
m = as.matrix(tdm)
# conteo de palabras en orden decreciente
wf <- sort(rowSums(m),decreasing=TRUE)
# crea un data frame con las palabras y sus frecuencias
dm <- data.frame(word = names(wf), freq=wf)
wordcloud(dm$word, dm$freq, random.order=FALSE,
colors=brewer.pal(8, "Dark"))
```

En el código anterior se utiliza el algoritmo de modelo vectorial ya que se está representando a la colección de datos extraída de *Twitter* como un vector de dimensión n y se está analizando la frecuencia de las palabras en la colección de datos. Estos resultados son representados de forma gráfica mediante una nube de palabras.

Análisis de Sentimiento

Así también, se realiza un análisis de sentimiento con el fin de determinar la polaridad de los *tweets* emitidos por los usuarios en cuanto a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca. Adicionalmente, se usan las librerías `RCurl` y `stringr`. Para esto se utilizó el siguiente código:

```
library(RCurl)

library(stringr)

#se obtiene solo el texto de los tweets

texto = laply(tweets, function(t) t$getText() )

#cargamos en memoria el archivo con las palabras positivas

positive <- scan("D:/Unidad de Titulación/sentiment/positive-
words.txt", what='character',comment.char=';')

#cargamos en memoria el archivo con las palabras negativas

negative <- scan("D:/Unidad de Titulación/sentiment/negative-
words.txt", what='character',comment.char=';')

#se aplica la función encargada de examinar la polaridad de los
tweets

resultado <- score.sentiment(texto, positive, negative)

resultado$score

#se grafica en un histograma el resultado obtenido previamente

hist(resultado$score, col='magenta',main='Análisis de
Sentimiento',ylab='Frecuencia',xlab='Polaridad')
```

Para realizar el análisis de sentimiento se implementa una función encargada de encontrar la polaridad de los datos extraídos, en esta etapa usamos la librería `plry`. La función se encuentra descrita a continuación:

```
score.sentiment = function(sentences, pos.words, neg.words,
.progress='none'){

 library(plyr)

  library(stringr)

  scores = laply(sentences, function(sentence, pos.words,
neg.words) {

  #se realiza depuración de datos con expresiones regulares

  #se eliminan signos de puntuación

   sentence = gsub('[[:punct:]]', '', sentence)

#se eliminan signos de control

 sentence = gsub('[[:cntrl:]]', '', sentence)

    sentence = gsub('\\d+', '', sentence)

    # se convierte en minúsculas

    sentence = tolower(sentence)
```

```
    # se divide el texto por palabras
    word.list = str_split(sentence, '\\s+')
    words = unlist(word.list)
    # comparar nuestras palabras a los diccionarios de términos
positivos y negativos
    pos.matches = match(words, pos.words)
    neg.matches = match(words, neg.words)
    # match() devuelve la posición del término que coincide con
la palabra del diccionario
    # devuelve verdadero o falso
    pos.matches = !is.na(pos.matche s)
    neg.matches = !is.na(neg.matches)
     #si es verdadero será tratado como 1 y si es falso como 0,
esto se irá acumulando
    score = sum(pos.matches) - sum(neg.matches)
     return(score)
  }, pos.words, neg.words, .progress=.progress )
    scores.df = data.frame(score=scores, text=sentences)
  return(scores.df)
}
```

En el código anterior se describe el algoritmo que permite ir comparando las palabras del diccionario de palabras positivas y negativas con los datos extraídos de *Twitter*, con el fin de generar un gráfico que represente la polaridad de la colección de datos. Para representar estos resultados se utiliza un histograma.

Reglas de Asociación

Como un tercer procesamiento de minería de textos se ha utilizado el algoritmo a-priori de las reglas de asociación, este algoritmo se encuentra implementado en la librería `arules` y `arulesViz` de R. El fin de esta minería es encontrar patrones de relaciones entre las palabras emitidas a través de los *tweets* de los usuarios. Se presenta el código necesario para implementar dicho algoritmo:

```
library("arules")
library("arulesViz")
```

```
library("grid")

#carga el archivo de texto de los tweets para ser procesados

trans = read.transactions(file="D:/Unidad de
Titulación/Tranviap1.txt",format="basket", sep=" ",
rm.duplicates = TRUE);

summary(trans)

# extrae las reglas de asociación del archivo

rules <- apriori(trans, parameter=list(support=0.04,
confidence=0.5))

#determina el número de reglas encontradas

length(rules)

#muestra las reglas ordenadas por el factor lift

inspect(head(sort(rules, by="lift")))

#dibuja las reglas en un gráfico x,y según confianza y soporte

plot(rules, measure=c("support" , "confidence") ,
shading="lift", main="Reglas de Asociacion")

