ANÁLISIS DE DATOS DE SENTIMIENTOS ENFOCADOS AL SERVICIO DE TRASPORTE MASIVO TRANSMILENIO S.A APLICANDO TECNOLOGÍAS BIG DATA

PRESENTADO POR:
LUIS ENRIQUE RIOS RODRIGUEZ

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS
INGENIERIA DE SISTEMAS
BOGOTA D.C
2019

ANÁLISIS DE DATOS DE SENTIMIENTOS ENFOCADOS AL SERVICIO DE TRASPORTE MASIVO TRANSMILENIO S.A APLICANDO TECNOLOGÍAS BIG DATA

PRESENTADO POR:
LUIS ENRIQUE RIOS RODRIGUEZ

Trabajo de grado desarrollado como requisito para optar por el título de Ingeniero de Sistemas

DIRECTOR:
CELIO GIL AROS
INGENIERO DE SISTEMAS

FUNDACIÓN UNIVERSITARIA LOS LIBERTADORES
FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS
INGENIERIA DE SISTEMAS
BOGOTA D.C
2019

Nota de Aceptación

_____

_____

_____

_____

Presidente del Jurado

_____

Jurado

_____

Jurado

Bogotá D.C, 27 de noviembre de 2019

DEDICATORIA

Dedico este trabajo de grado primeramente a DIOS, por todo lo que él me ha dado durante toda mi vida, alzando mis manos hacia el cielo y diciendo en voz alta y entendible que todo lo que tengo se lo debo a él, que gracias a una de las muchas bendiciones que tengo es haciendo referencia en la oportunidad de tener a mi lado a mi madre Bella Aurora Rodriguez y a mi padre Luis Evelio Rios Jaramillo dos seres humanos que me han apoyado en todo el sentido de la palabra, guiándome por el camino correcto y siempre aconsejándome en cómo afrontar las adversidad con la frente en alto, inculcándome pasión por lo que estudie, pasión por lo que hago y con base a todo aquello y mucho más doy gracias a ellos porque hoy estoy logrando llegar a este último escalafón de la primera parte de mi formación profesional de mi vida, , a mis dos hermosos hijos Sara Michelle Rios Ariza y Christopher Rios Britel por inspirarme en ser un guerrero a pesar de mis errores y caídas brindando con su amor la fuerza para seguir en la lucha diaria por cumplir las metas y objetivos propuestos, a mi esposa hermosa Linda Deysy Britel Alzate que ha estado a mi lado para apoyarme y darme fuerzas en momentos donde las cosas no salían como lo esperaba, también ayudándome cada día en entender acerca de los caminos de DIOS y su maravillosa voluntad, dándome esa fuerza para nunca dejar de creer, también agradezco a cada una de las relaciones interpersonales que tuve y que aún tengo como lo son mis compañeros y docentes, al recordar a cada uno de ellos siento la alegría y la gran satisfacción de saber que durante este camino que me permitieron recibir apoyo, conocimiento, cariño en cada una de las circunstancias presentadas.

Luis Enrique Rios Rodriguez

AGRADECIMIENTOS

Primero que todo el agradecimiento a DIOS por permitirme llegar al cumplimiento de mi sueño profesional, porque sin la ayuda de él no logramos llegar al cumplimento de los objetivos y metas trazadas en cada uno de los ámbitos de nuestra vida, adicionalmente agradezco a todos los docentes de la Fundación Universitaria Los Libertadores, personas que cuentan con gran calidad humana y profesional, que en todo momento me brindaron de sus técnicas y metodologías de conocimientos, un agradecimiento especial a los docentes ingenieros Celio Gil Aros, Javier Daza Piragauta y Luis Alexis Plazas, docentes que me ayudaron mucho en mi formación profesional, que me aportaron el conocimiento necesario para ser lo que hoy soy, y que siempre tuvieron una relación de maestros y amigos, son una gran bendición y un excelente recurso humano para la institución.

Finalizando reitero el agradecimiento a mi hermosa y maravillosa familia compuesta por mis padres, mis hijos y mi esposa que gracias a toda la paciencia, tolerancia y amor me ayudaron a ser fuerte y soñador en todo el proceso de mi formación.

Queda en mi ser todos aquellos recuerdos, dando la importancia a todos aquellos recuerdos en los que caí y alguien llego para levantarme y animarme, con gran nostalgia llega el momento de nuevas metas las cuales pienso trabajar y afrontar con todo lo que he recibido por parte de mi excelente universidad LOS LIBERTADORES.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

ACRÓNIMOS

JSON: (JavaScript Objeto Notation), formato de texto para el intercambio de datos.

APIS: (application programming interface), la interfaz de programación de aplicaciones.

TB: (Terabyte) es una unidad de medida informática.

CRISP – DM (CRISP-DM: Towards a Standard Process Model for Data Mining), Modelo de fases.

GPS: (Global Positioning System): Sistema americano de navegación y localización mediante satélites.

CD: (compact disc) disco compacto.

XML: (Standard Generalized Markup Language), un lenguaje que permite la organización y el etiquetado de documentos.

HTML: (HyperText Markup Language) lenguaje de marcado que se utiliza para el desarrollo de páginas de Internet.

SMS: (Short Message Service), es un servicio disponible en los teléfonos móviles que permite el envío de mensajes.

ISACA: (Information Systems Audit and Control Association) Asociación de Auditoría y Control de Sistemas de Información.

OLAP: (On-Line Analytical Processing) es una solución utilizada en el campo de la llamada Inteligencia de negocios.

NFC:(Near Field Communication) comunicación de campo cercano.

RFID: (Radio Frequency Identification) identificación por radiofrecuencia, es un método de almacenamiento.

QR: (Quick Response code) código de respuesta rápida.

CRM: (Customer Relationship Management) administración de la relación con los clientes.

HPC: (High Performance Computing) la computación de alto rendimiento.

RAM: (Random Access Memory) Memoria de Acceso Aleatorio.

KDD: (Knowledge Discovery in Databalese) es un campo de la estadística y las ciencias de la computación.

MAC: (Media Access Control) es un identificador de 48 bits.

MATLAB: (MATrix LABoratory) laboratorio de matrices») es un sistema de cómputo numérico que ofrece un entorno de desarrollo integrado.

GNU: (not Unix) no es Unix.

RESUMEN

Este documento busca que los lectores conozcan acerca del análisis de sentimiento, este término también se conoce como minería de opinión (*Opinion Mining*), y está diseñado para determinar el tono emocional de una serie de palabras en menciones online, este es extremadamente útil en la monitorización de las redes sociales ya que permite hacernos una idea de la opinión pública general sobre ciertos temas. Este proyecto se enfocó a recolectar, analizar y validar datos de la entidad número uno de transporte masivo *Transmilenio S.A* compañía que presta su servicio en la ciudad de Bogotá D.C - Colombia. Adicionalmente que el mismo sirva como herramienta que permita realizar el análisis y almacenamiento de información Big Data.

ABSTRACT

This document seeks readers to know about sentiment analysis, this term is also known as opinion mining (Opinion Mining), and is designed to determine the emotional tone of a series of words in online mentions, this is extremely useful in the monitoring of social networks as it allows us to get an idea of the general public opinion on certain topics. This project focused on collecting, analyzing and validating data from the number one mass transit entity Transmilenio S.A company that provides its service in the city of Bogotá D.C - Colombia. Additionally, it serves as a tool that allows the analysis and storage of Big Data information.

INTRODUCCIÓN

Durante los últimos años los sistemas de información constituyen uno de los principales ámbitos de estudio en el área de organización de empresas. El entorno donde las compañías desarrollan sus actividades se vuelve cada vez más complejo. La creciente globalización, el proceso de internacionalización de la empresa, el incremento de la competencia en los mercados de bienes y servicios, la rapidez en el desarrollo de las tecnologías de información, el aumento de la incertidumbre en el entorno y la reducción de los ciclos de vida de los productos originan que la información se convierta en un elemento clave para la gestión, así como para la supervivencia y crecimiento de la organización empresarial. Si los recursos básicos analizados hasta ahora eran tierra, trabajo y capital, ahora la información aparece como otro insumo fundamental a valorar en las empresas.

En la actualidad, el poder de la información de una empresa puede incrementarse por su fiabilidad, volumen, accesibilidad y la capacidad que tiene dicha empresa para darle utilidad en un tiempo razonable, con el objetivo de ayudar en la toma de decisiones inteligentes. Big Data surge del hecho de grandes volúmenes de datos para procesarlos, analizarlos, descubrir patrones y otros aspectos fundamentales para la toma de decisiones. “La empresa que tiene la mejor información, sabe cómo encontrarla y puede utilizarla es la que triunfa más rápido” (Michel Daconta, Leo Obrst y Kevin T. Smith, 2004, The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management).

1. ASPECTOS DE INVESTIGACIÓN

1.1. TITULO

ANÁLISIS DE DATOS DE SENTIMIENTOS ENFOCADOS AL SERVICIO DE TRASPORTE MASIVO TRANSMILENIO S.A APLICANDO TECNOLOGÍAS BIG DATA.

1.2. ESTADO DEL ARTE

Según Molina (2005, p. 73), “el estado del arte es una modalidad de la investigación documental que permite el estudio del conocimiento acumulado (escrito en textos) dentro de un área específica”. Sobre esto, Vélez y Galeano (2002), citado por Londoño et al.

(2014, p 18), precisan: “El Estado del arte se percata de las investigaciones recientes respecto a las categorías de análisis de la investigación, partiendo de un lectura y análisis intra e intertextual en un tiempo y espacio geográfico determinado”.

Además, Schwarz (2013), citado por Londoño et al. (2014, p. 19), agrega como el estado del arte.

Contiene la base más profunda de la investigación que permite descubrir conocimiento nuevo al revisar la literatura asociada al tema de investigación de manera que pueda determinarse quiénes, cómo, cuándo, dónde y por qué han tratado de resolver el problema de investigación, determinar su actualización y verificar si el tema sigue vigente, así como descubrir hasta dónde ha avanzado el conocimiento validado más reciente sobre el tema en el que se está trabajando.

Cuando se elabora un estado del arte, se identifican de una forma rápida y acertada las fronteras del conocimiento respecto al problema de investigación, lo que significa que cualquier desviación y aspecto por estudiar traslada casi directamente al investigador al desarrollo de los nuevos conocimientos. De lo anterior se puede decir que el propósito principal del estado del arte es mostrar el estado actual del conocimiento en un determinado campo o tema específico, con el fin de orientar adecuadamente las investigaciones científicas a realizar. En la construcción del documento y sus respectivos apartes, se debe evidenciar la intertextualidad en la escritura investigativa. En este sentido, Sánchez (2011, p. 62) acerca de la intertextualidad plantea que “se trata de una actividad de construcción de un texto con base en otros textos”. Esto hace referencia a la adecuada utilización de un texto a partir del conocimiento previo que se tenga de otros textos, reconociendo el aporte que otros han realizado sobre lo que se está investigando. La compilación de resultados de otras investigaciones que sobre el tema de

investigación elegido se han efectuado, permite ir suministrando los insumos para la construcción del nuevo documento que hace parte del anteproyecto o proyecto de investigación.

1.2.1 Antecedentes

A través de búsqueda de investigaciones y fuentes de información, registro de memorias de congresos, revistas especializadas en el área de interés, bases de datos electrónicas y tesis se encuentra un gran número de información que hace referencia al análisis de datos de sentimientos o también conocido como análisis de datos de emociones, hoy en día la mayoría de entidades han aplicado o están aplicando métodos y/o técnicas de *Big Data*, esto con el fin de mejorar la calidad de sus servicios y la optimización de procesos.

Algunos ejemplos de compañías conocidas que decidieron ingresar a las tecnologías del Big Data son:

Alpina, tiene ahora menos interés en publicidad sobre medios tradicionales, excepto la televisión. Se enfoca en las redes sociales y su estrategia de transformación digital la llevó a establecer una nueva forma de segmentación del público, ya no basada en edades ni estratos socioeconómicos, como el mercadeo convencional, sino en intereses y estilos de vida, por lo cual crearon páginas en Facebook.

Carvajal representa un emblemático caso de compañías que salieron airosas tras enfrentar el reto de la revolución digital. Fundada en 1904, llegó a tener 23 negocios diferentes, la mayoría orientados a impresión, basados en la cultura del papel. Para ajustarse al cambio de la tecnología, Carvajal se salió de algunos de ellos y profundizó en otros. Se retiró del negocio de directorios telefónicos, de la edición de libros y la impresión de valores, y se concentró en empaques, pulpa y papel, soluciones educativas y tecnología. De ese modo, absorbió a lo largo de 15 años de transformación las turbulencias que parecían traer el advenimiento del mundo digital. “Carvajal se transformó, no sin dolores, pero es claro que hizo bien la tarea”, explica Bernardo Quintero, presidente del Grupo Carvajal.

Gómez, A. J. (2014, abril, 29). Así se han montado las empresas colombianas a la onda digital, Semana,100,1-4. Recuperado de http://www.semana.com/100-empresas/articulo/100-empresas-2017-empresas-colombianas-se-renuevan-con-tecnologia/523421.

Además de encontrar casos de éxito de algunas compañías colombianas también esta investigación se enfocó en la extracción de datos en redes sociales, esto con el objetivo de analizar validar y proponer soluciones a aquellas problemáticas que afectan un dominio en específico.

Minería de medios sociales: Para (Zafarani et al., 2014), las redes sociales rompen las fronteras entre el mundo real y el mundo virtual. Ahora se pueden integrar las teorías sociales con métodos computacionales para estudiar cómo interactúan los individuos (también conocidos como átomos sociales) y como se forman las comunidades (es decir, las moléculas sociales). La singularidad de los datos de los medios de comunicación social requiere nuevas técnicas de minería de datos que puedan manejar eficazmente contenido generado por los usuarios. El estudio y desarrollo de estas nuevas técnicas están bajo el ámbito de la minería de medios sociales, una disciplina emergente bajo el paraguas de la minería de datos.

Análisis de Sentimiento: Para la firma (Brandwatch, 2015), el análisis de sentimiento es el proceso de determinar el tono emocional que hay detrás de una serie de palabras, y se utiliza para intentar entender las actitudes, opiniones y emociones expresadas en una mención online.

El análisis sentimiento es extremadamente útil en la monitorización de las redes sociales ya que permite hacernos una idea de la opinión publica general sobre ciertos temas. Herramientas de monitorización de las redes sociales como Brandwatch Analytics hacen que este proceso sea mucho más rápido y fácil que nunca gracias a la capacidad de monitorizar en tiempo real. Los beneficios del análisis de sentimiento son numerosos e importantes. La habilidad de extraer información de datos de las redes sociales es una práctica que ya están adoptando organización

Minería de opinión: Para (Liu, 2012), la opinión sería el centro de todas las actividades humanas, es un aspecto importante que influye en el comportamiento humano. La minería de opinión, representa un estudio computacional de ´ las opiniones, actitudes y evaluaciones con respecto a una entidad y sus aspectos. La entidad se refiere a un producto, servicio, organización o a una persona determinada y los aspectos a los atributos de la entidad. De igual manera, para (Rafea and Mostafa, 2013) y (Shahheidari et al., 2013), la minería de opinión es un área de investigación dedicada a la ´ extracción de los sujetos dominantes o el análisis de subjetividad de un texto dado.

(Medhat et al., 2014) en su artículo" Algoritmos y aplicaciones de análisis de sentimientos", exponen que el análisis ´ de sentimiento y la minería de opinión, en realidad, expresan un significado mutuo, son intercambiables, complementarias y representan el estudio computacional de las opiniones, actitudes y emociones de las personas hacia una entidad. La entidad puede representar individuos, eventos o temas. La minería de extracción y análisis relacionadas con la opinión de la gente sobre una entidad mientras que análisis de sentimiento identifica el sentimiento expresado en un texto y luego lo analiza. Por lo tanto, el objetivo de análisis del sentimiento es encontrar opiniones, identificar los sentimientos que expresan, y luego clasificar su polaridad.

Un ejemplo de este preprocesamiento puede verse en el trabajo de Yu y Wang (2015). Algunos de las acciones que llevaron a cabo fueron: eliminar las URL del mensaje, pues no aportan información de sentimiento; tokenizar o extraer las palabras del tweet; pasar a minúsculas, y quitar las stopwords o palabras vacías, esto es, aquellas palabras que no aportan información de sentimiento como los artículos y las preposiciones

Según Medhat et al. (2014), lo más habitual en el preprocesamiento es quitar las stopwords y lematizar. La unicación de las formas léxicas supone identificar términos con el mismo significado semántico, como sería el caso de una palabra en singular y otra en plural, o las diferentes formas verbales. Por ejemplo: la palabra "comerían" pasaría a ser "comer", de tal forma que se identifica un mismo común sentido semántico en todas las formas verbales de "comer" que aparezcan en el corpus.

El TASS 2015 estableció diversas tareas de análisis de sentimiento (García Cumbreras, Martínez Cámara, Villena Román y García Morera, 2016). En nuestro caso, nos vamos a centrar en la más sencilla: la clasificación según la polaridad del sentimiento. Para ello, se disponía de un corpus de 68.000 tweets escritos en castellano por 150 personalidades del mundo de la política, economía, comunicación y cultura, el cual había sido redactado entre noviembre 2011 y marzo 2012 (Villena-Román, Lana-Serrano, Martínez-Cámara y González-Cristóbal,2013). Nos referiremos a él como corpus TASS.

A continuación, se realizó el proceso de extracción de atributos, que constituye el corazón de todo el modelo. Tuvo las siguientes fases (Play Hurtado, 2013):

- Se consideraron solo unigramas de lemas con una frecuencia mínima preestablecida. Un unigramas equivale a una palabra del mensaje del tweet.
- Los hashtags, las menciones a usuarios, los números, las fechas y los signos de puntuación fueron unificados respectivamente en una sola característica.
- Se sustituyeron los emoticonos por su correspondiente categoría: happy, sad, tongue, wink y otros.
- Se excluyeron los términos pertenecientes a ciertas categorías morfosintácticas poco significativas para el análisis de sentimiento.
- Se utilizó como recurso externo varios diccionarios de polaridad.

Cabe mencionar que uno de los estudios interesantes, el cual brinda evidencia de la importancia de aplicar técnicas de extracción de datos, con fines de hallar soluciones ante problemáticas presentadas en un entorno, obedece a tesis de grado con el nombre de *seguimiento a las barras bravas del futbol en la ciudad de Pereira, basado en tecnologías big data* del señor González Marín estudiante de la Universidad Tecnológica de Pereira, este trabajo hace extracción y análisis de datos de sentimientos en Twitter referente al comportamiento de las barras bravas de futbol en Pereira, dando como resultado un estudio técnico – social el cual nos muestra de forma organizada y detallada cual son las principales actividades que generan malestares al interior de la ciudad.