#realiza una representación gráfica de la relación entre las
palabras

plot(rules, method="graph",control=list(type="items"),
main="Palabras Asociadas" );
```

En el código descrito se describe el proceso llevado a cabo para encontrar patrones entre las palabras presentes en los *tweets*. En la función *rules* se coloca los parámetros de soporte y confianza, soporte hace referencia a la probabilidad de que un conjunto de palabras estén presentes en la colección de datos y confianza es la probabilidad de que en el consecuente esté presente el antecedente. En el caso de estudio se utilizó un soporte bajo debido a que las palabras poseen gran número de sinónimos por lo que aún no es posible unificar los términos para obtener términos únicos y aumentar la probabilidad. Los resultados son representados mediante un gráfico de coordenadas x e y; mediante una representación gráfica que muestra la relación entre las palabras.

Conclusiones

La minería de textos permite procesar grandes colecciones de datos con el fin de encontrar patrones que permitan la toma de decisiones. Se ha tomado un conjunto de *tweets* provenientes del medio social *Twitter* referentes al proyecto Tranvía Cuatro Ríos de Cuenca

para minarlos textualmente y obtener resultados que muestren la opinión de las personas en cuanto a la implementación de dicho proyecto. Cabe destacar que los datos han sido depurados en un principio para de esta manera trabajar con datos puros y descartando datos irrelevantes. Así, por ejemplo, se ha eliminado signos de puntuación, caracteres especiales, enlaces, nombres de usuarios, palabras que por ende estarían presentes como: tranvía, Cuenca, proyecto. El algoritmo de modelo vectorial ha permitido encontrar la frecuencia de las palabras y se ha representado en una nube de palabras para su posterior análisis.

Capítulo IV: Resultados

Introducción

En este capítulo se presentarán los resultados obtenidos luego de aplicar técnicas de minería de textos en los *tweets* emitidos por los usuarios de *Twitter* referente a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca. Los resultados se presentarán en una nube de palabras, en la cual el tamaño de las palabras es proporcional a la frecuencia de las mismas en la colección de los datos.

Mediante el análisis de sentimiento se determinará si una opinión expresada por un usuario tiene un criterio positivo o negativo facilitándose el análisis de los datos previamente obtenidos.

4.1 Análisis de Resultados

Con el fin de obtener resultados objetivos se ha realizado el procesamiento en determinadas fechas durante el mes de diciembre. A continuación se presenta las diferentes nubes de palabras obtenidas según las diferentes fechas.

Al recolectar los *tweets* desde el 14 de diciembre de 2015 hasta el 18 de diciembre de 2015 se presenta la siguiente nube de palabras:

Ilustración 7: nube de palabras al 18 de diciembre de 2015

Fuente: Material Obtenido de los tweets emitidos por los usuarios con el hashtag #tranviacuenca, transformación en R

La nube de palabras evidencia que existe una opinión negativa en cuanto a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca, debido a que las palabras con mayor frecuencia son malestar y retrasos.

Se recolectan los *tweets* desde el 21 de diciembre de 2015 hasta el 25 de diciembre de 2015 obteniendo la siguiente nube de palabras:

Ilustración 8 nube de palabras al 25 de diciembre de 2015

Fuente: Material Obtenido de los tweets emitidos por los usuarios con el hashtag #tranviacuenca, transformación en R

El sentimiento de negatividad continúa presente y las palabras con mayor frecuencia continúan siendo *malestar* y *retrasos.* Además cabe señalar que la época en la que fueron tomados los datos es una época en la que aumenta la movilización de las personas por el área comercial de la ciudad y por donde se realizan las obras de construcción del proyecto Tranvía Cuatro Ríos de Cuenca.

A partir del 28 de diciembre de 2015 hasta el 01 de enero de 2015 se presenta la siguiente nube de palabras:

Ilustración 9: nube de palabras al 01 de enero de 2016

Fuente: Material Obtenido de los tweets emitidos por los usuarios con el hashtag #tranviacuenca, transformación en R

En esta nube de palabras la palabra con mayor frecuencia es *avance*. Se aprecia que los tópicos más relevantes sobre la implementación del proyecto Tranvía Cuatro Ríos de Cuenca se refiere al proceso de avance que ha tenido el proyecto hasta la fecha, posiblemente a que se rindió cuentas del mismo por parte de la alcaldía de la ciudad.

Se realiza una cuarta nube de palabras desde el 04 de enero de 2016 hasta el 08 de enero de 2015:

Ilustración 10: nube de palabras al 08 de enero de 2016

Fuente: Material Obtenido de los tweets emitidos por los usuarios con el hashtag #tranviacuenca, transformación en R

En esta nube de palabras las palabras con mayor frecuencia son: *tráfico, vehicular, ciudad.* El sentimiento de negatividad está presente nuevamente y se evidencia que los tópicos tratados en este tiempo se relacionan con los problemas que ha generado la implementación del proyecto Tranvía Cuatro Ríos de Cuenca.

Así también, se presenta un histograma que grafica la polaridad de los *tweets* emitidos durante las 4 semanas de análisis:

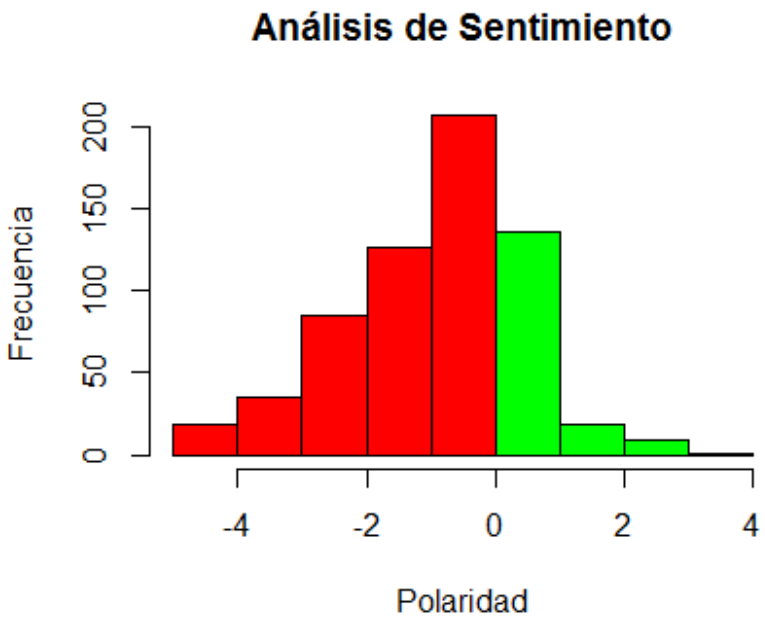


Ilustración 11 histograma al 08 de enero de 2016

Fuente: Material Obtenido de los tweets emitidos por los usuarios con el hashtag #tranviacuenca, transformación en R

El sentimiento de negatividad es mayor al de positivismo en los *tweets* recolectados, por lo que se puede determinar que la implementación del proyecto Tranvía Cuatro Ríos de Cuenca no goza de aceptación.

En un segundo procesamiento, se aplicó el algoritmo a-priori con el fin de encontrar reglas de asociación en la colección de *tweets* recolectados. A continuación se presenta los resultados en un gráfico de coordenadas:

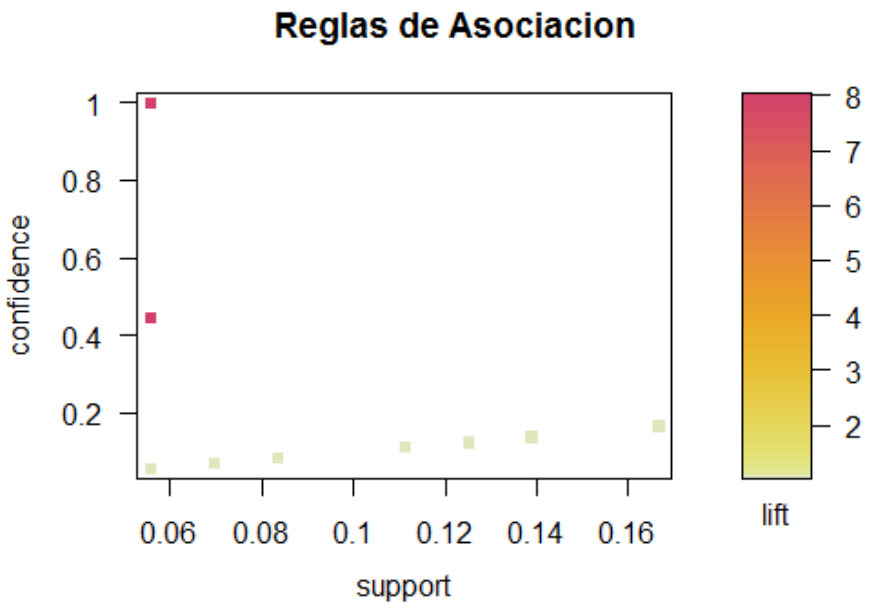


Ilustración 12 Gráfico de validación de variables de soporte y confianza al 2 de Febrero de 2016

Fuente: Material Obtenido de los tweets emitidos por los usuarios con el hashtag #tranviacuenca, transformación en R

El gráfico anterior describe los niveles de soporte y confianza que se han recolectado de los *tweets* emitidos por los usuarios. El nivel de soporte representa el porcentaje de que un conjunto de palabras estén presentes en la colección de datos. El nivel de confianza representa la probabilidad de que en el consecuente esté presente el antecedente.

A través de este gráfico se establece los niveles óptimos para encontrar las palabras asociadas. Se observa que los niveles de soporte abarcados no sobrepasan el 20% de la confianza. Así también, cuando la confianza es alta, no se supera el 6% de soporte. Este comportamiento se debe a que los datos minados son datos no estructurados, en los cuales existen palabras cuyo significado es el mismo, pero cuya representación es diferente, este es el caso de los sinónimos. Por ejemplo:

Un usuario puede emitir un *tweet:* "El estado del proyecto se debe a una mala planificación".

Otro usuario puede publicar: "El proyecto tuvo una pésima planificación". Las palabras mala y pésima están refiriéndose al mismo criterio de negatividad pero se escriben de forma diferente, ocasionando que el comportamiento de soporte y confianza sea muy bajo, esto se observa en el grafico anterior, ya que los puntos están dispersos, sin embargo mantienen cercanía con el límite del eje x e y, debido a la variedad de las palabras. A diferencia de una base de datos relacional, que contiene una colección de datos estructurados donde el nivel de soporte y confianza responden a datos codificados en todas las instancias, es decir, el comportamiento se diferencia sustancialmente al que se presenta en el minado de texto, que expresa una diversidad de términos que pese a que son diferentes, apuntan a un sentimiento similar. Por ejemplo, en una base de datos, si una persona compra pan, el nombre y el código se mantiene para una o varias compras, lo que beneficia al comportamiento del algoritmo de asociación, en el minado de texto no sucede aquello.

Los patrones encontrados se representan mediante el siguiente gráfico:

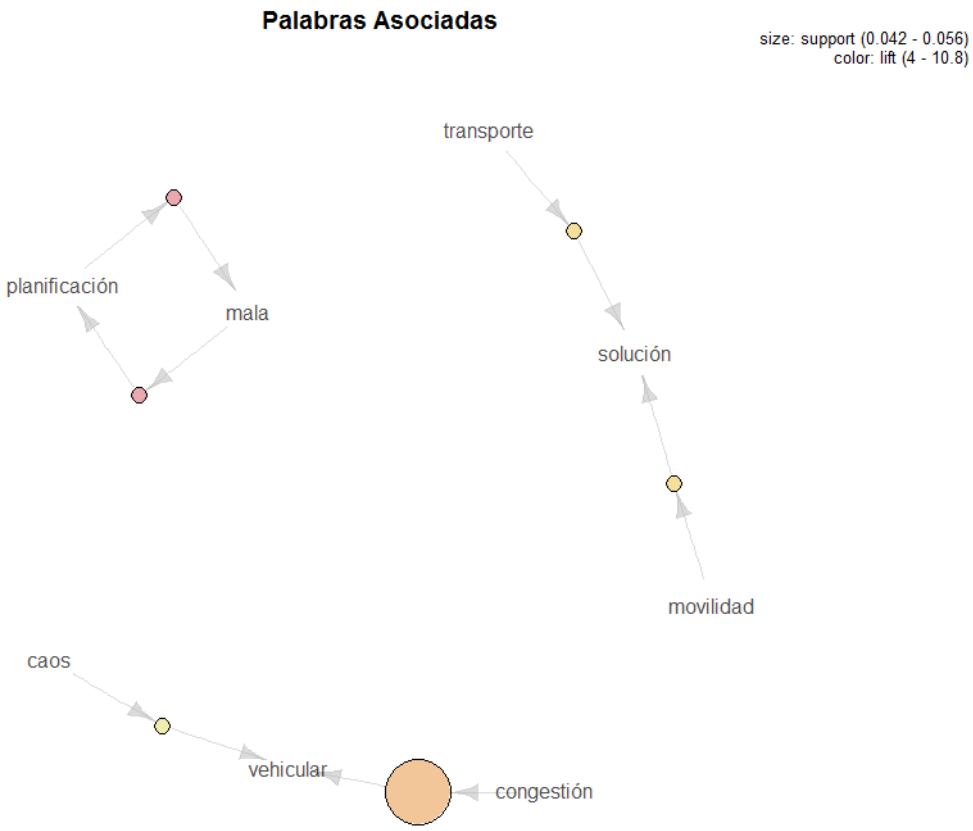


Ilustración 13 Reglas de Asociación al 2 de Febrero de 2016

Fuente: Material Obtenido de los tweets emitidos por los usuarios con el hashtag #tranviacuenca, transformación en R

En el gráfico anterior se observa que cada vez que en un *tweet* se emita la palabra "congestión", esta va acompañada de la palabra "vehicular" generándose el patrón (congestión, vehicular). Así también, se presenta el patrón (caos, vehicular), por lo tanto si un usuario usa la palabra "caos" va acompañado de la palabra "vehicular".

Otros dos patrones encontrados son: (planificación, mala) y (mala, planificación), se determina que cada vez que se emita la palabra "planificación" se presenta la palabra "mala" y viceversa.

Además, al presentarse en un *tweet* la palabra "transporte" se presenta la palabra "solución", este es el patrón (transporte, solución). El último patrón encontrado es:

(movilidad, solución) en donde la palabra “movilidad” va acompañada de la palabra “solución”.

A través de estos patrones encontrados en la colección de datos se determina que la posición de los usuarios de *Twitter* en cuanto a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca es negativa, ya que las palabras asociadas están vinculadas a ese sentimiento.

Conclusiones

El modelo vectorial es un algoritmo de minería de textos que se encarga de encontrar la frecuencia de las palabras en una colección de datos, por ello se aplicó este algoritmo en el caso de estudio. Se procesaron los datos y se obtuvo nubes de palabras que representan las palabras más utilizadas por los usuarios al referirse a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca. Se evidencia que las palabras más connotadas son “malestar”, “retrasos”, “congestión”, “vehicular”, “caos” por lo que se interpreta un criterio negativo por parte de los usuarios. Así también, se logró representar a través de un histograma la polaridad de los *tweets* observando que la negatividad es mayor en los usuarios.

Por otro lado, se empleó el algoritmo a-priori encargado de encontrar patrones en la colección de datos. Para esto se emplean los factores de soporte y confianza; en el caso de estudio estos factores son muy bajos debido a que los usuarios podrían utilizar sinónimos para dar su criterio, por lo que las palabras no podrán ser tomadas como un ítem codificado; a diferencia de una base de datos transaccional, en la que cada ítem tiene un nombre único. Como resultado obtenemos un gráfico que representa el nivel de soporte y confianza que existe en la colección de datos y un gráfico que evidencia las relaciones que existe entre las palabras: (caos, vehicular), (congestión, vehicular), (planificación, mala), (mala, planificación), (transporte, solución) y (transporte, movilidad). Los patrones encontrados están orientados a la negatividad que tiene la gente en cuanto a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca.

Conclusiones

Esta investigación se ha basado en los algoritmos de la minería textual que permitan encontrar palabras frecuentes, la polaridad de las mismas y la relación que existe entre unas y otras palabras. Se ha aplicado a datos no estructurados, los cuales no pertenecen a una estructura relacional, jerárquica o de red sino por el contrario provienen del lenguaje natural de las personas. A través del lenguaje *R* se ha realizado la minería a los datos mediante la utilización de las siguientes librerías: `twitteR` que se encarga de la extracción de los *tweets*, `gsub` para la depuración de los datos, `tm` y `wordcloud` para realizar el algoritmo de modelo vectorial, `RCurl`, `stringr y plry` para determinar la polaridad de la colección de datos y las librerías `arules y arulesViz` para el algoritmo de asociación.

Para la aplicación específica de la investigación, se tomó como caso de estudio el proyecto de Tranvía Cuatro Ríos de Cuenca, en concreto los datos no estructurados provenientes de las opiniones que las personas emitían a través del medio social *Twitter*. Mediante el modelo vectorial se han encontrado los términos más relevantes de las opiniones de los usuarios en la implementación del proyecto Tranvía Cuatro Ríos de Cuenca. Cabe señalar que, al momento del análisis, el proyecto, planificado de 2014 hasta 2016, se encuentra en fase de construcción.

Antes de realizar una minería en la colección de datos fue necesario la depuración de los mismos mediante un proceso de limpieza de datos; por ejemplo, se eliminan los signos de puntuación o los conectores del lenguaje como por ejemplo: “y”, “con”, entre otros. Así también, es conveniente eliminar enlaces, números y espacios vacíos. Adicionalmente, se ha limpiado los datos irrelevantes como las palabras: “tranvía”, “Cuenca”, “proyecto” que estuvieron presentes sin marcar una relevancia en la investigación.