1.3. DEFINICIÓN DEL PROBLEMA

El sistema de transporte masivo Transmilenio S.A fue implementado en la ciudad de Bogotá D.C Colombia en el año 2000 fue vista en principio como la solución a los problemas de movilidad que se presentan en las ciudades y puntos más estratégicos de una ciudad. A pesar de ser una mejora para la ciudad el sistema Transmilenio ha venido presentando varios problemas detallados a continuación:

Alto precio de los pasajes, ineficacia para satisfacer la demanda creciente (pocos buses), protestas, bloqueos, largas filas para adquirir la compra del pasaje y del mismo modo para ingresar y salir de una estación, problemas de corrupción, escándalos de abuso sexual, personas que se cola, accidentes de atropellos que terminan en muerte, vendedores ambulantes, entre otros), todos aquellos factores mencionados han intensificaron y desembarcaron en una crisis multidimensional (financiera, técnica, sociocultural, urbanística y político-administrativa) en este sistema de transporte masivo. No obstante, aunque se han estudiado estos problemas, los estudios se han centrado en las explicaciones desde el punto de vista técnico y financiero, dejando de lado la importancia de un análisis desde disciplinas relevantes para la planeación urbana, como la ciencia política.

Se hace necesario el desarrollo de un análisis de datos que, desde el aspecto político, administrativo y social, evidencie las razones estructurales que tienen hasta hoy en día al sistema de transporte masivo Transmilenio sumergido en una crisis.

Para este análisis de datos nos enfocaremos y apoyaremos con los recursos que nos proveen las famosas redes sociales, ya que esta contiene grandes volúmenes de datos que las tecnologías de la información generan. En el año 2000, se almacenaron en el mundo 800.000 Petabytes. Se espera que para el año 2020 se alcancen 35 Zettabytes. Solo Twitter genera más de 9 terabytes (TB) de datos cada día. Facebook 10 TB (Joyanes, 2013), toda esta información adquiere connotación de información masiva y por ello mismo se hace difícil de procesar, almacenar y entender desde un punto de vista computacional debido a la variedad de formas que expresa ya que no son solo datos estructurados para las bases de datos relacionales (datos tradicionales)

que se suelen utilizar en las mayoría de las entidades corporativas y/o instituciones, estos datos no tienen formatos fijos o no tienen estructura uniforme, son los denominados datos semiestructurados y no estructurados respectivamente (Joyanes, 2013).

Con base a lo anterior y poner en marcha el análisis de datos de sentimientos nos adentramos a la era del Big data el cual describe el gran volumen de datos, tanto estructurados como no estructurados, que inundan una empresa en el día a día. Pero no es la cantidad de datos lo que importa. Lo que las organizaciones hacen con los datos es lo que importa. Los grandes datos se pueden analizar para obtener información que conduzca a mejores decisiones y movimientos estratégicos de negocios.

1.4. JUSTIFICACIÓN

El análisis de sentimientos, con la ayuda del big data, se ha convertido en la herramienta más fiable y precisa de la actualidad. El avance de la sociedad y la llegada de la era digital ha obligado a afinar más las técnicas de análisis para responder a las necesidades del mercado. Las compañías y agencias de comunicación entienden el valor de esta información y por ello la solicitan para sus estrategias comerciales, la atención al cliente y la detección de tendencias. Gracias a los estudios basados en este sistema, las empresas pueden detectar intenciones y el contagio social, determinar el éxito de un negocio, realizar una segmentación de los usuarios y conocer mejor al cliente. Por ellos, se convierte en una información con un valor incalculable.

Los avances en big data influyen directamente en la sociedad. La creación de aplicaciones de este tipo que mejoran notablemente su método es sólo un ejemplo de lo que la recopilación de datos unidos al análisis puede llegar a hacer. En el uso del big data está una de las claves para el desarrollo en cualquier campo, contar con este dotará a los diferentes sectores de nuevos instrumentos capaces de mejorar y facilitar su trabajo.

Cada vez la situación de Transmilenio es más compleja haciendo que esto genere pensamientos de poca credibilidad en cuanto a mejoras, es tan alto el índice de inconformidad que ha llegado al límite de generar grandes problemas entre la sociedad, mala administración financiera y bastante inconformismo entre todos los ciudadanos que habitan en Bogotá D.C.

Esta temática no es nueva y ya no logra sorprender a nadie, es tanto que se ha convertido en el vivir de cada día entre los que suelen utilizar el sistema para trasladarse a su rumbo de destino, los casos van desde una simple y larga fila hasta con la muerte provocados por hurto y/o por esquivar la responsabilidad de pagar un pasaje en el sistema.

A través de la era digital se ha tratado de crear conciencia tanto para el usuario como para las administrativas del sistema, llegando a pensar que toda esta información pasa desapercibida por los mismos encargados del sistema.

El uso cada vez mayor de terminales móviles tipo Smartphone conectados a la web es una oportunidad que nos permite realizar seguimiento a las acciones y situaciones que provocan este tipo de inconvenientes en el sistema, por medio de un modelo basado en tecnologías Big Data.

1.5. PREGUNTAS DE INVESTIGACIÓN

¿Cómo a través de la interpretación del Análisis de Sentimientos, en conjunto con las técnicas del Big Data garanticen el mejoramiento al servicio del transporte masivo TRANSMILENIO S.A. en la ciudad de Bogotá?

1.6. OBJETIVOS

1.6.1. Objetivo General

Proponer un modelo de seguimiento y verificación al comportamiento de los acontecimientos que ocurre dentro y fuera del sistema de transporte masivo Transmilenio, soportado en tecnologías Big Data y restringido a la red social Twitter.

1.6.2. Objetivos Específicos

1. Elaboración del estado del arte
2. Determinar, seleccionar y poblar un modelo de recolección de información de la red social Twitter, en torno a eventos específicos.
3. Evaluar y seleccionar las técnicas de filtrado y de análisis de datos, aplicables a la información proveniente de la red social Twitter.
4. Filtrar y analizar la información obtenida de la red social Twitter, aplicando las técnicas seleccionadas de Big Data, para predecir posibles problemas reiterativos en el sistema de TransMilenio.
5. Demostrar que las técnicas de análisis de datos de sentimientos son recomendado y muy útiles para hallar las causas y posibles soluciones a problemas, independiente sea el campo factor de la necesidad.

1.7. IMPACTO

El uso de técnicas de recolección y análisis de datos como parte de una solución de inteligencia de negocio que suplirá las necesidades de mitigación de los distintos eventos generados en la funcionalidad del sistema, llegando a lograr la reducción de errores, y a cambio generando avance estratégico que permita el desarrollo y apoyo a la toma de decisiones en beneficio mutuo (Ciudadanía & Compañía).

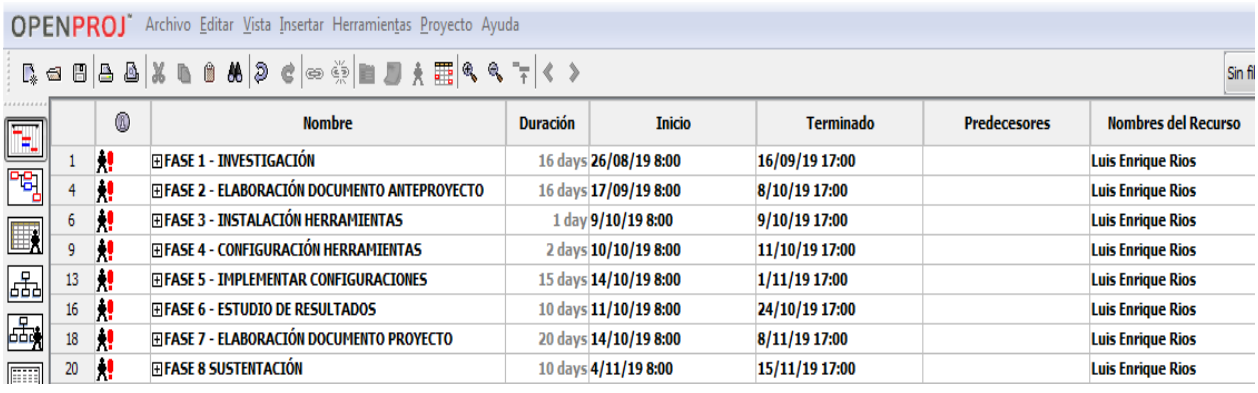
1.8. DELIMITACIÓN

1.8.1. Espacial.

Este proyecto se realizará en las instalaciones de la Fundación Universitaria Los Libertadores ubicada en la Carrera 16 A # 63 A – 80, Barrio Chapinero, contando con el apoyo de dirección del proyecto.

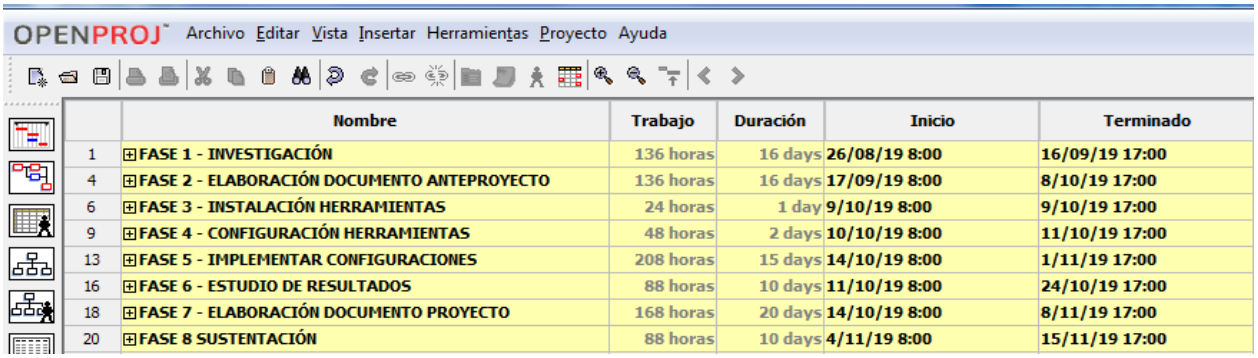
1.8.2. Cronológica.

El proyecto tendrá una duración aproximadamente de (2) Meses y (20) días a partir calendario de aprobación.



		Nombre	Duración	Inicio	Terminado	Predecesores	Nombres del Recurso
1		FASE 1 - INVESTIGACIÓN	16 days	26/08/19 8:00	16/09/19 17:00		Luis Enrique Rios
4		FASE 2 - ELABORACIÓN DOCUMENTO ANTEPROYECTO	16 days	17/09/19 8:00	8/10/19 17:00		Luis Enrique Rios
6		FASE 3 - INSTALACIÓN HERRAMIENTAS	1 day	9/10/19 8:00	9/10/19 17:00		Luis Enrique Rios
9		FASE 4 - CONFIGURACIÓN HERRAMIENTAS	2 days	10/10/19 8:00	11/10/19 17:00		Luis Enrique Rios
13		FASE 5 - IMPLEMENTAR CONFIGURACIONES	15 days	14/10/19 8:00	1/11/19 17:00		Luis Enrique Rios
16		FASE 6 - ESTUDIO DE RESULTADOS	10 days	11/10/19 8:00	24/10/19 17:00		Luis Enrique Rios
18		FASE 7 - ELABORACIÓN DOCUMENTO PROYECTO	20 days	14/10/19 8:00	8/11/19 17:00		Luis Enrique Rios
20		FASE 8 SUSTENTACIÓN	10 days	4/11/19 8:00	15/11/19 17:00		Luis Enrique Rios

Print 1. Fase consolidada. Fuente Autor.

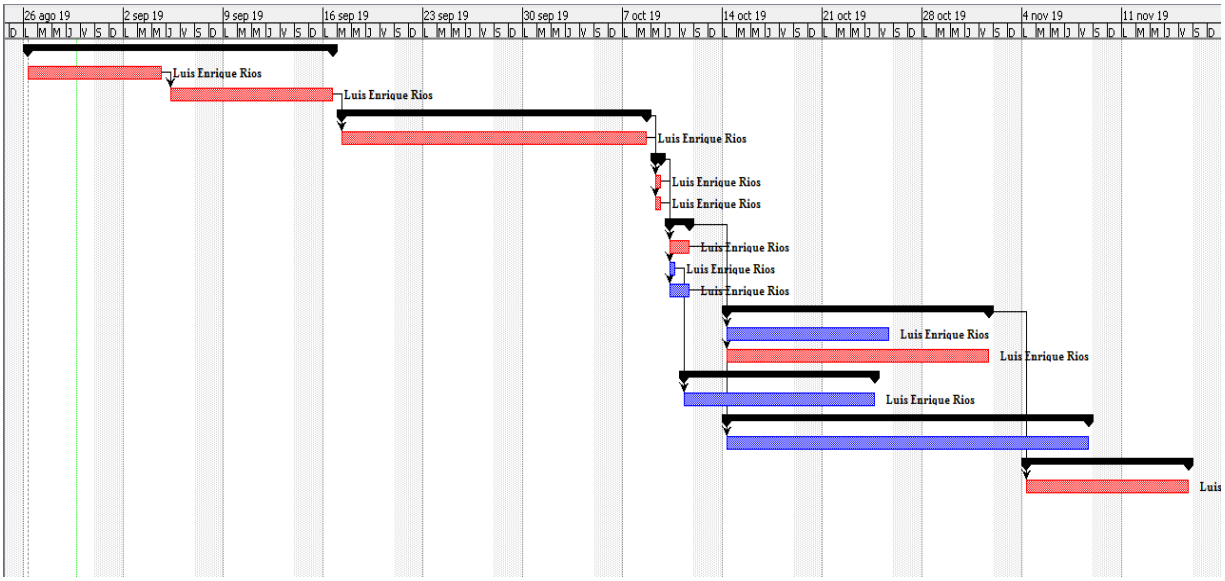


	Nombre	Trabajo	Duración	Inicio	Terminado
1	FASE 1 - INVESTIGACIÓN	136 horas	16 days	26/08/19 8:00	16/09/19 17:00
4	FASE 2 - ELABORACIÓN DOCUMENTO ANTEPROYECTO	136 horas	16 days	17/09/19 8:00	8/10/19 17:00
6	FASE 3 - INSTALACIÓN HERRAMIENTAS	24 horas	1 day	9/10/19 8:00	9/10/19 17:00
9	FASE 4 - CONFIGURACIÓN HERRAMIENTAS	48 horas	2 days	10/10/19 8:00	11/10/19 17:00
13	FASE 5 - IMPLEMENTAR CONFIGURACIONES	208 horas	15 days	14/10/19 8:00	1/11/19 17:00
16	FASE 6 - ESTUDIO DE RESULTADOS	88 horas	10 days	11/10/19 8:00	24/10/19 17:00
18	FASE 7 - ELABORACIÓN DOCUMENTO PROYECTO	168 horas	20 days	14/10/19 8:00	8/11/19 17:00
20	FASE 8 SUSTENTACIÓN	88 horas	10 days	4/11/19 8:00	15/11/19 17:00

Print 2. Fase Consolidada. Fuente Autor.

	Nombre	Trabajo	Duración	Inicio	Terminado	Entorno de Trabajo
1	FASE 1 - INVESTIGACIÓN	136 horas	16 days	26/08/19 8:00	16/09/19 17:00	
2	Investigación Big Data	64 horas	8 days	26/08/19 8:00	4/09/19 17:00	
	Luis Enrique Rios	*64 horas*	*8 days*	*26/08/19 8:00*	*4/09/19 17:00*	*Plano*
3	Investigación lenguaje de programación R	64 horas	8 days	5/09/19 8:00	16/09/19 17:00	
	Luis Enrique Rios	*64 horas*	*8 days*	*5/09/19 8:00*	*16/09/19 17:00*	*Plano*
	Luis Enrique Rios	*8 horas*	*1 day*	*26/08/19 8:00*	*26/08/19 17:00*	*Plano*
4	FASE 2 - ELABORACIÓN DOCUMENTO ANTEPROYE(	136 horas	16 days	17/09/19 8:00	8/10/19 17:00	
5	Guia estandar de idea principal	128 horas	16 days	17/09/19 8:00	8/10/19 17:00	
	Luis Enrique Rios	*128 horas*	*16 days*	*17/09/19 8:00*	*8/10/19 17:00*	*Plano*
	Luis Enrique Rios	*8 horas*	*1 day*	*17/09/19 8:00*	*17/09/19 17:00*	*Plano*
6	FASE 3 - INSTALACIÓN HERRAMIENTAS	24 horas	1 day	9/10/19 8:00	9/10/19 17:00	
7	Instalación lenguaje de programación R	8 horas	1 day	9/10/19 8:00	9/10/19 17:00	
	Luis Enrique Rios	*8 horas*	*1 day*	*9/10/19 8:00*	*9/10/19 17:00*	*Plano*
8	Instalación entono de desarrollo RStudio	8 horas	1 day	9/10/19 8:00	9/10/19 17:00	
	Luis Enrique Rios	*8 horas*	*1 day*	*9/10/19 8:00*	*9/10/19 17:00*	*Plano*
	Luis Enrique Rios	*8 horas*	*1 day*	*9/10/19 8:00*	*9/10/19 17:00*	*Plano*
9	FASE 4 - CONFIGURACIÓN HERRAMIENTAS	48 horas	2 days	10/10/19 8:00	11/10/19 17:00	
10	Instalación de paquetes	16 horas	2 days	10/10/19 8:00	11/10/19 17:00	
	Luis Enrique Rios	*16 horas*	*2 days*	*10/10/19 8:00*	*11/10/19 17:00*	*Plano*
11	Llamar las librerías	8 horas	1 day	10/10/19 8:00	10/10/19 17:00	
	Luis Enrique Rios	*8 horas*	*1 day*	*10/10/19 8:00*	*10/10/19 17:00*	*Plano*
12	Adquirir valores para las variables a traves de Twitter	16 horas	2 days	10/10/19 8:00	11/10/19 17:00	
	Luis Enrique Rios	*16 horas*	*2 days*	*10/10/19 8:00*	*11/10/19 17:00*	*Plano*
	Luis Enrique Rios	*8 horas*	*1 day*	*10/10/19 8:00*	*10/10/19 17:00*	*Plano*
13	FASE 5 - IMPLEMENTAR CONFIGURACIONES	208 horas	15 days	14/10/19 8:00	1/11/19 17:00	
14	Técnicas de filtrado	80 horas	10 days	14/10/19 8:00	25/10/19 17:00	
	Luis Enrique Rios	*80 horas*	*10 days*	*14/10/19 8:00*	*25/10/19 17:00*	*Plano*
15	Conocimiento	120 horas	15 days	14/10/19 8:00	1/11/19 17:00	
	Luis Enrique Rios	*120 horas*	*15 days*	*14/10/19 8:00*	*1/11/19 17:00*	*Plano*
	Luis Enrique Rios	*8 horas*	*1 day*	*14/10/19 8:00*	*14/10/19 17:00*	*Plano*
16	FASE 6 - ESTUDIO DE RESULTADOS	88 horas	10 days	11/10/19 8:00	24/10/19 17:00	
17	Recopilación y lectura de la información	80 horas	10 days	11/10/19 8:00	24/10/19 17:00	
	Luis Enrique Rios	*80 horas*	*10 days*	*11/10/19 8:00*	*24/10/19 17:00*	*Plano*
	Luis Enrique Rios	*8 horas*	*1 day*	*11/10/19 8:00*	*11/10/19 17:00*	*Plano*
18	FASE 7 - ELABORACIÓN DOCUMENTO PROYECTO	168 horas	20 days	14/10/19 8:00	8/11/19 17:00	
19	Documentación completa referente a todo el proyecto	160 horas	20 days	14/10/19 8:00	8/11/19 17:00	
	Luis Enrique Rios	*8 horas*	*1 day*	*14/10/19 8:00*	*14/10/19 17:00*	*Plano*
20	FASE 8 SUSTENTACIÓN	88 horas	10 days	4/11/19 8:00	15/11/19 17:00	
21	Elaborar y preparar sustentación	80 horas	10 days	4/11/19 8:00	15/11/19 17:00	
	Luis Enrique Rios	*80 horas*	*10 days*	*4/11/19 8:00*	*15/11/19 17:00*	*Plano*
	Luis Enrique Rios	*8 horas*	*1 day*	*4/11/19 8:00*	*4/11/19 17:00*	*Plano*

Print 4. Fase General Cronograma. Fuente Autor.