Una vez depurados los datos, se los ha analizado mediante el modelo vectorial. Los resultados de este análisis han sido representados en el gráfico “nube de palabras”, donde algunos términos observados son: “malestar”, “tráfico”, “vehicular”, entre otras. Se evidencia un sentido negativo por parte de los usuarios respecto a la implementación del proyecto Tranvía Cuatro Ríos de Cuenca. Así también, se representó en un histograma la polaridad de los *tweets* obteniendo mayor cantidad de *tweets* negativos que positivos.

Para encontrar la asociación entre las palabras emitidas en los *tweets* se ha utilizado el algoritmo de asociación a-priori, donde se encontraron patrones que demuestran la postura negativa por parte de los usuarios respecto al proyecto Tranvía Cuatro Ríos de Cuenca. Se han encontrado algunos patrones, tales como: (planificación, mala) y (mala, planificación) donde cada vez que un usuario utilice la palabra "mala" va acompañada de la palabra "planificación", y viceversa. Este algoritmo utiliza los factores de: soporte y confianza, en el caso de estudio estos factores son muy bajos ya que los datos minados son no estructurados, están basados en un lenguaje natural, el cual es difícil codificar como sucede en una base de datos transaccional. Sin embargo, es posible categorizar los términos identificando los valores discretos más frecuentes, con el fin de aumentar los valores de soporte y confianza. Para un posterior estudio se recomienda realizar este proceso previamente a la aplicación del algoritmo, mejorando así la obtención de resultados y la toma de decisiones.

La minería de texto es un área poco explotada en nuestro entorno, a pesar del gran aporte y utilidad que tiene. Puede aplicarse en diferentes áreas, como por ejemplo: medir la popularidad y grado de aceptación de un producto, un cantante, un político o en el caso de estudio de un proyecto. Mediante esta técnica se puede obtener información valiosa para la toma de decisiones e incluso información que a simple vista estaba oculta. Por ejemplo: al analizar las transacciones de una base de datos se puede obtener patrones de compra de los consumidores y a través de estos patrones generar promociones en los productos aumentando las ventas y por ende las ganancias, definitivamente son técnicas y métodos que aportan en la toma de decisiones de las personas que tiene a cargo la responsabilidad de promover el funcionamiento de las organizaciones e instituciones y el bienestar de la sociedad.

Referencias

Agrawal, R., Imielinski, T., & Swami, A. (1993). Minning association rules between sets of items in large databases. *ACM SIGMOD Record*, Vol. 22, No:2, pp. 207-216.

Álvarez Díaz, E. (2015). Publicación de Cartografía en Internet.

Bailón-Moreno, R. (2013). Bailón-Moreno, R. (2013). Ingeniería del conocimiento y vigilancia tecnológica aplicada a la investigación en el campo de los tensioactivos. Desarrollo de un modelo ciencimétrico unificado.

Bianchi Gallo, B., & Valverde Corrado, M. C. (2009). Un caso de estudio en Calidad de Datos para Ingeniería de Software Empírica.

Cámara, E. M., Valdivia, M. M., & Urena, L. A. (2011). Análisis de Sentimientos. *IV Jornadas de Tratamiento de la Información Multilingue y Multimodal 7 y 8 de abril de 2011*, 61.

Caparrrini, F. S. (16 de Julio de 2015). Obtenido de http://www.cs.us.es/~fsancho/?e=75

Castro, M., & Lizasoain, L. (2012). Las técnicas de modelización estadística en la investigación educativa: minería de datos, modelos de ecuaciones estructurales y modelos jerárquicos linealeales. . *Revista Española de pedagogía*, 131-148.

Concha, U. R., & Alarcón, L. (2005). Minería de Uso Web para predicción de usuarios en la universidad. *Revista de investigación de Sistemas e Informátic*, 2 (3), 7-13.

cran.r-project.org. (29 de Diciembre de 2015). Obtenido de cran.r-project.org: https://cran.r-project.org/web/packages/gsubfn/gsubfn.pdf

Del Alcanzar Ponce, J. P. (21 de Noviembre de 2015). *Formación Gerencial*. Obtenido de Formación Gerencial: http://blog.formaciongerencial.com/2014/05/16/ranking-redes-sociales-ecuador-mayo-2014/

Domínguez, D. C. (2010). Las Redes Sociales.Tipología, uso y consumo de las redes 2.0 en la sociedad digital actual. . *Documentación de las Ciencias de la Información*, 33, 45-68.

El Comercio. (15 de abril de 2015). *El comercio.com.* Obtenido de El Comercio.com: http://www.elcomercio.com/tendencias/facebook-redessociales-ecuador-inec-usuarios.html

Fainholc, B. (2011). Un análisis contermporáneo del Twitter. *Revista de Educación a Distancia*, 26, 1-12.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. . *Cambridge University Press*.

Fernández, C. B. (2012). Twitter y la Ciberpolítica. *Disertaciones: Anuario electrónico de estudios en Comunicación Social*, 5(1), 9-24.

Garre, M., Cuadrado, J., Silicia, M., Rodriguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Innovación, Calidad e Ingeniería del Software*, 3(1), 6-22.

Guevara, R. (2011). Minería de textos en la red social Twitter.

Hernandez, J. A. (2013). Generación, Tratamiento y Análisis de Información en las Organizaciones. *Universidad Autónoma del Estado de Morelos*, 174.

Jimenez, A., Berzal, F., & Cubero, J. C. (2010). POTMiner: mining order, unordered, and partially-ordered trees. . *Knowledge and information systems*, 199-224.

Llewellyn, B. (9 de Diciembre de 2013). *BlogSpot.* Obtenido de BlogSpot: http://www.profitpt.com/wp-content/uploads/2010/11/New-Picture-6.png

Lobos, C., & Bartolome, L. (2012). Metodología de búsqueda de sub-comunidades mediante análisis de redes socialeas y minería de datos.

Mansilla, P. S., Costaguta, R., & Missio, D. (2014). Aplicación de Algoritmos de Clasificaación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos. . *Revista Iberoamericana de Inteligencia Artificial*, 17 (53), 57-67.

Martinez Luna, G. L. (2011). Minería de datos: Cómo hallar una aguja en un pajar. *Ingenierías*, Vol. XIV, No. 53.

Mendoza, M., Ortiz, I., & Rojas, V. (2011). Categorización de texto en bases documentales a partir de modelos computacionales livianos. *Revista Signos*, 44(77), 251-274.

Molina López, J. M., & García Herrero, J. (2006). Aplicaciones prácticas utilizando Microsoft Excel y Weka. *Técnicas de Análisis de Datos*.

Ocerín, J. M. (2011). La minería de datos: Análisis de bases de datos en la empresa.

Pérez López, C., & Santín González, D. (2007). *Minería de Datos. Técnicas y Herramientas.* Madrid: Thomson Ediciones Paraninfo S.A.