Print 4. Diagrama Gantt. Fuente Autor.

1.9 ALCANCE

El alcance de este proyecto es lograr reunir toda la información necesaria a cerca de una de las problemáticas que afecta la capital del país, con base a lo anterior, nos enfocaremos en Twitter una de las redes sociales con más usuarios, nuestro eje central circula a través de la extracción de datos, aplicando tácticas y técnicas de las tecnologías que hacen parte de Big Data.

1.10. RECURSOS

1.10.2. Recurso Hardware

Estos recursos son básicos, y no se necesita de hardware avanzado.

No.	Concepto
1	Pentium (R) Dual Core CPU E5700 @ 3.00GHz
2	Disco Duro 500 GB

Tabla 1. Recurso Hardware. Fuente Autor.

1.10.3. Recurso Software

El software requerido es libre y de pago, en este proyecto las herramientas contemplan periodos de tiempo de evaluación.

No.	Concepto
1	Ejecutable software lenguaje R Vo. 3.6.1
2	Ejecutable software RStudio Vo. 3.6.1
3	Microsoft Office 2016.
4	
5	

Tabla 2. Recurso Software. Fuente Autor.

1.11. METODOLOGÍA

Como metodología del presente proyecto, se toma la definición del modelo de proceso estándar para minería de datos CRISP – DM (CRISP-DM: Towards a Standard Process Model for Data Mining), en las fases y tareas respectivas del mismo, así como para sus salidas.

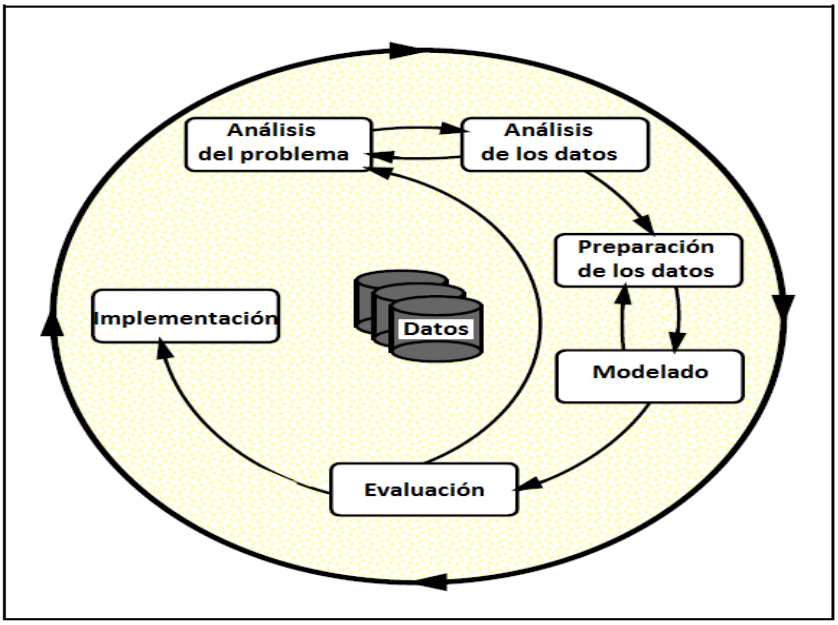


Figura 1. Fases modelo de proceso estándar para minería de datos CRISP – DM.

1.11.1. Análisis del problema

Al pasar el tiempo, el sistema de transporte masivo Transmilenio S.A ha generado bastante polémica entre entidades del sector público y privado, llegando así a perjudicar su eje central que es el ciudadano, aquel que debe utilizar de este medio para lograr optimizar recorridos que con otros sistemas de transporte es más complejo.

En el sistema de transporte se ve a diario corrupción, vandalismo, casos de agresión personal, abuso de la ley entre muchas otras más.

1.11.2. Análisis de los datos

Durante esta etapa se hace necesario el entendimiento de los datos contenidos en los Data Set producidos por la red social Twitter, reconocer la estructura de sus mensajes y contemplarlos frente a los objetivos del proyecto.

1.11.3. Preparación de los datos

En esta etapa es de suma importancia saber cómo realizar los filtros de los datos obtenidos mediante la técnica aplicada, en este se detallará la información referente de comentarios positivos y negativos del sistema de transporte, haciendo énfasis a problemática interna y externa.

1.11.4. Modelado

Etapa de creación y modelamiento del seguimiento de los casos más relevantes de la problemática.

1.11.5. Evaluación

En esta etapa se analiza y se verifica la etapa anteriormente nombrada como modelado, con el fin de corregir y/o agregar algún otro tipo de modelo.

1.11.6. Implementación

Etapa final de la metodología empleada, en este paso se obtiene el modelo completo y final.

2. MARCO TEÓRICO

2.1. DEFINICIÓN DE BIG DATA

El primer cuestionamiento que posiblemente llegue a su mente en este momento es ¿Qué es Big Data y porqué se ha vuelto tan importante? pues bien, en términos generales podríamos referirnos como a la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos. Entonces ¿Cuánto es demasiada información de manera que sea elegible para ser procesada y analizada utilizando Big Data? Analicemos primeramente en términos de bytes:

Además del gran volumen de información, esta existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de una oportunidad para Big Data.

Es importante entender que las bases de datos convencionales son una parte importante y relevante para una solución analítica. De hecho, se vuelve mucho más vital cuando se usa en conjunto con la plataforma de Big Data. Pensemos en nuestras manos izquierda y derecha, cada una ofrece fortalezas individuales para cada tarea en específico. Por ejemplo, un beisbolista sabe que una de sus manos es mejor para lanzar la pelota y la otra para atraparla; puede ser que cada mano intente hacer la actividad de la otra, más, sin embargo, el resultado no será el más óptimo.

2.2. FUENTE PROVENIENTE DE INFORMACIÓN

Los seres humanos estamos creando y almacenando información constantemente y cada vez más en cantidades astronómicas. Se podría decir que, si todos los bits y bytes de datos del último año fueran guardados en CD's, se generaría una gran torre desde la Tierra hasta la Luna y de regreso.

Esta contribución a la acumulación masiva de datos la podemos encontrar en diversas industrias, las compañías mantienen grandes cantidades de datos transaccionales, reuniendo información acerca de sus clientes, proveedores, operaciones, etc., de la misma manera sucede con el sector público. En muchos países se administran enormes bases de datos que contienen datos de censo de población, registros médicos, impuestos, etc., y si a todo esto le añadimos transacciones financieras realizadas en línea o por dispositivos móviles, análisis de redes sociales (en Twitter son cerca de 12 Terabytes de tweets creados diariamente y Facebook almacena alrededor de 100 Petabytes de fotos y videos), ubicación geográfica mediante coordenadas GPS, en otras palabras, todas aquellas actividades que la mayoría de nosotros realizamos varias veces al día con nuestros "smartphones", estamos hablando de que se generan alrededor de 2.5 quintillones de bytes diariamente en el mundo.

De acuerdo con un estudio realizado por Cisco, entre el 2011 y el 2016 la cantidad de tráfico de datos móviles crecerá a una tasa anual de 78%, así como el número de dispositivos móviles conectados a Internet excederá el número de habitantes en el planeta. Las naciones unidas proyectan que la población mundial alcanzará los 7.5 billones para el 2016 de tal modo que habrá cerca de 18.9 billones de dispositivos conectados a la red a escala mundial, esto conllevaría a que el tráfico global de datos móviles alcance 10.8 Exabytes mensuales o 130 Exabytes anuales. Este volumen de tráfico previsto para 2016 equivale a 33 billones de DVDs anuales o 813 cuatrillones de mensajes de texto.

Pero no solamente somos los seres humanos quienes contribuimos a este crecimiento enorme de información, existe también la comunicación denominada máquina a máquina (M2M machine-to-machine) cuyo valor en la creación de grandes cantidades de datos también es muy importante. Sensores digitales instalados en contenedores para determinar la ruta generada durante una entrega de algún paquete y que esta información sea enviada a las compañías de transportación, sensores en medidores eléctricos para determinar el consumo de energía a

intervalos regulares para que sea enviada esta información a las compañías del sector energético. Se estima que hay más de 30 millones de sensores interconectados en distintos sectores como automotriz, transportación, industrial, servicios, comercial, etc. y se espera que este número crezca en un 30% anualmente.

2.3. TIPOS DE DATOS

Muchas organizaciones se enfrentan a la pregunta sobre ¿qué información es la que se debe analizar?, sin embargo, el cuestionamiento debería estar enfocado hacia ¿qué problema es el que se está tratando de resolver?

Si bien sabemos que existe una amplia variedad de tipos de datos a analizar, una buena clasificación nos ayudaría a entender mejor su representación, aunque es muy probable que estas categorías puedan extenderse con el avance tecnológico.

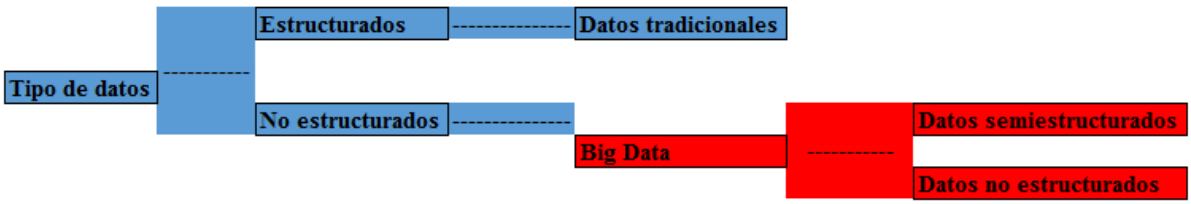


Figura 2. Datos Estructuras & No Estructurados

Fuente: Autor

2.3.1. Datos estructurados

Son aquellos datos con formato y campos fijos, en el que el formato es anticipadamente definido, para ser almacenados en bases de datos relacionales; este tipo de datos guardan un orden específico lo que facilita trabajar con ellos.

2.3.2. Datos semiestructurados

Son aquellos datos que no tienen formatos fijos, pero que contienen etiquetas, marcadores o separadores que permiten entenderlos; se procesan a base de reglas para extraer la información en piezas. Los lenguajes XML y HTML son ejemplos de texto con etiquetas.

2.3.3. Datos no estructurados

Son aquellos datos que no tienen formatos predefinidos, es decir no tienen estructura uniforme. Se tiene poco o ningún control sobre su estructura los correos electrónicos, mensajes instantáneos SMS, WhatsApp, Viber, fotos, audios, videos entre otros.

2.4. TIPOS DE ANÁLISIS DE DATOS

Según ISACA (Information Systems Audit and Control Association - Asociación de Auditoría y Control de Sistemas de Información) en 2011, la analítica de datos "implica los procesos y actividades diseñados para obtener y evaluar datos para extraer información útil". La analítica de datos es la ciencia que examina datos en bruto o crudos para obtener conclusiones acerca de la información contenida en ellos. Existen muchas herramientas de software diseñadas para analítica de datos (herramientas como CRM (Customer Relationship Management - Administración de la Relación con los Clientes), OLAP (On-Line Analytical Processing), tableros de control, entre otros), en las que se utilizan técnicas de consultas (quering), reportes (reporting), visualización, lógica difusa, minería de datos, análisis de datos predictivos, streaming de audio, video o fotografía, entre otros. La analítica de datos está influenciada por todo tipo de dispositivos como GPS, sensores, dispositivos tipo internet de las cosas, chips NFC y RFID, códigos de barras, códigos QR; además, por medios sociales como Twitter, Facebook, blogs, es decir la WEB 2.0. En general la tendencia SoLoMo (Social, Localización, Movilidad)

genera gran cantidad de información en forma de mapas, coordenadas, estados, etiquetas, rutas; almacenados en soluciones cloud cada vez con mayor frecuencia y en más organizaciones alrededor del mundo.

Figura 3. Tipos de Datos Big Data

2.4.1. Web and Social Media

Incluye contenido web e información que es obtenida de las redes sociales como Facebook, Twitter, LinkedIn, etc, blogs.

2.4.2. Machine-to-Machine (M2M)

M2M se refiere a las tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún evento en particular (velocidad, temperatura, presión, variables meteorológicas, variables químicas como la salinidad, etc.) los cuales transmiten a través de redes alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estos eventos en información significativa.

2.4.3. Big Transaction Data

Incluye registros de facturación, en telecomunicaciones registros detallados de las llamadas (CDR), etc. Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

2.4.4. Biometrics

Información biométrica en la que se incluye huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc. En el área de seguridad e inteligencia, los datos biométricos han sido información importante para las agencias de investigación.

2.4.5. Human Generated

Las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc.

2.5. CARACTERÍSTICAS DE BIG DATA

Las características más importantes del Big Data perfectamente se pueden clasificar en cuatro magnitudes, más conocidas como las cuatro V del Big Data, relativas a volumen, variedad, velocidad y veracidad. A estas cuatro V, podemos añadir tres más, como pueden ser la de Viabilidad y Visualización. Pero si hablamos de V en Big Data no podemos dejar pasar la principal característica del análisis de datos que es la V de Valor de los datos, ya no es de las tradicionales cuatro V de Big Data, sino de las 7 "V" del Big Data:

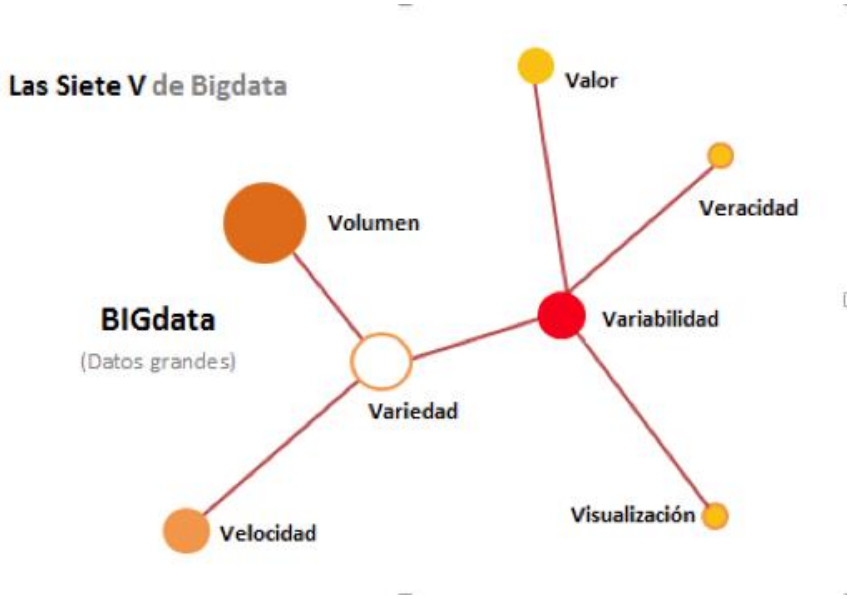


Figura 4. Diseño 7V Big Data.

2.5.1. Volumen

Es la primera característica intrínseca al Big Data, si se usa este término se hace referencia a un alto volumen datos, inmanejable sin un sistema de almacenamiento y procesamiento.

Millones de transacciones al día en Internet, trillones de fotos, billones de usuarios compartiendo intereses y haciendo compras, almacenamiento de archivos digitales que sobrepasan todo el conocimiento humano pueden dar una idea del inmenso volumen

2.5.2. Velocidad

Tras asimilar el gran volumen de datos disponibles gracias a la tecnología, la segunda característica de las 7v del Big Data es la velocidad. El asunto toma ya dos dimensiones, no se

trata de cuánta información sino de cuándo se genera, cuánto tarda en almacenarse y en procesarse, al punto tal que, si no es a tiempo real, ya es tarde.

2.5.3. Variedad

Mensajes por correo electrónico, datos en formularios de contacto, interacciones en redes sociales, comentarios en YouTube, transacciones en línea, compras en e–commerce, comportamiento de usuarios al navegar una página de Internet, mensajes SMS, histórico de compras en un CRM, frecuencia de compras en un establecimiento físico, cantidad de visitas, series preferidas en televisión por demanda. Sumados al volumen y a la velocidad, la variedad de formatos y tipos de datos exigen una tarea de categorización profunda y amplia.

2.5.4. Variabilidad

La variedad se refiere a las formas, tipos y fuentes en las que se registran los datos. Estos datos pueden ser datos estructurados y fáciles de gestionar como son las bases de datos, o datos no estructurados, entre los que se incluyen documentos de texto, correos electrónicos, datos de sensores, audios, vídeos o imágenes que tenemos en nuestro dispositivo móvil, hasta publicaciones en nuestros perfiles de redes sociales, artículos que leemos en blogs, las secuencias de clic que hacemos en una misma página, formularios de registro e infinidad de acciones más que realizamos desde nuestro Smartphone, Tablet y ordenador.

2.5.5. Veracidad

Cuando hablamos de veracidad nos referimos a la incertidumbre de los datos, es decir, al grado de fiabilidad de la información recibida. Es necesario invertir tiempo para conseguir datos de calidad, aplicando soluciones y métodos que puedan eliminar datos imprevisibles que puedan surgir como datos económicos, comportamientos de los consumidores que puedan influir en las decisiones de compra.

2.5.6. Visualización

Modo en el que los datos son presentados. Una vez que los datos son procesados (los datos están en tablas y hojas de cálculo), necesitamos representarlos visualmente de manera que sean legibles y accesibles, para encontrar patrones y claves ocultas en el tema a investigar.

2.5.7. Valor

El dato no es valor. Tampoco tienes valor por el mero hecho de recopilar gran cantidad de información. El valor se obtiene de datos que se transforman en información; esta a su vez se convierte en conocimiento, y este en acción o en decisión. El valor de los datos está en que sean accionables, es decir, que los responsables de la empresa puedan tomar una decisión (la mejor decisión) en base a estos datos.

2.6. BIG DATA ANALYTICS

El término big data se refiere al almacenamiento digital de información que tienen un gran volumen, velocidad y variedad. Big Data Analytics es el proceso de usar software para descubrir tendencias, patrones, correlaciones u otras ideas útiles en esos grandes almacenes de datos.

El análisis de datos no es nuevo. Ha existido durante décadas en la forma de software de inteligencia empresarial y minería de datos. Con el paso de los años, ese software ha mejorado de forma espectacular, por lo que puede manejar volúmenes de datos mucho más grandes, ejecutar consultas más rápidamente y ejecutar algoritmos más avanzados. La firma de investigación de mercado Gartner categoriza las herramientas de big data y analytics en cuatro categorías diferentes.

2.6.1. Análisis descriptivo

Estas herramientas les dicen a las compañías lo que sucedió. Crean informes simples y visualizaciones que muestran lo que ocurrió en un momento particular o durante un período de tiempo. Estas son las herramientas analíticas menos avanzadas.

2.6.2. Diagnóstico analítico

Las herramientas de diagnóstico explican por qué sucedió algo. Más avanzadas que las herramientas descriptivas de informes, les permiten a los analistas profundizar en los datos y determinar la raíz de las causas para una situación dada.

2.6.3. Análisis Predictivo

Entre las herramientas de big data analytics más populares disponibles en la actualidad, las herramientas de análisis predictivo utilizan algoritmos altamente avanzados para pronosticar lo que podría suceder a continuación. A menudo, estas herramientas hacen uso de la inteligencia artificial y la tecnología.

2.6.4. Análisis prescriptivo

Un paso por encima del análisis predictivo, el análisis prescriptivo les dice a las organizaciones qué deben hacer para lograr un resultado deseado. Estas herramientas requieren capacidades de aprendizaje automático muy avanzadas, y pocas soluciones en el mercado actual ofrecen verdaderas capacidades preceptivas.

2.7. GENERALIDADES DE SOCIAL BIG DATA

Social Big Data se refiere al análisis de la gran cantidad de información que se produce en las redes sociales. Algunos datos que muestran esto son:

- **Google:** más de 3.7 millones de búsquedas por minuto.
- **Facebook:** más de 977.000 logins artículos y 34.000 «Me gusta» por minuto.
- **YouTube:** más de 4.3 millones de videos vistos en un minuto.
- **Twitter:** más de 481.000 tuits por minuto.
- **WhatsApp:** más de 38 millones de mensajes enviados en un minuto.
- **Correos electrónicos:** más de 187 millones emails enviados por minuto.
- **NetFlix:** más de 266.000 de horas de video vistas en un minuto.

2.7.1. Beneficios del Social Big Data para las empresas

El Social Big Data es un concepto que se reúne una serie de funciones las cuales ayudan a interpretar los datos.

2.7.2. Conocer el comportamiento del consumidor

El marketing puede aprovechar la gran cantidad de datos de las redes sociales para conocer con precisión la conducta y preferencias de los consumidores en Internet.

2.7.3. Ajustar la comunicación con el cliente

El análisis de la enorme cantidad de datos de las redes sociales permite ajustar la comunicación con los clientes, determinando el momento preciso y el contexto adecuado.

2.7.4. Ventaja competitiva

El Big Data permite a las empresas encontrar una ventaja competitiva respecto a la competencia al procesar esta cantidad inmensa de datos.

2.7.5. Planificar y Anticipar

Los análisis de las conversaciones en Social Big Data permiten identificar las tendencias del consumidor, planificar nuevos productos.

2.7.6. Realizar innovación

El análisis del Social Big Data permite obtener otros insights del mercado, necesarios para el proceso de innovación.

2.7.7. Mejorar el servicio al cliente

La escucha de la gran cantidad de conversaciones ayuda a prestar un mejor servicio y a integrar los canales sociales con los canales de atención al cliente, mejorando enormemente la experiencia de uso.

2.8. HERRAMIENTAS DE ANALÍTICA SOCIAL

No monitorizar las acciones que se desarrollan en las redes sociales es algo parecido a navegar en el mar sin rumbo fijo y tener muy pocas posibilidades de llegar al objetivo final: alcanzar la orilla.

Toda acción que se lleve a cabo en Social Media debería ser medida adecuadamente y en base a los objetivos planteados. Acción por acción, campaña por campaña, estrategia por estrategia, cualquier paso debería darse en constante monitorización.

La mayor parte de las redes sociales ofrece sus propias herramientas de analítica. En casos como el de Facebook, éstas son tremendamente completas y serán más que suficientes para la mayor parte de los pequeños negocios que tengan presencia en esta plataforma.

Sin embargo, en ocasiones, ya sea por el tamaño de la empresa, los objetivos planteados o porque la red social cuya actividad se quiere monitorizar no proporcione los informes adecuados, habrá que recurrir a herramientas de monitorización externas que ayuden en la tarea.

2.8.1. Hootsuite

Sin duda es una de las herramientas de gestión multiplataforma más populares. Entre otras funciones, Hootsuite también proporciona completos informes para *Facebook*, *Twitter* e *Instagram*, tanto en general como de cada una de las publicaciones que se haga en ella.

Hootsuite presenta adecuadamente toda la información en claras gráficas fácilmente customizables y exportables. Además, tiene herramientas útiles para equipos de Social Media. Por ejemplo, señala cuánto tiempo tarda en darse respuesta a los comentarios de los usuarios e incluso monitoriza el tiempo de resolución de las cuestiones planteadas por mensaje directo.

2.8.2. Sprout Social

Se trata de otra de las principales herramientas para todo profesional del Social Media, que proporciona analítica en profundidad y resume la actividad gráficos fáciles de entender. Mide la actividad en redes sociales como. *Facebook*, *Twitter*, *Instagram* y *LinkedIn* y es capaz de compararla con lo que están realizando los máximos competidores.

Resulta interesante la manera en la que establece comparaciones entre las publicaciones orgánicas y aquellas que han sido pagadas. Sprout Social se puede probar de forma gratuita durante 30 días, pero los planes de pago son más caros que en otras herramientas ya que parten de 99 dólares al mes.

2.8.3. Unión Metrics

Esta herramienta monitoriza *Facebook*, *Twitter* e *Instagram* y se centra en el valor del contenido, teniendo en cuenta por ejemplo a qué horas del día es más activa la comunidad en torno a la marca o quiénes son los usuarios más influyentes.

La plataforma proporciona análisis de la competencia y recomendaciones en cuanto al contenido que se debe publicar y puede llegar a proporcionar informes geolocalizados, y diferenciados por idioma, en los planes más completos.

2.8.4. Rival IQ

Es una de las más completas, pero también de las de mayor coste. Permite monitorizar publicaciones, me gustas y comentarios en casi todas las plataformas sociales, como *Facebook*, *Instagram*, *Twitter* e incluso *Pinterest*. También puede emplearse para comparar toda la información con *Google Analytics* y saber de dónde llega el tráfico a los perfiles sociales.

Rival IQ permite establecer gráficas personalizadas para cada uno de los perfiles sociales, así como alertas y envíos de informes periódicamente. Se puede probar durante 14 días de forma gratuita, pero los planes de pago parten de los 199 dólares mensuales.

2.8.5. Awario

Se trata de una herramienta de monitorización que proporciona analíticas muy potentes. Mide las menciones que se producen, establece el «share of voice» de la compañía y proporciona datos como el crecimiento y el alcance, pudiendo segmentar por lenguaje, ubicación.

También tiene en cuenta el tipo de usuarios que emplean determinadas palabras clave o el nombre de la compañía, y es capaz de identificar un listado de personas influyentes para la marca. Con Awario, además, se pueden crear alertas y establecer criterios de comparación con marcas de la competencia.

2.8.6. Snaplytics

Si buscas una herramienta que monitorice tu actividad en *Snapchat* quizá debas decantarte por esta, que, además, también ofrece informes completos sobre *Instagram*. Permite analizar las Stories, el contenido efímero en una y otra plataforma, y establecer niveles de engagement, de adquisición, de alcance presentados en interesantes reportes que se pueden exportar en formato base de datos.

2.8.7. Squarelovin

Está más indicada específicamente para monitorizar acciones en Instagram: me gustas, engagement, mejores horas para publicar, mejores filtros, hashtags más efectivos… Proporciona información cada hora sobre cómo están funcionando cada uno de los posts que se publican. Se puede crear una cuenta gratuita, aunque será necesario pagar para acceder a algunas opciones.

2.9. ANÁLISIS DE SENTIMIENTOS

Es un modelo conceptual establecido para la clasificación de emociones. Son varias las teorías que pretenden establecer categorías a los sentimientos expresados por las personas a partir de sus acciones, interpretaciones, comportamientos u opiniones. La carga subjetiva implícita en estos procesos impide definir con precisión que se entiende por sentimiento o emoción. Teniendo en cuenta la complejidad de asignar una clasificación binaria a los sentimientos expresados por los seres humanos, mediante sistemas automatizados que se

fundamentan en la aplicación de algoritmos matemáticos. O sea que, el término Minería de Opinión se refiere al proceso de detectar expresiones subjetivas en textos.

Las emociones y los sentimientos son cualidades que representan el grado de afectividad de los seres humanos, posibilitando reflejar el estado de ánimo. Al expresarnos manifestamos nuestro estado emocional, y en función de este adoptamos determinadas actitudes. Por tanto, al comentar sobre un tema en específico, ya sea de forma presencial o por escrito, se establece una comunicación emocional donde quedan expresadas las intenciones del emisor, hablante o escritor.

Los términos más utilizados en la literatura para denominar la clasificación de documentos basada en la opinión son los siguientes:

- Análisis de sentimientos
- Minería de Opinión
- Análisis de la subjetividad
- Brandmonitoring
- Buzzmonitoring
- Conversation mining
- Online consumer intelligence
- User generated content.

Una característica determinante al realizar técnicas de Minería de Opinión, es tener en cuenta el contexto desde el cual el usuario se manifiesta, debido a que existen palabras que presentan una orientación semántica en sí misma, como es el caso de los términos pobre o excelente. Un término con evidentes implicaciones negativas puede tener una interpretación positiva, en función del contexto en el que se utiliza. Estos términos son reconocidos en la literatura sobre el tema como expresiones polares.

2.9.1. Aplicación de la minería de opinión

La técnica Minería de Opinión engloba diferentes tareas, encaminadas al procesamiento de texto no estructurado, para ello se persigue dar una valoración cuantitativa a expresiones subjetivas asociadas a opiniones y sentimientos, además se busca identificar el grado de

polaridad positivo, negativo o neutro en el que se califica a todo tipo de “entidades”. Las tareas que define el autor son las siguientes:

2.9.2. Clasificación de la subjetividad

“Tarea cuyo objetivo es la identificación de fragmentos de texto que poseen un significado o una carga subjetiva, expresada por parte de la persona que ha escrito el texto, ya sea una opinión, la expresión de un sentimiento, etc.”

Esta tarea permite distinguir entre comentarios que expresan una opinión sobre un asunto determinado, de aquellas que se limitan a narrar un hecho sin brindar un punto de vista. La clasificación de la subjetividad es considerada con frecuencia, como un paso previo antes de realizar otras actividades de la Minería de Opinión como es el caso del cálculo de la polaridad.

2.9.3. Clasificación de la intensidad

Esta tarea pretende clasificar los textos de entrada de acuerdo a la intensidad emocional expresada. De esta manera, la mayoría de las aproximaciones que abordan este problema trabajan con lo que se denomina clasificación en tres clases de intensidad: positivo, neutro y negativo.

Los estudios se basan en que hay palabras que poseen mayor carga emotiva que otras, este autor pone como ejemplo los términos cáncer y refriado, que, aunque ambos tienen obvias implicaciones negativas, mientras que el primero expresa malestar o incomodidad.

2.9.4. Minería de Opinión basada en tópicos/características

Generalmente, este tipo de sistemas suelen evaluar documentos que recogen opiniones sobre productos o servicios donde ciertos aspectos de esos productos o servicios condicionan más que otros la carga afectiva global de la opinión¨.

Se refiere a la capacidad de un sistema de determinar las distintas características del producto tratadas en la opinión escrita por el usuario, y para cada una de esas características mencionadas en la opinión, ser capaces de extraer una polaridad.

2.9.5. Clasificación de la polaridad

Tarea que pretende, como última finalidad, clasificar fragmentos de texto, que pueden ser desde documentos hasta sintagmas, en positivo o negativo dependiendo de su significado emocional.

La presente investigación se centra en la tarea de clasificar la polaridad por tanto es válido profundizar en su ejecución. La clasificación de la polaridad se aborda desde dos aproximaciones, que permiten capturar el significado emocional de un texto; a estas metodologías se les denomina como supervisadas y no supervisadas. La primera de ellas se basa en el aprendizaje máquina o aprendizaje automático, es decir en el entrenamiento de un sistema con términos que han sido clasificados previamente.

2.10. ARQUITECTURA DE BIG DATA

La arquitectura de Big Data se apoya en componentes que se organizan en torno a capas.

Figura 5. Arquitectura de referencia de Big Data propuesta por Sunil Soares.

2.10.1. Ciclo de vida Big Data

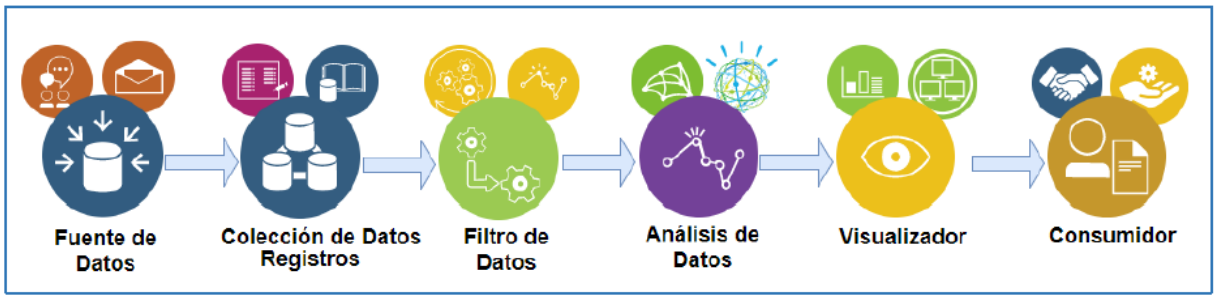


Figura 6. Ciclo de Vida Big Data (Fuente: Defining the Big Data)

En la figura (Ver Figura 6) se muestra el ciclo de vida de Big Data, el cual parte desde una fuente de datos, de allí se sacan los datos que se necesitan y se crea una colección de registros, con dicha colección se realiza un filtro de datos para luego analizarlos con diferentes algoritmos y obtener un resultado del análisis, esta información es procesada de manera que pueda ser visualizada de una forma entendible para el usuario, quien es finalmente el que consume el servicio y saca ventaja de ello.

2.10.2. Usos de Big Data

	Descubrimiento Científico	Nueva Tecnología	Manufactura y Transporte	Servicios Personales Campañas	Medio Ambiente Infraestructura	Cuidado de la Salud
Ciencia	+++++	++++	+	-	++	+++
Telecomunicaciones	+	++++	++	+	++++	+
Industria	++	++++	+++++	-	-	++
Negocio	+	+++	++	-	+	++
Vida Medio Ambiente	++	++	++	++	+++++	+
Social Media	+	++	-	++++	++	-
Cuidado de la Salud	+++	++	-	-	++	+++++

Figura 7. Usos de Big Data (Fuente: Defining the Big Data Architecture Framework

La tabla de usos de Big Data (Ver Figura 7) Se realiza una comparación entre el origen de los datos de Big Data (Columnas) y el objetivo de uso de los datos de Big Data (Filas), el símbolo más (+) representa el uso que se le ha dado al dato de acuerdo con su origen; el símbolo menos (-) representa que no tiene uso el dato.

2.11. COMPONENTES DE SOPORTE A LAS TECNOLOGÍAS BIG DATA

Hadoop es una estructura de software de código abierto para almacenar datos y ejecutar aplicaciones en clústeres de hardware comercial. Proporciona almacenamiento masivo para cualquier tipo de datos, enorme poder de procesamiento y la capacidad de procesar tareas o trabajos concurrentes virtualmente ilimitados.

Hadoop está inspirado en el proyecto de Google File System (*GFS*) y en el paradigma de programación *MapReduce*, el cual consistió en dividir dos tareas (*Mapper Reducer*) la manipulación de los datos distribuidos a nodos de un cluster logrando un alto paralelismo en el procesamiento.

Hadoop está compuesto de tres piezas *Hadoop Distributed File System (HDFS), Hadoop MapReduce y HadoopCommon.*

Historia de hadoop

A medida que la World Wide Web creció a finales de los 1900 y principios de los 2000, se crearon buscadores (o motores de búsqueda) e índices para ayudar a localizar información relevante dentro de contenido basado en texto. En sus primeros años, los resultados de las búsquedas eran entregados por humanos. Pero a medida que la Web creció de docenas a millones de páginas, se requirió de la automatización. Se crearon los rastreadores Web, muchos como proyectos dirigidos por universidades, y entonces se iniciaron las primeras compañías de buscadores (Yahoo, AltaVista, etc.).