Ponce, G. (27 de noviembre de 2012). *WordPress*. Obtenido de WordPress: https://gerardoponce.wordpress.com/2012/11/27/que-es-un-bot-en-las-redes-sociales/

Rodriguez Aldape, F. (2013). Cuantificación del interés de un usuario en un tema mediante minería de texto y análisis de sentimiento. *Doctoral dissertation. Universidad Autónoma de Nuevo León*.

Sánchez Arteaga, J. M. (2011). Estudio y Análisis del uso de las redes sociales en la ciudad de Cuenca y elaboración de un manual de buenas prácticas de usuario. *Doctoral dissertation*.

Sánchez, D., Miranda, M., & Cerda, L. (2004). Reglas de asociación aplicadas a la detección de fraude con tarjetas de crédito. *In Anctas del XII Congreso Español sobre Tecnologías y Lógica Fuzzy*, pp. 15-17.

Sharma, K., Shrivastava, G., & Kumar, V. (2011). Web Mining: Today and Tomorrow. *Electronics Computer Technology (ICECT)*, Vol. 1, pp. 399-403.

Ureña, A., Ferrari, A., Blanco, D., & Valdecasa, E. (2011). Las Redes Sociales en Internet. *Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información*.

Anexos

Los anexos que se incluyen en este documento son los siguientes:

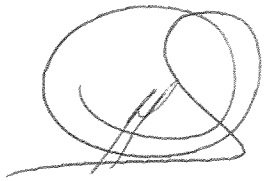
- Anexo 1: Aprobación de protocolo del Trabajo de Titulación
- Anexo 2: Convocatoria a la sustentación del Protocolo del Trabajo de Titulación
- Anexo 3: Acta Sustentación de Protocolo/Denuncia del Trabajo de Titulación
- Anexo 4: Rúbrica para la evaluación del protocolo del Trabajo de Titulación
- Anexo 5: Solicitud de aprobación del protocolo del Trabajo de Titulación
- Anexo 6: Diseño del Trabajo de Titulación

Doctora Jenny Ríos Coello, Secretaria de la Facultad de Ciencias de la Administración de la Universidad del Azuay,

C E R T I F I C A:

Que, el Consejo de Facultad en sesión del 05 de noviembre de 2015, conoció la petición del (los) estudiante(s) **María Belén Arias** con código(s) **60460**, registrado(s) en la Unidad de Titulación Especial, quien(es) denuncia(n) su trabajo de titulación denominado: **"MINERIA DE TEXTOS EN MEDIOS SOCIALES"**en la modalidad: Proyecto de investigación y presentado como requisito previo a la obtención del título de Ingniera de Sistemas y Telemática .- El Consejo de Facultad acoge el informe de la Junta Académica y aprueba la denuncia. Designa como Director(a) a Ing. Marcos Orellana Cordero y como miembro del Tribunal Examinador a Ing. Francisco Salgado Arteaga. De conformidad con el cronograma de la Unidad de Titulación el (los) peticionario(s) debe presentar su trabajo de titulación hasta el 11 de marzo de 2016.

Cuenca, 06 de noviembre de 2015



Dra. Jenny Ríos Coello
Secretaria de la Facultad de Ciencias de la Administración

Oficio Nro. 157-2015-DIST-UDA

Cuenca, 28 de Octubre de 2015

Señor Ingeniero
Xavier Ortega Vázquez
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMNISTRACIÓN
Presente.-

De mis consideraciones:

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 28 de octubre del 2015, revisó la documentación del proyecto de tesis denominado "Minería de textos en medios sociales", presentado por la estudiante María Belén Arias, estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por el Ing. Marcos Orellana, previo a la obtención del título de Ingeniera de Sistemas y Telemática.

La Junta considera que la documentación cumple con las normas legales y reglamentarias de la Universidad y de la Facultad de Ciencias de la Administración y avala la aprobación por parte del tribunal designado, así por su digno intermedio, el conocimiento y aprobación por parte del Consejo de Facultad.

Atentamente,

Ing. Marcos Orellana Cordero
Director Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

Catalina Astudillo

Oficio Nro. 156-2015-DIST-UDA

Cuenca, 28 de Octubre de 2015

Señor Ingeniero
Xavier Ortega Vázquez
DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMNISTRACIÓN
Presente.-

De nuestras consideraciones:

La Junta Académica de la Escuela de Ingeniería de Sistemas y Telemática, reunida el día 28 de octubre del 2015, recibió el proyecto de tesis titulado "Minería de textos en medios sociales", presentado por la estudiante María Belén Arias, estudiante de la Escuela de Ingeniería de Sistemas y Telemática, y revisado por el Ing. Marcos Orellana, previo a la obtención del título de Ingeniera de Sistemas y Telemática.

Por lo expuesto, y de conformidad con el Reglamento de Graduación de la Facultad, recomienda como director y responsable de aplicar cualquier modificación al diseño del trabajo de graduación posterior a al Ing. Marcos Orellana y como miembro del Tribunal a Francisco Salgado Ph.D.

Atentamente,

Ing. Marcos Orellana Cordero
Director Escuela de Ingeniería de Sistemas y Telemática
Universidad del Azuay

Catalina Astudillo

CONVOCATORIA

Por disposición de la Junta Académica de Ingeniería de Sistemas y Telemática, se convoca a los Miembros del Tribunal Examinador, a la sustentación del Protocolo del Trabajo de Titulación: "Minería de texto en medios sociales. Caso de estudio de proyecto Tranvía de Cuenca" presentado por la estudiante **María Belén Arias** con código 60460 previa a la obtención del grado de Ingeniero de Sistemas y Telemática, para el día **MIERCOLES 28 DE OCTUBRE DE 2015 A LAS 08H30.**

Cuenca, 27 de abril de 2015

Dra. Jenny Ríos Coello
Secretaria de la Facultad

Ing. Marcos Orellana Cordero

Ing. Francisco Salgado Arteaga

UNIVERSIDAD DEL AZUAY

ACTA
SUSTENTACIÓN DE PROTOCOLO/DENUNCIA DEL TRABAJO DE TITULACIÓN

1.1 **Nombre del estudiante:** María Belén Arias Zhañay
1.2 Código 60460
1.3 **Director sugerido**: Ing. Marcos Orellana Cordero
1.4 **Codirector (opcional)**: _____
1.5 **Tribunal:** Ing. Francisco Salgado ArteaGA
1.6 **Título propuesto:** : "Minería de texto en medios sociales. Caso de estudio de proyecto Tranvía de Cuenca"
1.7 **Resolución**:

1.7.1 Aceptado sin modificaciones ✓

1.7.2 Aceptado con las siguientes modificaciones:

_____

_____

_____

- Responsable de dar seguimiento a las modificaciones: Ing. Marcos Orellana Cordero

1.7.3 No aceptado
- Justificación:

_____
_____
_____

Tribunal

Ing. Marcos Orellana Cordero

Ing. Francisco Salgado Arteaga

Srta. María Belén Arias Zhañay

Dra. Jenny Ríos Coello
Secretario de Facultad

Fecha de sustentación: Miércoles 28 de octubre de 2015 al as 08h30

UNIVERSIDAD DEL AZUAY

RÚBRICA PARA LA EVALUACIÓN DEL PROTOCOLO DE TRABAJO DE TITULACIÓN

1.1 Nombre del estudiante: María Belén Arias Zhañay
1.1.1 **Código** 60460
1.2 Director sugerido: Ing. Marcos Orellana Cordero
1.3 Codirector (opcional):.
1.4 Título propuesto: "Minería de texto en medios sociales. Caso de estudio de proyecto Tranvía de Cuenca"