Uno de estos proyectos fue un buscador Web de código abierto llamado Nutch – idea original de Doug Cutting y Mike Cafarella. Deseaban generar resultados de búsquedas en la Web a mayor velocidad distribuyendo datos y cálculos en diferentes computadoras de modo que se pudieran procesar múltiples tareas de manera simultánea. Durante este tiempo, estaba en progreso otro proyecto de buscador llamado Google. Éste se basaba en el mismo concepto –

almacenar y procesar datos de manera distribuida y automatizada de modo que se pudieran generar resultados de búsquedas en la Web a mayor velocidad.

En 2006, Cutting se unió a Yahoo y se llevó con él el proyecto Nutch, así como también ideas basadas en los trabajos iniciales de Google con la automatización del almacenaje y procesamiento de datos distribuidos. El proyecto Nutch fue dividido – la parte del rastreador Web se mantuvo como Nutch y la parte de cómputo y procesamiento distribuido se convirtió en Hadoop (en honor del elefante de juguete del hijo de Cutting). En 2008, Yahoo presentó Hadoop como proyecto de código abierto. Hoy día, la estructura y el ecosistema de tecnologías de Hadoop son gestionados y mantenidos por la Apache Software Foundation (ASF) sin fines de lucro, que es una comunidad global de programadores de software y otros contribuyentes.

A continuación, se describen las características de importancia al utilizar Hadoop.

- **Capacidad de almacenar y procesar enormes cantidades de cualquier tipo de datos, al instante:** con el incremento constante de los volúmenes y variedades de datos, en especial provenientes de medios sociales y la Internet de las Cosas (IoT), ésa es una consideración importante.
- **Poder de cómputo:** El modelo de cómputo distribuido de Hadoop procesa big data a gran velocidad. Cuantos más nodos de cómputo utiliza usted, mayor poder de procesamiento tiene.
- **Tolerancia a fallos:** El procesamiento de datos y aplicaciones está protegido contra fallos del hardware. Si falla un nodo, los trabajos son redirigidos automáticamente a otros modos para asegurarse de que no falle el procesamiento distribuido. Se almacenan múltiples copias de todos los datos de manera automática.
- **Flexibilidad:** A diferencia de las bases de datos relacionales, no tiene que procesar previamente los datos antes de almacenarlos. Puede almacenar tantos datos como desee y decidir cómo utilizarlos más tarde. Eso incluye datos no estructurados como texto, imágenes y videos.
- **Bajo costo:** La estructura de código abierto es gratuita y emplea hardware comercial para almacenar grandes cantidades de datos.
- **Escalabilidad:** Puede hacer crecer fácilmente su sistema para que procese más datos son sólo agregar nodos. Se requiere poca administración.

2.11.1. Hadoop Distributed File System (HDFS)

El tema de la computación de alto rendimiento (HPC o High Performance Computing) lleva años dando vueltas y soluciones ya maduras y establecida (Tanto gestores de colas como Condor, Oracle Grid Engine, Torquen en la parte de cluster como Globus o Glite en la parte de grid computing).

Lo que realmente aporta Hadoop es una capacidad de gestionar grandes cantidades de datos. Los cluster tradicionales están orientados a tener que dar mucha potencia de cálculo gestionando relativamente poco espacio en disco, pero ¿qué pasa cuando una base de datos tiene 100 TB o 1 PB? En estos casos se necesita algo más potente como Hadoop.

El HDFS (Hadoop Distributed File System) es quizás el componente principal de Hadoop, ya que permite crear sistemas de ficheros empleando servidores "Commodity" ofreciendo redundancia, capacidad y rendimiento (solo para ficheros muy grandes) y lo mejor de todo es que estos servidores Commodity son los que hacen la computación, permitiendo el paradigma de "llevar los datos a la computación", uno de los factores principales del rendimiento de Hadoop.

Los datos en el cluster de Hadoop son divididos en pequeñas piezas llamadas bloquees y distribuidas a través del cluster. De esta manera, las funciones map y reduce pueden ser ejecutadas en pequeños subconjuntos y esto provee de la escalabilidad necesaria para el procesamiento de grandes volúmenes.

La siguiente figura ejemplifica como los bloques de datos son escritos hacia HDFS. Observe que cada bloquee es almacenado tres veces y al menos un bloque se almacena en un diferente Rack para lograr redundancia.

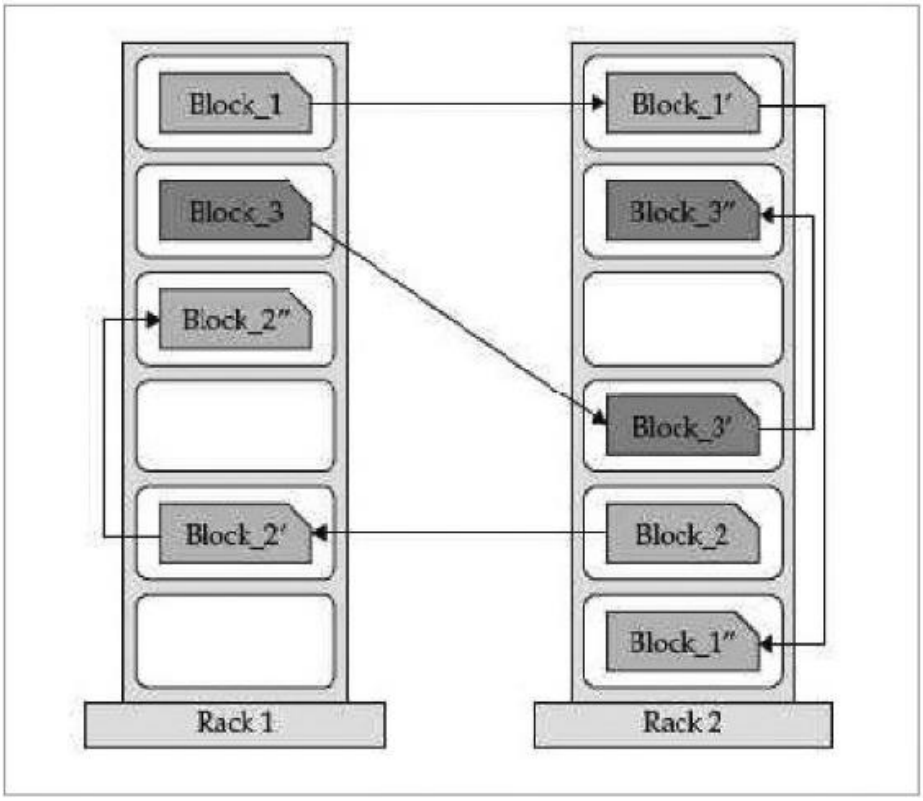


Figura 8. Demostración bloque y Rack.

NameNode

NameNode es el nodo maestro en la arquitectura Apache Hadoop HDFS que mantiene y administra los bloques presentes en los DataNodes (nodos esclavos). NameNode es un servidor de alta disponibilidad que administra el espacio de nombres del sistema de archivos y controla el acceso a los archivos por parte de los clientes. Discutiré esta característica de alta disponibilidad de Apache Hadoop HDFS en mi próximo blog. La arquitectura HDFS está construida de tal manera que los datos del usuario nunca residen en el NameNode. Los datos residen solo en DataNodes.

Funciones de NameNode

Es el demonio maestro que mantiene y gestiona los DataNodes (nodos esclavos) Registra los metadatos de todos los archivos almacenados en el clúster, por ejemplo, la ubicación de los bloques almacenados, el tamaño de los archivos, los permisos, la jerarquía, etc. Hay dos archivos asociados con los metadatos:

- **FsImage:** contiene el estado completo del espacio de nombres del sistema de archivos desde el inicio de NameNode.
- **EditLogs:** contiene todas las modificaciones recientes realizadas en el sistema de archivos con respecto a la imagen Fs más reciente.
- Registra cada cambio que tiene lugar en los metadatos del sistema de archivos. Por ejemplo, si un archivo se elimina en HDFS, NameNode lo registrará inmediatamente en EditLog.
- Regularmente recibe un Heartbeat y un informe de bloque de todos los DataNodes en el clúster para garantizar que los DataNodes estén activos.
- Mantiene un registro de todos los bloques en HDFS y en qué nodos se encuentran estos bloques.
- NameNode también es responsable de cuidar el factor de replicación de todos los bloques que discutiremos en detalle más adelante en este blog tutorial HDFS.
- En caso de falla de DataNode, NameNode elige nuevos DataNodes para nuevas réplicas, equilibra el uso del disco y gestiona el tráfico de comunicación a los DataNodes.

DataNode

Los DataNodes son los nodos esclavos en HDFS. A diferencia de NameNode, DataNode es un hardware básico, es decir, un sistema no costoso que no es de alta calidad o alta disponibilidad. DataNode es un servidor de bloques que almacena los datos en el archivo local ext3 o ext4.

Funciones de DataNode

- Estos son demonios esclavos o procesos que se ejecutan en cada máquina esclava.

Los datos reales se almacenan en DataNodes.

- Los DataNodes realizan las solicitudes de lectura y escritura de bajo nivel de los clientes del sistema de archivos.
- Envían latidos al NameNode periódicamente para informar el estado general de HDFS, de forma predeterminada, esta frecuencia se establece en 3 segundos.

Nombre de nodo secundario

Además de estos dos demonios, hay un tercer demonio o un proceso llamado Secondary NameNode. El NameNode secundario funciona simultáneamente con el NameNode primario como un demonio auxiliar. Y no se confunda acerca de que el NameNode secundario es un NameNode de respaldo porque no lo es.

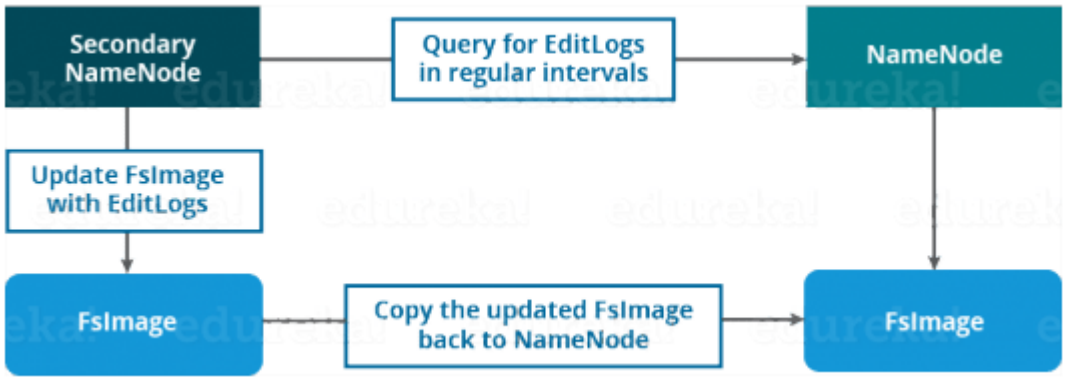


Figura 9. Arquitectura DataNode.

Funciones del nodo de nombre secundario

- El NameNode secundario es uno que lee constantemente todos los sistemas de archivos y metadatos de la RAM del NameNode y los escribe en el disco duro o en el sistema de archivos.
- Es responsable de combinar EditLogs con FsImage del NameNode.
- Descarga los EditLogs del NameNode a intervalos regulares y se aplica a FsImage. La nueva FsImage se copia de nuevo en NameNode, que se usa cada vez que se inicia NameNode la próxima vez.

Por lo tanto, NameNode secundario realiza puntos de control regulares en HDFS. Por lo tanto, también se llama CheckpointNode.

Bloques

Ahora, como sabemos, los datos en HDFS están dispersos en los DataNodes como bloques. Echemos un vistazo a qué es un bloque y cómo se forma.

Los bloques no son más que la ubicación continua más pequeña en su disco duro donde se almacenan los datos. En general, en cualquiera de los sistemas de archivos, almacena los datos como una colección de bloques. Del mismo modo, HDFS almacena cada archivo como bloques

que están dispersos por todo el clúster de Apache Hadoop. El tamaño predeterminado de cada bloque es de 128 MB en Apache Hadoop 2. x (64 MB en Apache Hadoop 1 .x) que puede configurar según sus necesidades.

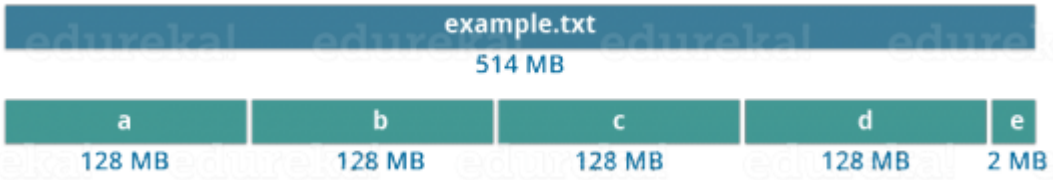


Figura 10. Ejemplo Particiones de Bloques.

No es necesario que en HDFS, cada archivo se almacene en múltiplos exactos del tamaño de bloque configurado (128 MB, 256 MB, etc.). (Ver Figura 10) "example.txt" de tamaño 514 MB como se muestra en la figura anterior. Se está utilizando la configuración predeterminada del tamaño de bloque, que es de 128 MB. Entonces, ¿cuántos bloques se crearán? 5, a la derecha. El primero cuatro bloques será de 128 MB. Pero, el último bloque tendrá solo 2 MB de tamaño.

Gestión de replicación

HDFS proporciona una forma confiable de almacenar grandes datos en un entorno distribuido como bloques de datos. Los bloques también se replican para proporcionar tolerancia a fallas. El factor de replicación predeterminado es 3, que nuevamente es configurable. Entonces, como puede ver en la figura a continuación, cada bloque se replica tres veces y se almacena en diferentes DataNodes (considerando el factor de replicación predeterminado):

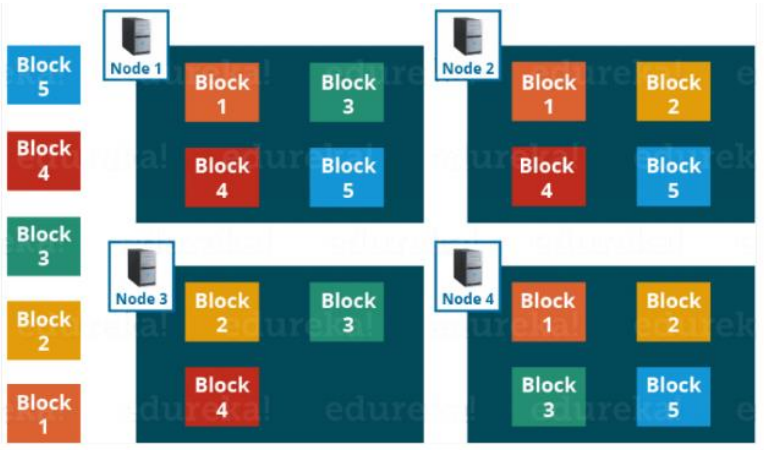


Figura 11. Replicación de Bloques.

Conciencia del estante

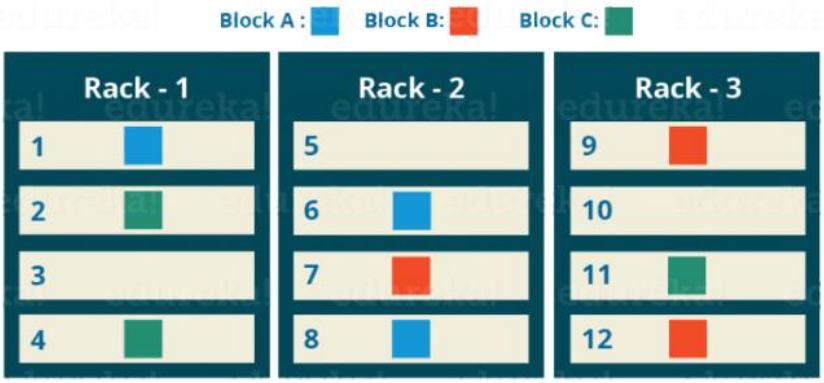


Figura 12. Algoritmo de Reconocimiento de Rack.

HDFS coloca la réplica y qué es el conocimiento del rack. Nuevamente, NameNode también garantiza que todas las réplicas no se almacenen en el mismo rack o en un solo rack. Sigue un algoritmo de conciencia de bastidor incorporado para reducir la latencia y proporcionar tolerancia a fallas. Teniendo en cuenta que el factor de replicación es 3, el Algoritmo de conocimiento del bastidor dice que la primera réplica de un bloque se almacenará en un bastidor local y las siguientes dos réplicas se almacenarán en un bastidor diferente (remoto) pero en un DataNode diferente dentro de ese (remoto) como se muestra en la figura anterior. Si tiene más réplicas, el resto de las réplicas se colocarán en DataNodes aleatorios siempre que no sea posible que haya más de dos réplicas en el mismo rack, si es posible.

2.11.2. Hadoop MapReduce

MapReduce es un framework que proporciona un sistema de procesamiento de datos paralelo y distribuido. Su nombre se debe a las funciones principales que son Map y Reduce, las cuales explicaremos a continuación. MapReduce está pensado para la solución práctica de algunos problemas que pueden ser paralelizados, pero se ha de tener en cuenta que no todos los problemas pueden resolverse eficientemente con MapReduce. MapReduce está orientado a resolver problemas con conjuntos de datos de gran tamaño, por lo que utiliza el sistema de archivos distribuido HDFS. El Framework MapReduce tiene una arquitectura maestra / esclavo. Cuenta con un servidor maestro o JobTracker y varios servidores esclavos o TaskTrackers, uno por cada nodo del clúster.

El JobTracker es el punto de interacción entre los usuarios y el framework MapReduce. Los usuarios envían trabajos MapReduce al JobTracker, que los pone en una cola de trabajos pendientes y los ejecuta en el orden de llegada. El JobTracker gestiona la asignación de tareas y delega las tareas a los TaskTrackers. Los TaskTrackers ejecutan tareas bajo la orden del JobTracker y también manejan el movimiento de datos entre la fase Map y Reduce.

Para ver las diferencias entre JobTracker y TaskTracker vamos a ver las características de cada uno.

JobTracker

- Capacidad para manejar metadatos de trabajos
- Estado de la petición del trabajo
- Estado de las tareas que se ejecutan en TaskTracker
- Decide sobre la programación
- Hay exactamente un JobTracker por cluster.
- Recibe peticiones de tareas enviadas por el cliente.
- Programa y monitoriza los trabajos MapReduce con TaskTrackers.