1.5 Revisores (tribunal): Ing. Francisco Salgado Arteaga
1.6 Recomendaciones generales de la revisión:

	Cumple totalmente	Cumple parcialmente	No cumple	Observaciones (*)
Línea de investigación				
1. ¿El contenido se enmarca en la línea de investigación seleccionada?	✓			
Título Propuesto				
2. ¿Es informativo?	✓			
3. ¿Es conciso?	✓			
Estado del arte				
4. ¿Identifica claramente el contexto histórico, científico, global y regional del tema del trabajo?	✓			
5. ¿Describe la teoría en la que se enmarca el trabajo	✓			
6. ¿Describe los trabajos relacionados más relevantes?	✓			
7. ¿Utiliza citas bibliográficas?	✓			
Problemática y/o pregunta de investigación				
8. ¿Presenta una descripción precisa y clara?	✓			
9. ¿Tiene relevancia profesional y social?	✓			
Hipótesis (opcional)				
10.¿Se expresa de forma clara?				
11.¿Es factible de verificación?				
Objetivo general				
12.¿Concuerda con el problema formulado?	✓			
13.¿Se encuentra redactado en tiempo verbal infinitivo?	✓			
Objetivos específicos				

14.¿Concuerdan con el objetivo general?	✓			
15.¿Son comprobables cualitativa o cuantitativamente?	✓			
Metodología				
16.¿Se encuentran disponibles los datos y materiales mencionados?	✓			
17.¿Las actividades se presentan siguiendo una secuencia lógica?	✓			
18.¿Las actividades permitirán la consecución de los objetivos específicos planteados?	✓			
19.¿Los datos, materiales y actividades mencionadas son adecuados para resolver el problema formulado?	✓			
Resultados esperados				
20.¿Son relevantes para resolver o contribuir con el problema formulado?	✓			
21.¿Concuerdan con los objetivos específicos?	✓			
22.¿Se detalla la forma de presentación de los resultados?	✓			
23.¿Los resultados esperados son consecuencia, en todos los casos, de las actividades mencionadas?	✓			
Supuestos y riesgos				
24.¿Se mencionan los supuestos y riesgos más relevantes?	✓			
25.¿Es conveniente llevar a cabo el trabajo dado los supuestos y riesgos mencionados?	✓			
Presupuesto				
26.¿El presupuesto es razonable?	✓			
27.¿Se consideran los rubros más relevantes?	✓			
Cronograma				
28.¿Los plazos para las actividades son realistas?	✓			
Referencias				
29.¿Se siguen las recomendaciones de normas internacionales para citar?	✓			
Expresión escrita				
30.¿La redacción es clara y fácilmente comprensible?	✓			
31.¿El texto se encuentra libre de faltas ortográficas?	✓			

(*) Breve justificación, explicación o recomendación.

- Opcional cuando cumple totalmente,
- Obligatorio cuando cumple parcialmente y NO cumple.

.....

.....

.....

Ing. Marcos Orellana Cordero

Ing. Francisco Salgado Arteaga

UNIVERSIDAD DEL AZUAY

Cuenca, 28 de Octubre del 2015

Señor Ingeniero

Xavier Ortega Vázquez

DECANO DE LA FACULTAD DE CIENCIAS DE LA ADMINNISTRACIÓN

Presente.-

De mis consideraciones:

Yo, María Belén Arias Zhañay con código 60460, solicito a usted que se me conceda la aprobación de mi tema de tesis "Minería de Texto en Medios Sociales. Caso de estudio del proyecto Tranvía de Cuenca", previo a la obtención del título de Ingeniero de Sistemas y Telemática.

Por la atención que preste a mi solicitud, mis más sinceros agradecimientos.

Atentamente,

Belén Arias

María Belén Arias Zhañay

60460

C.I. 0104508098

Edición autorizada de 30.000 ejemplares
Del 678.501 al 708.500 Nº 0698191

UNIVERSIDAD DEL AZUAY

Universidad del Azuay

Facultad de Ciencias de la Administración

Escuela de Sistemas y Telemática

Minería de Texto en Medios Sociales: Caso de estudio del proyecto Tranvía de Cuenca.

Trabajo de graduación previo a la obtención del título de Ingeniero de Sistemas y Telemática

Autor: Belén Arias

Director: Ing. Marcos Orellana

Cuenca, Ecuador

2015

Edición autorizada de 30.000 ejemplares
Del 708.501 al 738.500
Nº 0710856

UNIVERSIDAD DEL AZUAY

1. Datos generales

1.1 Nombre del estudiante: Arias Zhañay María Belén

1.1.1 Código: 60460

1.1.2 Contacto: teléfonos:

Convencional: 072806399

Celular: 0987211128

Correo electrónico: blncilla@hotmail.es

1.2 Director sugerido: Orellana Cordero, Marcos Ingeniero de Sistemas

1.2.1 Contacto: marore@uazuay.edu.ec

1.3 Asesor metodológico: *Salgado Arteaga Francisco*

1.4 Tribunal designado:

1.5 Aprobación:

1.6 Línea de Investigación de la carrera: Base de Datos

1.6.1 Código UNESCO: 1203.18 Sistemas de Información

1.6.2 Tipo de trabajo:

Proyecto de investigación

1.7 Área de estudio: Minería de Datos

1.8 Título propuesto: ***Minería de Texto en Medios Sociales.***

1.9 Subtítulo: (opcional) Caso de estudio del proyecto Tranvía de Cuenca.

Edición autorizada de 30.000 ejemplares
Del 708.501 al 738.500 Nº 0710859

1.10 Estado del proyecto:

La investigación aborda las nuevas técnicas de análisis de datos como la minería de textos, se pretende realizar la investigación con el fin de demostrar la capacidad de estas técnicas a través de un análisis del sentimiento mediante minería de texto que provoca la implementación del proyecto Tranvía de Cuenca. Esto se realizará a través de los *tweet* que se generen en el medio social *Twitter* de los habitantes de la ciudad.

2. Contenido

2.1 Motivación de la investigación:

El análisis de datos mediante técnicas nuevas como la minería de textos es el objetivo de esta investigación para encontrar patrones ocultos en datos no estructurados, tales como los textos expresados en un lenguaje natural (ej.: castellano, inglés).

La implementación del proyecto Tranvía de Cuenca genera comentarios y posiciones por parte de los ciudadanos. Estos comentarios demuestran diferentes puntos de vistas sobre el proyecto y su análisis bridará información a las autoridades para la toma de decisiones.

El análisis de estas opiniones permitirá cumplir con el objetivo principal de la investigación, ya que a través del uso de minería de texto se busca obtener indicadores que permitan evaluar la posición que tienen los ciudadanos en cuanto al proyecto. Se opta por este tipo de análisis ya que se podrá analizar datos no estructurados provenientes de los *tweet* generados por los usuarios en el medio social *Twitter*.

2.2 Problemática:

En la actualidad la sociedad de la información, ha contribuido a que la comunicación de los seres humanos se efectué también, en buena medida, a través de medios digitales. Aparecen los medios sociales, que manifiestan el criterio del ser humano y surgen nuevas técnicas para procesar el criterio de las personas que a simple vista están ocultas, estas técnicas están comprendidas bajo un ámbito denominado minería de textos. Para demostrar la eficacia de estas técnicas se utilizará el caso del proyecto Tranvía de Cuenca, ya que es necesario conocer la postura de los ciudadanos en cuanto al proyecto; no solo