TaskTracker: Servidor esclavo de MapReduce

- Ejecuta las solicitudes de trabajo de JobTrackers
- Obtiene el código que se ejecutará
- Aplica la configuración especifica del trabajo
- Comunicación con el JobTracker:
- Envíos de la salida, finalizar tareas, actualización de tareas, etc.

Figura 13. Framework MapReduce

En la Figura 13, los componentes principales del Framework, se evidencia las funciones de map son ejecutadas por los TaskTrackers, mientras que las funciones de Reduce son ejecutadas por el JobTracker.

Los componentes software que presenta el Framework MapReduce, vamos a entrar en detalle acerca de cada una las funciones Map y Reduce. Utilizaremos MapReduce para abordar problemas que pueden ser resueltos utilizando las operaciones de Map y Reduce, estas funciones están definidas con respecto a datos estructurados en tuplas del tipo (clave, valor).

Map

La función Map recibe como parámetros un par de (clave, valor) y devuelve una lista de pares. Esta función se encarga del mapeo y se aplica a cada elemento de la entrada de datos, por lo que se obtendrá una lista de pares por cada llamada a la función Map. Después se agrupan todos los pares con la misma clave de todas las listas, creando un grupo por cada una de las diferentes claves generadas. No hay requisito de que el tipo de datos para la entrada coincida con la salida y no es necesario que las claves de salida sean únicas.

Map (clave1, valor1) –> lista (clave2, valor2)

La operación de Map se paraleliza, el conjunto de archivos de entrada se divide en varias tareas llamado FileSplit, ver Figura 2, el tamaño típico de bloque es de 128MB. Las tareas se

distribuyen a los nodos TaskTrackers, y estos a su vez pueden realizar la misma tarea si hiciera falta.

Figura 14. Operación Split.

Reduce

La función Reduce se aplica en paralelo para cada grupo creado por la función Map (). La función Reduce se llama una vez para cada clave única de la salida de la función Map. Junto con esta clave, se pasa una lista de todos los valores asociados con la clave para que pueda realizar alguna fusión para producir un conjunto más pequeño de los valores.

Reduce (clave2, lista(valor2)) –> lista(valor2)

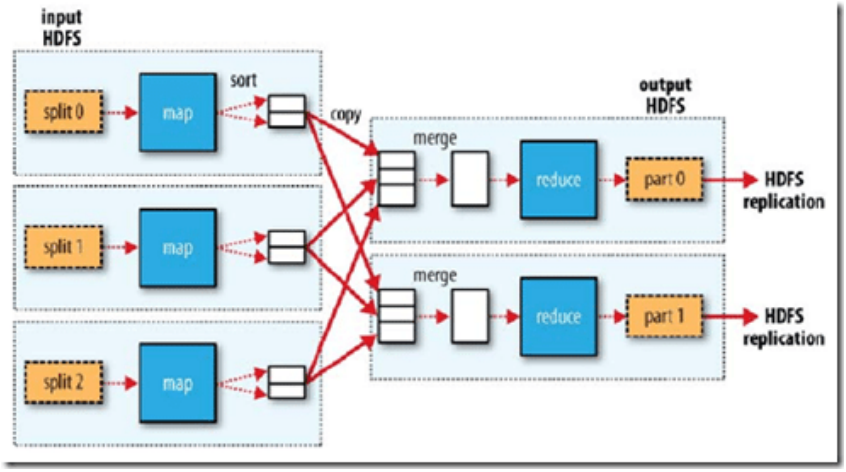


Figura 15. Funcionamiento de MapReduce

En la Figura 15 se ve que forma gráfica cómo funcionaría el proceso global de MapReduce.

Cuando se inicia la tarea reduce, la entrada se encuentra dispersa en varios archivos a través de los nodos en las tareas de Map. Los datos obtenidos de la fase Map se ordenan para que los pares clave-valor sean contiguos (fase de ordenación, sort fase), esto hace que la operación Reduce se simplifique ya que el archivo se lee secuencialmente. Si se ejecuta el modo distribuido estos necesitan ser primero copiados al filesystem local en la fase de copia. Una vez que todos los datos están disponibles a nivel local se adjuntan a una fase de adición, el archivo se fusiona (merge) de forma ordenado. Al final, la salida consistirá en un archivo de salida por tarea reduce ejecutada.

Por lo tanto, N archivos de entrada generará M mapas de tareas para ser ejecutados y cada mapa de tareas generará tantos archivos de salida como tareas Reduce hayan configuradas en el sistema.

3. TÉCNICAS Y TECNOLOGÍAS

3.1. Minería de Datos

La minería de datos es el proceso de búsqueda en grandes bases de datos para encontrar información útil que sirva para la toma de decisiones. También se utiliza el término en inglés «data mining».

Se puede entender como la tecnología y software utilizado para encontrar patrones de comportamiento dentro de la base de datos. La base fundamental de esto es que esos patrones ayuden a la toma de decisiones. Por ejemplo, podría ayudar a empresas, a conocer los patrones

de comportamiento de sus clientes. De manera que le facilitaría el establecimiento de estrategias para incrementar las ventas o reducir costes.

La minería de datos hace parte de un proceso más grande llamado Descubrimiento de Conocimientos en Base de Datos (KDD – Knowledge Discovery in Databalese), es un proceso de extracción de conocimiento. La minería de datos se restringe únicamente a la obtención de modelos.

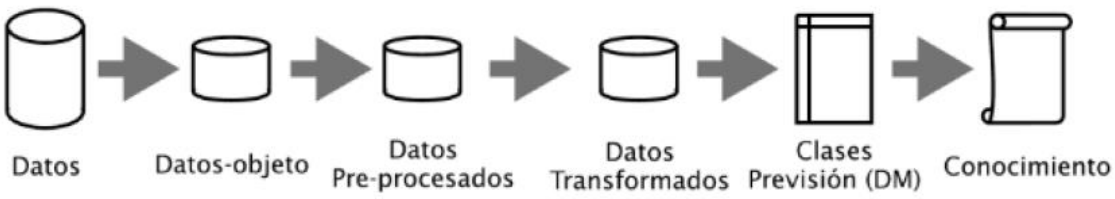


Figura 16. Proceso KDD. Fuente L. Vieira Braga

La figura 16 muestra el diagrama del proceso KDD, es importante notar que dentro del citado proceso la minería de datos es el penúltimo paso (previsión de modelos en minería de datos, en donde inicialmente se tienen los grandes datos de distintos medios, luego se seleccionan algunos mediante consultas, se procesan y transforman pensando en etapas siguientes, para luego pasar a la etapa de minería de datos en la que se establecen modelos y patrones, para que finalmente sean interpretados para presentarlos como el conocimiento obtenido.

En minería de datos se destacan dos metodologías, CRISP-DM (Cross Industry Standard Process for Data Mining), la cual es utilizada por Clementine – SPSS y SEMMA (Sample, Explore, Modify, Model, Access) empleada por SAS Enterprise Miner. Ambos sistemas pasan por las mismas etapas: recoger datos, depurarlos y analizarlos, para luego construir un modelo predictivo (con diferencias en las etapas de presentación e implementación).

Genéricamente un proyecto de minería de datos se realiza en las siguientes etapas:

- Definición del problema: conocer perfectamente lo que el estudio desea obtener, en común con todos los involucrados en el proyecto.

- Adquisición y evaluación de datos: adquirir, formatear y validar los datos, tomar muestras aleatorias.
- Extracción de características: identificar atributos que contribuyan en la solución, los que no se alteran deben salir del estudio, para producir un dataset representativo y confiable.
- Plan para el prototipo: desarrollo de hipótesis y del prototipo.
- Desarrollo del modelo: desarrollar modelos descriptivos y/o predictivos.
- Evaluación del modelo: considerar los resultados del prototipo.
- Implementación: presentación del producto final.
- Evaluación del retorno de la inversión: se evalúa si la inversión en el proyecto está generando utilidades para los inversionistas

3.Etapas de la minería de datos

Dentro de un proceso de minería de datos podemos encontrar cinco fases:

- **Objetivo y recolección de datos:** Lo primero de todo es centrarnos de en qué tipo de información queremos obtener. Imaginemos el ejemplo que un supermercado quiere conocer a qué hora del día es donde más asistencia de clientes hay. Este sería el objetivo y la información que quiere obtener el comercio en este caso.

- **Procesamiento y gestión de los datos:** Una vez que sabemos los datos que queremos recopilar ponemos a trabajar a los datos. Esta quizás sea la fase más complicada del proceso. Pues requiere seleccionar la muestra representativa sobre la que se va a realizar el análisis. Una vez escogida la muestra se debe analizar qué tipo de variables o modelo de regresión se va a realizar sobre la muestra.

- **Selección del modelo:** Está muy relacionado con la anterior fase. Se trata de crear un modelo o Algoritmo que nos arroje el mejor resultado posible. Para ello hay que hacer un análisis exhaustivo de las variables a incluir en el modelo. Esto se convierte en una tarea complicada ya que dependerá del tipo de información a analizar. Por ello, los mineros de datos llevan a cabo distintos exámenes del algoritmo como: regresión lineal, árbol de decisión, series temporales, red neuronal, etc.

- **Análisis y revisión de resultados:** Básicamente es analizar los resultados para comprobar si arrojan una explicación lógica. Explicación que facilite la toma de decisiones en base a la información suministrada por los resultados.

- **Actualización del modelo:** El último paso del proceso sería la actualización del modelo. Es muy importante que se vaya haciendo con el paso del tiempo para que no quede obsoleto. Las variables del modelo podrían pasar a ser no significativas y por tanto se requiere un control periódico del mismo.

3.2. Minería de Opinión

La minería de opinión también se conoce como, *análisis de opiniones, análisis de sentimientos, extracción de opinión, minería de sentimiento, análisis de subjetividad, análisis de emociones, o en inglés sentiment analysis u opinion mining* sin que haya diferencia entre estos términos.

La minería de opinión o análisis de sentimientos se ha estado estudiando a partir del año 2000.

Se ha abordado con técnicas de aprendizaje automático (Pang, Lee y Vaithyanathan, 2002) o no supervisado (Turney, 2002). Con estos trabajos pioneros en el área se determinó la polaridad de opiniones, aplicándolas a las opiniones sobre películas, posteriormente, por medio de técnicas combinadas se han realizado estudios en el mismo tema y desde luego en otros dominios (PlaFerran y Hustado Lluis, 2013).

El análisis de sentimientos surge de la necesidad de clasificar la orientación o la opinión de lo manifestado en un documento, en el presente proyecto examinar la subjetividad del documento es relevante, y aplicar técnicas de clasificación no son suficientes.

El análisis de sentimientos (AS) clasifica los documentos en función de la polaridad de la opinión que se expresa, identificando las opiniones positivas, negativas y neutras. Se utiliza para determinar la polaridad de las opiniones del documento o frase, y en determinar si el documento

contiene opiniones. El análisis se realiza aplicando aprendizaje automático o por medio del enfoque semántico.

El AS ha sido aplicado a muchos dominios, la mayoría de ellos para documentos en inglés,

siendo muy reducido este tipo de investigaciones en idioma español. Las universidades

españolas, Universidad de Sevilla y su grupo ITALICA, y la Universidad de Jaén con el grupo SINAI (Sistemas Inteligentes de Acceso a la Información) han realizado sendos trabajos, sin embargo, ésta última realizando el AS en lengua árabe también.

3.3. Microblogging

Un servicio de microblogueo del inglés Microblogging, también conocido como nanoblogueo, permite a sus usuarios enviar y publicar mensajes breves,2 generalmente solo de texto. Las opciones para el envío de los mensajes varían desde sitios web, a través de SMS, mensajería instantánea o aplicaciones ad hoc.

La principal y más popular característica de estos servicios es su sencillez y capacidad de sintetización, porque en la mayoría de sistemas de microblogueo el tope de escritura son alrededor de 140 caracteres. En esos 140 caracteres se pueden contar desde qué se está haciendo, interactuar con otros usuarios mediante respuestas y mensajes privados, anunciar cosas, promocionarse, hacer o mantener amistades y redes de amigos, encontrar trabajo, y otros.

3.4. Facebook

Facebook es una de las redes sociales más usadas en el mundo. Este proyecto, que en un principio era una plataforma o sitio web para estudiantes de la Universidad de Harvard, se convirtió en un proyecto exitoso que se extendió por todo el mundo. Tiene más de 1350 millones de usuarios alrededor del planeta, y sigue conquistando nuevos mercados gracias a los servicios que ofrece para individuos o empresas naturales o jurídicas. Por ello, en este artículo podrás conocer cómo se inició Facebook y conocer el éxito de esta red social hasta la actualidad

Facebook fue creado por Mark Zuckerberg, quien entonces era un estudiante de la Universidad de Harvard. En el 2003, Zuckerberg lanza un sitio web llamado Facemash.com, donde recopila los nombres y las fotografías de todos los estudiantes de dicha universidad. Sin embargo, por ese primer intento de directorio virtual, fue llevado ante los directores de Harvard,

pues se señalaba que había violado la política de privacidad y de propiedad intelectual de la universidad. Este sitio sólo estuvo disponible por algunas horas, y por esta acción suspendieron a Mark de clases, quien posteriormente se alejó del centro estudiantil.

Para el año siguiente, en el 2004, Mark Zuckerberg tiene su primer contacto con los hermanos gemelos Winklevoss y Divya Narendra, quienes eran estudiantes de la Universidad de Harvard. Cuando ellos se enteraron que Mark había extraído la información de los alumnos para ponerlo en público por medio de facemash.com, hablaron con él para concretar la idea de crear un directorio en línea.

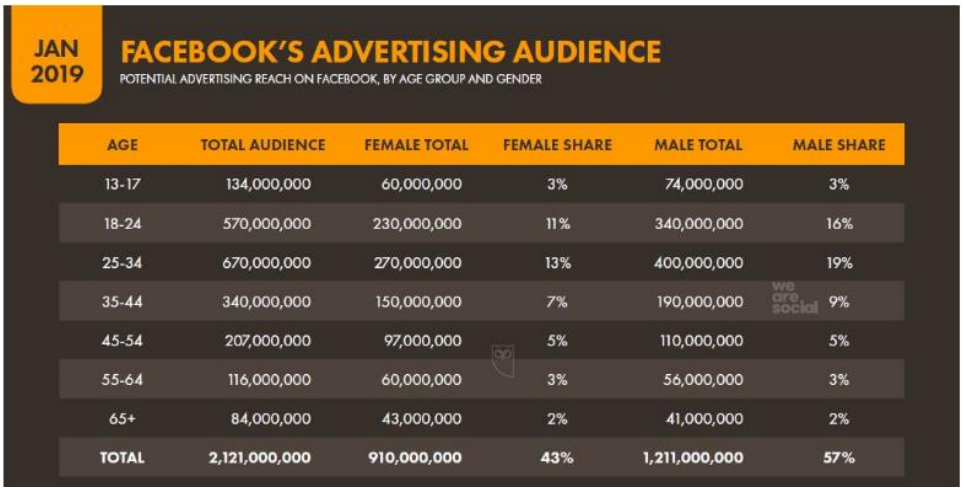
En el 2006, Facebook ya no era un sitio web solo para estudiantes de universidades de Estados Unidos, en donde podían compartir información, o se escogía el estudiante más popular, ahora estaba disponible para todas las personas que podían compartir sus gustos, preferencias con sus amigos, familiares o personas que tuvieran los mismos intereses.

Para el 2009, Facebook ya se convierte en una de las redes más usadas en el mundo con más de 250 millones de usuarios en el mundo. Este sitio web que se inició siendo un espacio virtual para estudiantes de la Universidad de Harvard, se convirtió en una plataforma en línea para todas las personas alrededor del mundo, y que cuenta con más de 124 idiomas.

Cuando se usa Facebook de forma adecuada en la estrategia de Social Media Marketing, se logra generar un importante tráfico de audiencia de valor para la marca hacia el sitio web, que es donde vendemos.

Crecimientos como este, nos obliga a los responsables de las redes sociales en las empresas a incluirla en las estrategias de Social Media Marketing de todas las empresas independiente de su tamaño, categoría y audiencia.

De acuerdo con el estudio de WeAreSocial y Hootsuite, el 43% del alcance potencial de los anuncios de Facebook son mujeres y el 57% son hombres. El 35% de la audiencia es menor de 25 años y más de un 90% acceden a través de dispositivos móviles. Lo anterior hace importante que todos los enlaces desde esta red social al sitio web o tienda electrónica deben tener buena experiencia de navegación en este tipo de dispositivos.



JAN 2019

FACEBOOK'S ADVERTISING AUDIENCE

POTENTIAL ADVERTISING REACH ON FACEBOOK, BY AGE GROUP AND GENDER

AGE	TOTAL AUDIENCE	FEMALE TOTAL	FEMALE SHARE	MALE TOTAL	MALE SHARE
13-17	134,000,000	60,000,000	3%	74,000,000	3%
18-24	570,000,000	230,000,000	11%	340,000,000	16%
25-34	670,000,000	270,000,000	13%	400,000,000	19%
35-44	340,000,000	150,000,000	7%	190,000,000	9%
45-54	207,000,000	97,000,000	5%	110,000,000	5%
55-64	116,000,000	60,000,000	3%	56,000,000	3%
65+	84,000,000	43,000,000	2%	41,000,000	2%
TOTAL	**2,121,000,000**	**910,000,000**	**43%**	**1,211,000,000**	**57%**

Figura 17. Estadísticas Facebook por Genero Fuente http://cort.as/-L28F

JAN 2019

FACEBOOK REACH RANKINGS

BASED ON FACEBOOK'S ADDRESSABLE ADVERTISING AUDIENCE

#	COUNTRY / TERRITORY	REACH	▲QOQ	▲QOQ
01	INDIA	300,000,000	+3.4%	+10,000,000
02	U.S.A.	210,000,000	0%	[UNCHANGED]
03=	BRAZIL	130,000,000	0%	[UNCHANGED]
03=	INDONESIA	130,000,000	0%	[UNCHANGED]
05	MEXICO	86,000,000	+1.2%	+1,000,000
06	PHILIPPINES	75,000,000	+2.7%	+2,000,000
07	VIETNAM	61,000,000	0%	[UNCHANGED]
08	THAILAND	50,000,000	-2.0%	-1,000,000
09	TURKEY	43,000,000	-2.3%	-1,000,000
10	U.K.	40,000,000	0%	[UNCHANGED]

#	COUNTRY / TERRITORY	REACH	▲QOQ	▲QOQ
11	EGYPT	39,000,000	0%	[UNCHANGED]
12	PAKISTAN	36,000,000	0%	[UNCHANGED]
13	FRANCE	35,000,000	+2.9%	+1,000,000
14	BANGLADESH	33,000,000	+3.1%	+1,000,000
15=	ARGENTINA	32,000,000	0%	[UNCHANGED]
15=	COLOMBIA	32,000,000	0%	[UNCHANGED]
15=	GERMANY	32,000,000	0%	[UNCHANGED]
18	ITALY	31,000,000	0%	[UNCHANGED]
19	JAPAN	25,000,000	0%	[UNCHANGED]
20	MALAYSIA	24,000,000	+4.3%	+1,000,000

Figura 18. Estadísticas Facebook por País Fuente http://cort.as/-L28F

3.5. Linkedin

Hoffman fundó Linkedin a finales de 2002 en Mountain View, California. En solo seis años, la compañía se convirtió en una marca reconocible en toda la América corporativa contando ya en ese momento con 350 empleados y valorándose en 2008 en mil millones de dólares.

Desde el principio, Hoffman, un graduado de Stanford comprendió la importancia de la construcción y el aprovechamiento de su Red.

A principios de 1990, después de una temporada estudiando filosofía en Oxford, su vida cambia de rumbo y vuelve a Silicon Valley. Allí comienza casi inmediatamente a aprovechar sus conexiones persiguiendo su sueño de crear una gran empresa de software.

Más allá de Linkedin, el empresario es también un inversor activo, asesorando y financiando más de 60 Esta red social orientada a grupos profesionales, también ha tenido un importante crecimiento en el último tiempo y ha evolucionado en los últimos años pasando de ser un canal de social media de reclutamiento a una donde se comparte información de valor agregado de las diferentes profesiones.

LinkedIn es otro canal de social media necesario para todas las empresas que quieran utilizar las redes sociales como un canal de comunicación y marketing.

El rango de edad más importante para esta red social es de 25 a 34 años y una presencia fuerte en el rango 35 a 54 años. nuevas empresas de Silicon Valley.

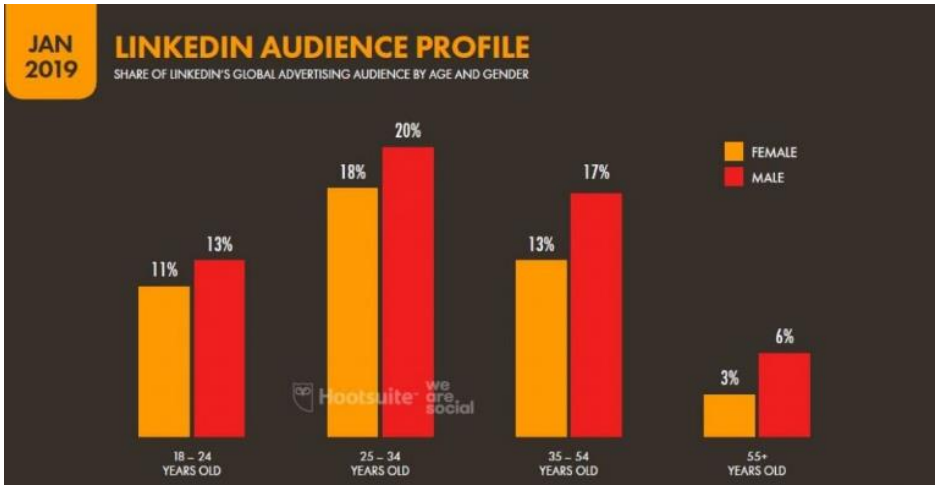


Figura 19. Estadísticas Linkedin por Edades Fuente http://cort.as/-L28F

3.6. Twitter

Twitter se ha ganado popularidad por crear sus tweets con sólo 140 caracteres. Se considera como un servicio de microblogging y como una red social que ha conquistado a sus 328 millones de usuarios activos, y quienes pueden crear y leer tweets de los followers. Algunos usuarios pueden acceder a Twitter a través de SMS, página web o desde la aplicación para dispositivos móviles. Ahora, esta red social ha incursionado en el mundo de la publicidad para fomentar

ayudar a las empresas. Por ello, te contaremos cómo nació y cuál ha sido su trayectoria que lo lleva al éxito.

La red social Twitter fue creado en marzo del 2006, pero fue lanzado en julio de ese mismo año. Este microblogging nació gracias a Jack Dorsey, Noah Glass, Biz Stone y Evan Williams, estos dos últimos habían sido colaboradores de Google. La idea se originó dentro de la compañía Odeo situado en San Francisco, donde se estaba llevando a cabo un servicio de radio online (postcasting), que no tuvo éxito debido al lanzamiento de un producto similar de iTunes.

Al inicio, este microblogging fue usado por los empleados de la compañía Odeo. Los creadores de esta red social fueron Evan Williams y Biz Stone, que tuvieron la colaboración de Jack Dorsey, Evan Henshaw.Path y Noah Glass. Este último fue despedido de la compañía, pero asegura que Twitter nació en su propia computadora. En cambio, Henshaw Path vendió su parte por 7000 dólares, y se compró un Wolkswagen para recorrer todo el país.

Para inicios del 2015, Twitter se renueva con su nueva aplicación Periscope, que ayuda a transmitir eventos en tiempo real, algo similar a Facebook Live. El microblogging ya tenía 200 millones de emisiones en directo luego de un año del lanzamiento de dicho servicio.

Poco a poco, Twitter se fue renovando y acercándose más a usuarios y empresas, y actualmente se ha convertido en una red social favorita para miles de personajes públicos, políticos, deportistas, periodistas, profesionales, y usuarios que la usan para conocer las tendencias en el mundo, entre otras funciones.

Esta red social orientada a grupos profesionales, también ha tenido un importante crecimiento en el último tiempo y ha evolucionado en los últimos años pasando de ser un canal de social media de reclutamiento a una donde se comparte información de valor agregado de las diferentes profesiones.

LinkedIn es otro canal de social media necesario para todas las empresas que quieran utilizar las redes sociales como un canal de comunicación y marketing.

El rango de edad más importante para esta red social es de 25 a 34 años y una presencia fuerte en el rango 35 a 54 años.

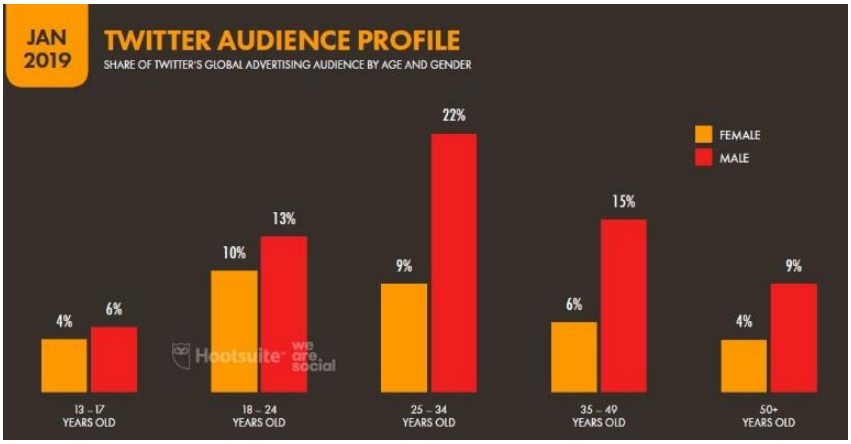


Figura 20. Estadísticas Twitter por Edades Fuente http://cort.as/-L28F

Figura 21Estadísticas Twitter por País Fuente http://cort.as/-L28F

3.7. Análisis de Sentimientos en Twitter

A partir del 2009 se cuenta con referencia de trabajos importantes en los que se aplica Análisis de sentimientos (AS) a Twitter, con la característica que es para idioma Inglés, como lo señala Eugenio Martínez Cámara en su artículo "Análisis de Sentimientos" citando por ejemplo los siguientes trabajos: "Twitter Sentiment Classication using Distant Supervision" en 2009, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series" en 2010

y “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment” en 2010.

- Charlas sin sentido (40%)
- Conversaciones (38%)
- Retweets (9%)
- Autopromoción (5%)
- Mensajes SPAM (4%)
- Noticias (4%)

En Colombia, la edad promedio de los usuarios de Twitter (twiteros) es de 24 años; en el mundo el 80% tienen menos de 29 años, el 43% tienen menos de 19 años y tan solo el 7% tiene más de 40 años. Cifras que concuerdan con los estudios que muestran la participación de los jóvenes dentro de las barras bravas, además son los jóvenes menores de edad los que generalmente inician los problemas escudándose en su edad para no tener problemas judiciales en Colombia (Colombia.com en 2002).

4. HERRAMIENTAS DE ANÁLISIS DE DATOS

4.1. Lenguaje R

R, entre otras cosas, es también un lenguaje de programación, aunque, generalmente, no se utiliza para programar; más bien, se utiliza interactivamente, el usuario lee datos, los revisa, los manipula y genera gráficos, modelos, ficheros Rmd, etc. Además, típicamente, en este proceso hay ciclos, la revisión de los datos conduce a reescribir su lectura, la modelización a modificar su manipulación, etc. Es inhabitual usar R para desarrollar programas largos con R al estilo de los que se crean con Java, C++ u otros lenguajes.

Así que, en R, normalmente, programar no consiste tanto en crear programas como en empaquetar código útil en bloques, i.e., funciones reutilizables.

Muchas de estas funciones son de usar y tirar, es decir, solo son útiles en un contexto determinado: pueden crearse para ser utilizadas en un único proyecto o en una parte muy concreta o pequeña del mismo. Otras son más generales y los usuarios pueden querer guardarlas para reutilizarlas en otros proyectos. No es infrecuente que los usuarios identifiquen determinadas necesidades, desarrollen funciones destinadas a satisfacerlas y acaben creando sus propios paquetes de funciones de R y redistribuyéndolas entre sus colegas. No obstante, la creación de paquetes, aunque no es complicada, queda fuera del alcance de este libro.

Existen dos grandes paradigmas de programación:

- Programación imperativa: variables, bucles, etc. Es la habitual en lenguajes como C, Fortran o Matlab.

- Programación funcional, donde las funciones son ciudadanos de primera clase. Lisp fue el lenguaje pionero en programación funcional y, actualmente, Haskell o Scala son lenguajes casi puramente funcionales; otros como Python, Java o C++, aunque imperativos, incorporan cada vez más elementos funcionales.

R permite combinar ambos. Y los combina, además, con la programación orientada a objetos. El objetivo de la sección será el de familiarizarnos con los aspectos imperativos y funcionales de la programación en R. Esta sección, de todos modos, no es una introducción a la programación. Se limita a mostrar la sintaxis que utiliza R para las construcciones (expresiones condicionales, bucles, definición de funciones, maps, etc.) habituales en otros lenguajes de programación, con alguno de los cuales se espera que esté familiarizado el lector.

Fue creado en la Universidad de Auckland de Nueva

Zelanda en 1.993 por Ross Ihaka y Robert Gentlemean del departamento de estadística, es software libre, un proyecto GPL/GNU. R es el resultado de la combinación de las fortalezas de sus predecesores lenguajes de programación para estadística S (de donde toma su apariencia) y Sheme (de donde obtiene la semántica). Se encuentra disponible para plataformas UNIX/LINUX, MAC y Windows.

R permite a sus usuarios crear sus propias funciones, permite integrarse a bases de datos. Por medio de bibliotecas puede ser utilizado desde Perl y Python, su poder se compara con GNU

Octave y MATLAB. Cuenta con una interfaz para interactuar con Weka (RWeka) y enriquecer el lenguaje R con algoritmos de minería de datos.

Figura 22. Logo Lenguaje R Fuente http://cort.as/-RNTX

En cuanto a las conexiones de R con Twitter y Facebook, se cuenta con APIs para obtener servicios de estas redes sociales de forma sencilla, son ellas TwitteR y Rfacebook respectivamente. Además, cuenta con más de 10 interfaces gráficas e IDEs como Sage, JGR, RStudio, Eclipse, Emacs, Vim, Scite, notepad++, entre otros. Su última versión es la 3.4.3, disponible desde el 30 de abril de 2017 en el sitio oficial https://www.r-project.org/. El logo oficial de R en 2016 aparece en la figura 22 con licencia bajo los términos de Creative Commons Attribution-ShareAlike 4.0 International license (CC-BY-SA 4.0).

4.2. Apis

Twitter cuenta con 3 API's (Application Programming Interface) en el lenguaje de programación R con las cuales se pueden obtener flujos de datos. Debido a la cantidad de mensajes que los usuarios de la red social generan por segundo, la que a su vez es almacenada como grandes datos en sus centros de almacenamiento, se requiere que al momento de obtener

información se tengan consideraciones precisas para extraer subconjuntos de datos relevantes para la investigación. Se deben considerar: la cadena de búsqueda, lenguaje, rangos de fechas, ubicación geográfica aproximada y tipo de conexión para realizar la consulta. A continuación, el listado de los tipos de conexión con que cuenta la API para la interacción con Twitter.

- ***Streaming API***

Proporciona un flujo continuo de tweets para conformar un subconjunto de datos en formato json casi en tiempo real de las publicaciones de los usuarios de Twitter, por medio de una conexión http permanente en el tiempo establecido. La velocidad de recepción de los tweets es dependiente del ancho de banda de los extremos en las conexiones y de la sobre carga en los servidores de Twitter.

- ***SearchTwitter API***

Suministra los tweets publicados en los últimos siete días de acuerdo a la consulta realizada, hasta donde puede recuperar los últimos 8000 mensajes aproximadamente, entrega el subconjunto de datos en formato json o atom. Permite aplicar filtros por lenguaje y localización. Está limitada hasta 150 peticiones por hora, por usuario o por IP.

- ***REST API***

Permite consultar en toda la base de datos disponible en Twitter, de la que puede recuperar los últimos 3200 tweets. Soporta formatos json, xml, rss, atom. Al igual que Search API, está limitada hasta 150 peticiones por hora, por usuario o por IP.

4.3.Tableau

Tableau es una herramienta de visualización de datos potente utilizada en el área de la Inteligencia de negocios (más conocida como Business Intelligence). Simplifica los datos en bruto en un formato muy fácil de entender.

La esencia de Tableau es simple y a la vez muy relevante: ayudar a las personas y empresas a ver y comprender todos sus datos. Y esto lo consigue ofreciendo a los usuarios toda una selección de herramientas útiles e intuitivas de inteligencia de negocios.

A través de funciones simples como la de arrastrar y soltar, cualquier persona puede acceder y analizar de forma sencilla datos, e incluso, crear informes y compartir esta información con otros usuarios.

Figura 23. Logo Tableau. Fuente https://n9.cl/jk2o

En Tableau, una historia es una secuencia de visualizaciones que se usan en conjunto para transmitir información. Puede crear historias para contar una narración de datos, proporcionar contexto, mostrar la relación de las decisiones con los resultados obtenidos o simplemente para agregar más atractivo al caso.

Una historia es una hoja, de modo que se rige por los mismos métodos que se usan para crear hojas de trabajo y dashboards, ponerles nombres y administrarlos (consulte Libros de trabajo y hojas). Al mismo tiempo, una historia también es un conjunto de hojas dispuestas en un orden específico. Las hojas que componen una historia se denominan puntos de la historia.

Al compartir una historia (por ejemplo, al publicar un libro de trabajo en Tableau Public, Tableau Server o Tableau Online), los usuarios pueden interactuar con la historia para mostrar nuevas conclusiones o plantear nuevas preguntas acerca de los datos.

5. PROCESO ETL

Para el desarrollo del problema de estudio, se plantea el siguiente modelo de proceso (ver la figura 24), el cual se encuentra en etapas, cada una de las etapas requiere que culmine la anterior. Este modelo de proceso es una particularización muy aproximada al modelo KDD, citado anteriormente en el presente documento y por muchos autores.

Figura 24. Modelo de Bloques. Fuente Autor.

5.1. Recolección de datos

Esta etapa es la inicial del proceso, el cual comprende en reunir información de las redes sociales, en este caso Twitter, red social que se centra en recibir comentarios y opiniones de diversas temáticas, la conexión en esta etapa se realiza con la API TwitteR, por medio de

consultas y servicios streaming utilizando el lenguaje de programación R en etapas siguientes al modelo.

5.2. Almacenamiento

En ésta etapa del modelo se realizan tres operaciones (extracción, transformación y carga), como se había mencionado con anterioridad, cada etapa es insumo para la siguiente. El resultado de ésta etapa es el conjunto de datos (dataset) obtenido de la red social.

- ***Extracción***

En esta etapa que permite crear conjuntos de datos específicos, para no traer cantidades de datos como los que administra la red social objetivo. Se realiza por medio de conexiones utilizando la API twitteR, por periodos de tiempo (o por defecto los últimos diez días), idioma de los mensajes (en éste caso español); la ubicación geográfica (desactivada en las consultas de

éste estudio debido al escaso número de mensajes con geo referenciación), acordes a la fecha y ubicación de las problemáticas del transporte masivo. Los atributos que permanezcan estáticos o no cambien no son tenidos en cuenta. En ésta etapa, las consultas se realizan mediante criterios específicos relacionados con el Dominio de la Investigación, los cuales son los nombres con los que se conoce al transporte masivo de TransMilenio ubicado en la capital de Colombia. La extracción puede ser a las líneas de tiempo de usuarios específicos, exponiendo una consulta al flujo de datos actual de la red social y consultando la base de datos en un rango de tiempo determinado.

- ***Transformación***

En esta sub etapa del modelo, no tenemos interacción directa con las respectivas redes sociales, el proceso de extracción suministra los datos, los cuales son transformados de lista de tweets a tuplas para ser almacenados en una matriz más general de datos llamada técnicamente como Data Frame, cada tupla se compone de 16 campos, todos suministrados por la red social.

- ***Carga***

Esta sub etapa permite crear el conjunto de datos de estudio o Data Set objetivo en un archivo externo tipo CSV (comma - separated values por su sigla en inglés, es decir valores separados por comas), los cuales permiten mejorar el tiempo de carga de la información en comparación con archivos similares almacenados en hojas de cálculo como Microsoft Office Excel.

5.3. Técnicas de filtrado

En ésta etapa del modelo se filtran los mensajes de texto cargados, denominados como Data Set, para que la computadora los pueda entender y analizar en etapas siguientes. Se eliminan de los mensajes los enlaces, retweets, menciones, palabras vacías, ya que ésta información no aporta mayores detalles al proceso.

5.4. Análisis

en esta etapa usando técnicas estadísticas en el lenguaje de programación R, se seleccionan aquellos mensajes que reúnen características de comportamientos inapropiados, promoción o aliento a contravenciones como riñas callejeras, hurtos, entre otros, como también reuniones en lugares públicos o convocatorias masivas de personas que pueden afectar la convivencia y seguridad ciudadana.

5.5. Conocimiento

Esta etapa concluye con la propuesta del modelo. Es la última etapa y es la encargada de los mensajes previamente cargados y que deben de generar una advertencia dentro del seguimiento inconvenientes y problemáticas más destacadas del trasporte masivo de la capital.

6. EXTRACION DE DATOS

6.1. Extracción de Datos con Lenguaje R

Como se ha mencionado a través de este documento, los datos que serán recolectados se encuentran en la red social de Twitter, en esta red social se encuentran temáticas y problemáticas de bastantes temas de intereses social y común.

6.2. Instalación de Librerías

El lenguaje de programación R se instala en arquitectura Windows, R se utiliza para la estadística computacional y la graficación. Puede ser descargado de la página oficial del proyecto: https://www.r-project.org. Luego de la descarga e instalación del lenguaje de programación R se debe instalar RStudio en su versión para escritorio; el cual es el Entorno de Desarrollo Integrado (IDE) para el lenguaje de programación R, se encuentra disponible para la descarga en:www.rstudio.com/products/RStudio/

Una vez instalado el lenguaje de programación y el entorno de desarrollo para él se debe proceder con la instalación de las librerías, es un procedimiento que solo se realiza una vez dentro de la preparación del sistema. Para instalar una librería se debe lanzar desde la línea de comandos dentro de paréntesis en comillas dobles anteponiendo el comando install.packages. Es indispensable tener conectividad a internet para obtener el paquete. Para el desarrollo del proyecto se han utilizado los siguientes paquetes, a continuación, con su comando de instalación:

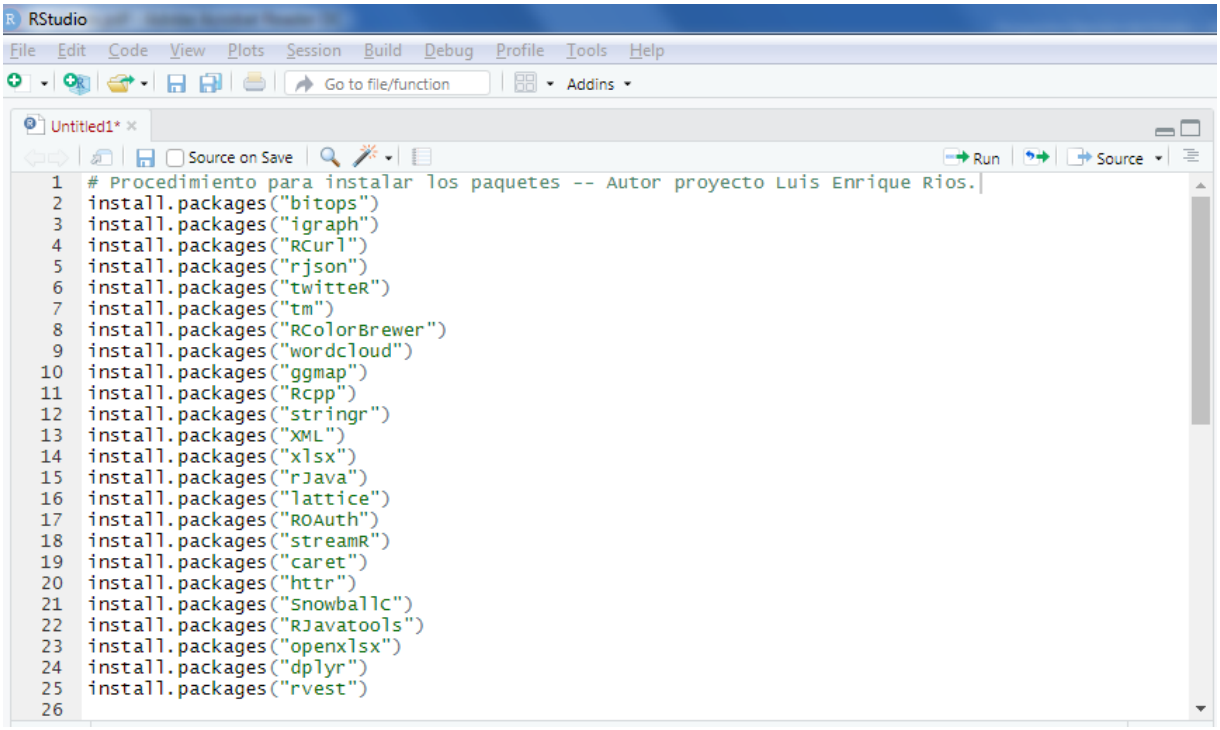


Figura 25. Instalación de Paquetes. Fuente Autor.

6.3. Carga de Librerías

Este procedimiento solo se realiza una sola vez y se debe realizar conectado al servicio de internet para lograr su respectiva descarga, y cada vez que se inicie el sistema R y RStudio es necesario llamar las librerías correspondientes a los respectivos paquetes instalados, se realiza desde la línea de comandos con el comando library encerrando en paréntesis el nombre de la librería, como se muestra a continuación:

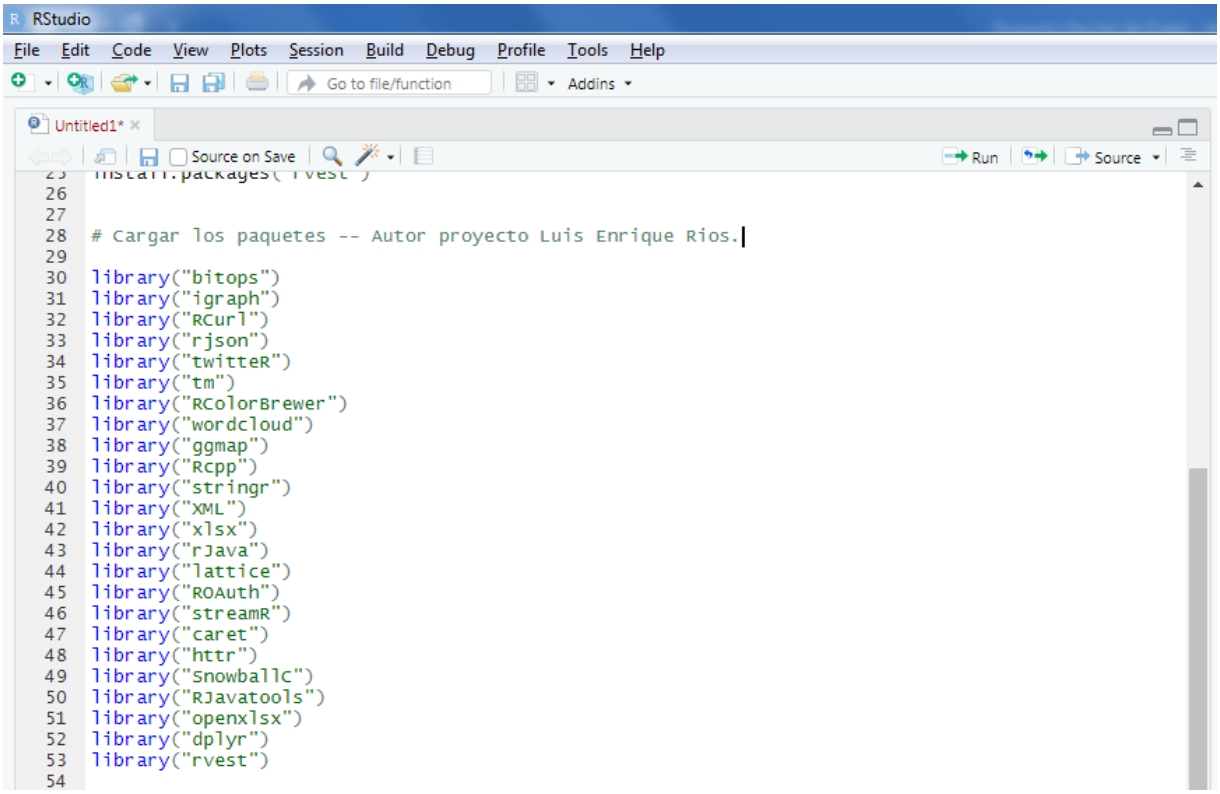


Figura 26. Carga de Paquetes. Fuente Autor.

6.4. Autenticación a Twitter

Para lograr conectarse a Twitter en necesario tener una cuenta, pero además se debe crear una aplicación de tipo desarrollador, para esto se debe ingresar a www.apps.twitter.com. de esta manera obtenemos los valores necesarios para ingresar a las variables que studio R requiere para el siguiente comando setup_twitter_oauth.

Figura 27 Registro con Twitter. Fuente Autor.

En el momento de ejecutar la sentencia anterior esta nos debe mostrar el siguiente mensaje [1] "Using direct authentication", de ser así el sistema nos está indicando que la conexión hacia twitter está correctamente establecida.

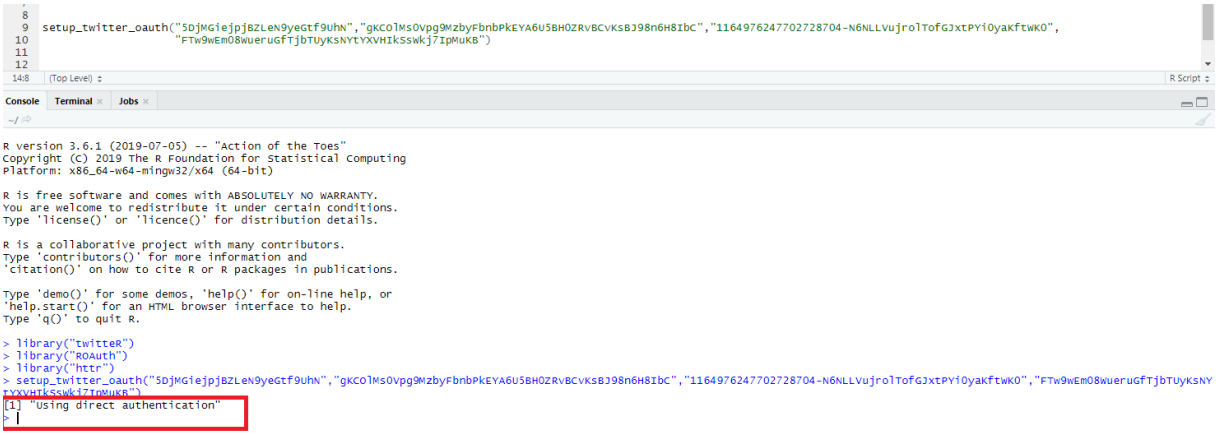


Figura 28. Conexión con Twitter. Fuente Autor.

6.5. Búsqueda de Información

En esta etapa se debe programar R para realizar la extracción de datos, la cual se realiza a través de la siguiente sintaxis

tweetsTrans <- searchTwitter ("#Transmilenio", n = 1000)

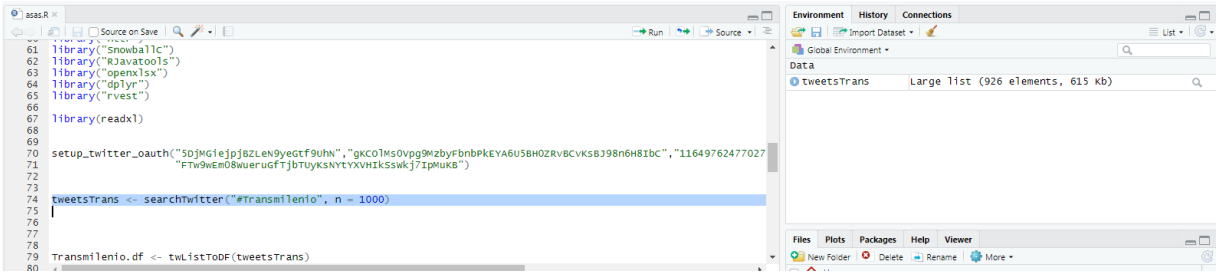


Figura 29 .Sintaxis Consulta de Información. . Fuente Autor.

tweetsTrans es el nombre de la variable donde vamos a almacenar los registros de la consulta, searchTwitter es la función que conecta con Twitter, seguido de #Transmilenio que es la palabra a filtrar y la letra n representa el número de registros que se desea consultar y extraer.

Después de ejecutar la sentencia el sistema nos crea una data con el número de registros que se extrajo en el paso anterior.

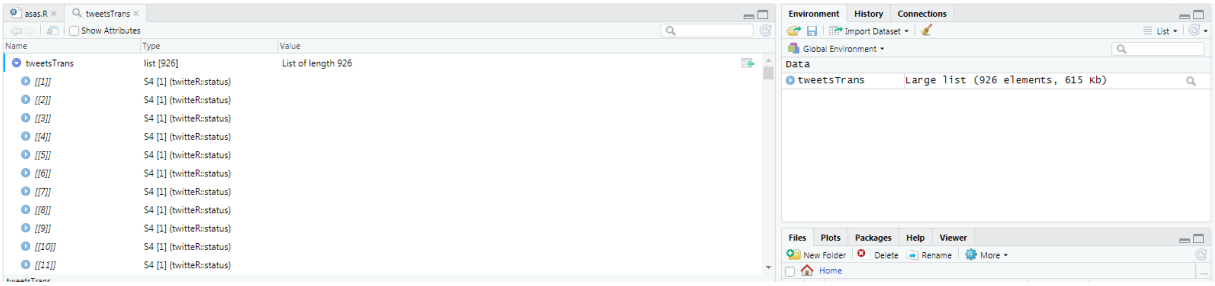


Figura 30. Creación Data. Fuente Autor.

6.6. Creación DataFrame

Para entender de forma correcta la información anterior se debe crear una tabla para organizar los registros, con la siguiente sintaxis se crea la tabla

Transmilenio.df <- twListToDF(tweetsTrans)

Donde Transmilenio.df es el nombre de la base de datos, twListToDF es la función y tweetsTrans la variable anteriormente creada.

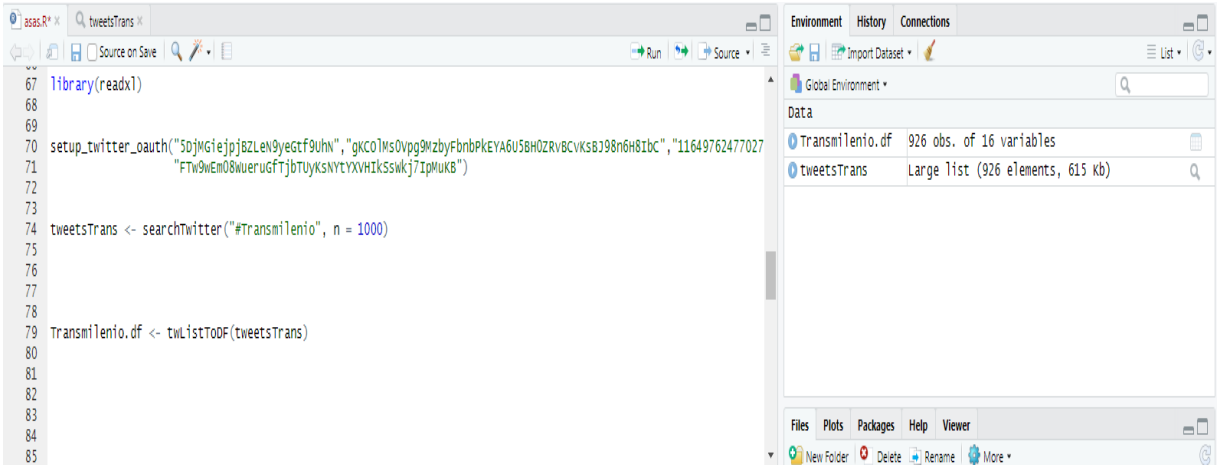


Figura 31. Creación de DataFrame. Fuente Autor.

6.7. Exportación de información

Para exportar la base de datos en formato Xlsm o CSV aplicamos la siguiente sintaxis

write.csv (Transmilenio.df, "C:/BDTransmi.CSV")

donde write.csv es la función de R, Transmilenio.df, el nombre de la base de datos, C:/BDTransmi es la ruta y el nombre con el cual se va a almacenar y CSV el formato del archivo.

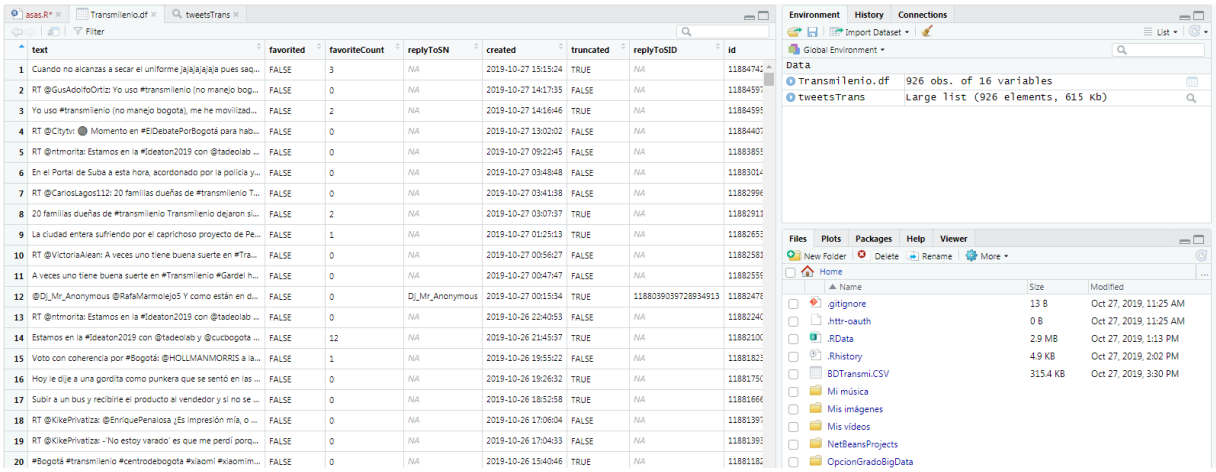


Figura 32. Información Organizada en la Tabla. Fuente Autor.

Se puede evidenciar que al ejecutar el paso anterior se genera y almacena la base de datos desde R