desde los medios tradicionales sino de medios digitales, en este caso *Twitter,* permitirá procesar la información no estructurada, analizarla y obtener la opinión que tiene el autor de un *tweet* en cuanto al caso de estudio.

2.3 Resumen:

La investigación tiene como objetivo obtener un análisis de la opinión expresada por parte de los usuarios sobre el caso Tranvía de Cuenca en el medio social *Twitter* a través del uso de algoritmos de minería de textos Para realizar el análisis de opinión se tomará los datos del medio social solamente de los usuarios que habitan en la ciudad de Cuenca y cuyos *tweets* se refieran al caso Tranvía de Cuenca. La extracción, almacenamiento, representación, visualización y análisis de los datos se realizará mediante el lenguaje R.

2.4 Indagación exploratoria y base conceptual:

Cada una de las actividades realizadas por el ser humano, genera reacciones tanto positivas como negativas, debido a que por cada tema que se plantea existen diferentes posiciones, que estas afectan a la forma en que una persona, una empresa o una organización son concebidas.

En la actualidad la sociedad de la información ha cambiado la manera de manifestarse de las personas, aparecen los medios sociales que son la reunión de personas que interactúan entre sí mediante medios digitales, donde pueden conocerse o no, pero existe retroalimentación entre ellos. Es ahí en la retroalimentación donde surge la opinión de los usuarios en uno u otro tema. (Domínguez, 2010)

Los medios sociales toman lo que cada usuario publica y esto no se queda ahí, sino empieza un diálogo virtual entre los mismos tomando sus referencias y opiniones, generando así una comunicación más realista, debido a que un usuario se siente mucho más cómodo de dar su idea a través de un medio social que al decirla personalmente. En realidad, un medio social refleja la verdadera personalidad de un individuo. Para Aciar y colaboradores (2015) los usuarios expresan sus opiniones en formato de texto libre

usando su lenguaje natural. A través de los medios sociales los usuarios se conectan, comparten y discuten cualquier tema, en cualquier lugar y en cualquier momento (Del Fresno & Otros, 2015).

Estos datos generados por un medio social son datos no estructurados. Un dato no estructurado es aquel que no está almacenado en una base de datos tradicional (estructura relacional, jerárquica o de red). Más bien es aquel dato que se toma de la web, de las cámaras móviles y videos, medios sociales, sensores de ciudades y edificios, etc. Es un dato que proviene de un texto normal, y se caracterizan por la variedad de su origen así como con la rapidez con la que incrementan su volumen. (Tascón, 2013).

El análisis de texto no estructurado era una tarea ardua y artesanal realizada por las personas; sin embargo, con el avance de la tecnología, surgen nuevos métodos y técnicas para simplificar esta tarea (Castillo & Otros, 2015).

Minería de datos se define como "conjunto de técnicas para la representación, análisis, manejo y descubrimiento de conocimiento a partir de diversas fuentes de datos (bases de datos, web, archivos, etc)" (McDonell & Otros, 2012). Involucra aspectos de la estadística y el conocimiento extraído se emplea para la toma de las decisiones.

Minería de texto son un conjunto de métodos que, mediante los datos aportados por los usuarios, obtiene información y encuentra los tópicos más hablados, que permiten categorizar a los usuarios según la información aportada. En la minería de texto cada opinión es tratada como un documento. (Lobos & Bartolome, 2012)

Así también, minería de texto según García-García es aquella que ofrece con sistemas automáticos una solución, sustituyendo o completando el trabajo de personas sin que importe la cantidad de texto. Encuentra información que antes era desconocida. Pueden ser patrones que de otro modo serían extremadamente difícil de descubrir. (García-García, & otros, 2014)

Así también, la minería de textos consiste en extraer la información que será útil de documentos no estructurados e identificar patrones interesantes de conocimiento

(Rodriguez Aldape, 2013). La minería de textos es una extensión de la minería de datos ya que muchas de las técnicas utilizadas en la minería de datos pueden ser aplicadas en la de textos. Ésta, busca obtener conocimiento a partir de grandes repositorios de datos.

Fielman y Sanger la dividen en cuatro áreas: (i) tareas de preprocesamiento, (ii) operaciones de minería, (iii) presentación y navegación, y (iv) técnicas de refinamiento.

(i) Tareas de pre-procesamiento: incluye rutinas, procesos y métodos requeridos para obtener y preparar los datos para las operaciones de descubrimiento de conocimiento.

(ii) Operaciones de minería: incluye la detección de patrones, análisis de tendencias y algoritmos de descubrimeinto de conocimiento.

(iii) Presentación y navegación: incluye herramientas de visualización y navegación de los patrones orientados al usuario.

Técnicas de refinamiento: incluye los métodos que filtran información redundante, la ordenan, resumen, generalizan y agrupan. (Feldman & Sanger, 2007)

Las técnicas de clasificación de minería de textos consisten en asignar objetos a categorías predefinidas. (Mansilla, Costaguta, & Missio, 2014)

El análisis de sentimiento es el área que se encarga de clasificar las opiniones de las personas. Estas opiniones reflejan su punto de vista y se clasifican en tres grupos: positivas, negativas y neutras. (Rodriguez Aldape, 2013). Aparece el concepto de *Social Media Mining* que se encarga de extraer, almacenar, representar, visualizar y analizar los datos que un usuario genera en un medio social obteniendo patrones significativos que permiten la toma de decisiones.

2.5 Objetivo general:

- Aplicar técnicas de minería de datos para el análisis de sentimiento de los usuarios del medio social Twitter en el caso Tranvía de Cuenca.

2.6 Objetivos específicos:

- Sistematizar información sobre conceptos y métodos de minería de textos.
- Obtener y depurar información no estructurada (*tweets* en lenguaje natural) del medio social Twitter de los ciudadanos.
- Aplicar algoritmos de minería de textos para analizar los *tweets* a través del lenguaje *R*.
- Representar objetivamente los resultados importantes en un reporte ilustrativo

2.7 Metodología:

La base teórica de este proyecto se va a desarrollar con la recopilación de información procedente de referencias en los siguientes portales web: Google Académico y bases digitales; para sintetizar las características y funcionalidades más relevantes de la minería de texto.

Para realizar el análisis se obtendrán los datos del medio social Twitter, se utilizarán algoritmos de minería de textos, los gráficos y análisis de sentimiento serán representados mediante el uso del lenguaje *R*. Los datos corresponderán a los usuarios del medio social *Twitter* que habiten en la ciudad de Cuenca y que la información que publiquen en un *tweet* sea referente al proyecto Tranvía de Cuenca.

2.8 Alcances y resultados esperados:

El alcance de este trabajo será obtener el análisis de opinión en los usuarios del medio social *Twitter* que se refieran al caso de estudio Tranvía de Cuenca. Mediante el análisis de gráficos que permitan obtener indicadores que se refieran al criterio de los usuarios.

2.9 Supuestos y riesgos:

Supuestos	Riesgos	Posible Solución

La investigación se muestra viable debido a que existen referencias bibliográficas en las que se puede sustentar su desarrollo.	La información no sea la suficiente para permitir el desarrollo de la investigación.	Utilizar referencias bibliográficas en inglés.
En cada uno de los capítulos de la investigación, se cumplan con los plazos acordados	Se producen atrasos que no cumplen con los plazos acordados.	Solicitar más tiempo para concluir con la investigación.
Los datos son recopilados sin problema.	Los datos no son generados para poder realizar el análisis.	Verificar nuevas técnicas
Se procede con la aplicación de algoritmos de minería de texto.	La aplicación de algoritmos no genera resultados esperados.	Cambiar de algoritmo para la evaluación de los datos.
El análisis de sentimientos con *R* se realiza de forma satisfactoria.	El análisis no puede realizarse ni generar resultados.	Cambiar el lenguaje para el análisis de sentimientos.

2.10 Presupuesto:

Rubro-Denominación	**Costo USD**	**Justificación**

Proveedor de internet	$120	Para la indagación exploratoria acerca de minería de textos sobre sus últimas investigaciones, conseguir artículos, consultar bases de datos bibliográficas para localizar referencias, etc.
Impresiones	$50	Para realizar la impresión de los documentos
Materiales de Oficina (papel, lápiz)	$10	Materiales que se utilizará para apuntar ideas, registrar pruebas, etc.

2.11 Financiamiento:

El financiamiento se realizará por parte del estudiante.

2.12 Esquema tentativo:

Capítulo I: Introducción

1.1 Introducción

1.2 Objetivos

1.3 Justificación

1.4 Alcance y Limitaciones

1.5 Marco de referencia

Capitulo II: Marco Teórico

1.1 Minería de datos

1.2 Minería de textos

1.3 Minería web

1.4 Medios sociales

UNIVERSIDAD DEL AZUAY

1.5 Twitter

1.6 Algoritmos

Capítulo III: Experimentación

1.1 Obtención de datos

1.2 Procesamiento

1.3 Presentación de Resultados

Capitulo IV: Resultados

1.1 Análisis de Sentimiento

1.2 Análisis de Resultados

Conclusiones

Referencias

2.13 Cronograma:

Objetivo Específico	**Actividad**	**Resultado esperado**	**Tiempo (semanas)**
Sistematizar información sobre conceptos y métodos de minería de textos.	Investigación teórica basada en artículos científicos, en las diferentes bibliotecas virtuales.	Disponer de buenas referencias teóricas que avalen la propuesta para la investigación.	4 semanas
Obtener y depurar información no estructurada (tweets en lenguaje natural) del medio social	Extraer la información del medio social Twitter.	Obtener los datos que sean necesarios para el análisis.	4 semanas

Twitter de los ciudadanos.			
Aplicar algoritmos de minería de textos para analizar los tweets a través del software R.	Selección del algoritmo a utilizar. Implementación del algoritmo. Prueba del algoritmo. Procesar los datos en el software R.	Obtener los resultados tras la aplicación de los algoritmos. Obtener el análisis de sentimientos.	4 semanas
Representar objetivamente los resultados importantes utilizando materiales ilustrativos.	Creación de gráficas.	Obtener gráficas que representen los resultados.	2 semanas

2.14 Referencias:

Aciar, S., Duque, N. D., & Aciar, M. (2015). Procesamiento de Opiniones de Usuarios Respecto a Objetos de Aprendizaje Basado en Ontologías y Minería de Texto. . Conferencias LACLO, 5 (1).

Castillo, J. J., Cardenas, M. E., Curti, A., & Casco, O. (2015). Software para asistencia en la creación de corpus para sistemas de análisis de texto no estructurado. *In XVII Workshop de Investigadores en Ciencias de la Computación.*

Del Fresno, M., Daly, A. J., & Supovitz, J. (2015). Desvelando climas de opinión por medio del Social Media Mining y Análisis de Redes Sociales en Twitter. *El caso de los Common Core State Standars. Revista Redes*.

Domínguez, D. C. (2010). Las Redes Sociales.Tipología, uso y consumo de las redes 2.0 en la sociedad digital actual. . *Documentación de las Ciencias de la Información*, 33, 45-68.

Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. . *Cambridge University Press*.

García-García, A., García-Massó, X., Ferrer-Sapena , A., González, L. M., Peset, F., Villamón, M., & Aleixandre, R. (2014). Text mining versus redes neuronales. Dos métodos de análisis aplicados al caso de las políticas de las revistas sobre datos.

Lobos, C., & Bartolome, L. (2012). Metodología de búsqueda de sub-comunidades mediante análisis de redes socialeas y minería de datos.

Mansilla, P. S., Costaguta, R., & Missio, D. (2014). Aplicación de Algoritmos de Clasificaación de Minería de Textos para el Reconocimiento de Habilidades de E-tutores Colaborativos. . *Revista Iberoamericana de Inteligencia Artificial*, 17 (53), 57-67.

McDonell, R., De la Fuente Aragón, M. V., & McDonnell, R. (2012). Minería de Datos Aplicada a la Gestión de la Información Urbanística Data Mining Applied to Urban Information Management. *In 6th Internacional Conference on Industrial Engineering and Industrial Management*, pp. 1476-1483.

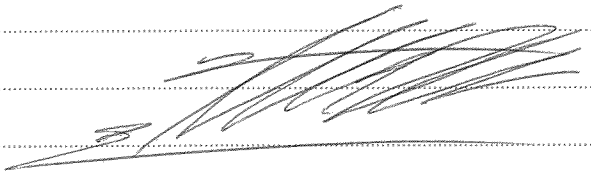
Rodriguez Aldape, F. (2013). Cuantificación del interés de un usuario en un tema mediante minería de texto y análisis de sentimiento. *Doctoral dissertation. Universidad Autónoma de Nuevo León*.

Edición autorizada de 30.000 ejemplares
Del 708.501 al 738.500 Nº 0710864

Tascón, M. (2013). Introducción: Big Data. Pasado, presente y futuro. . *Telos: Cuadernos de comunicación e innovación*, (95) 47-50.

2.15 Firma de responsabilidad

2.16 Firma de responsabilidad



2.17 Fecha de entrega: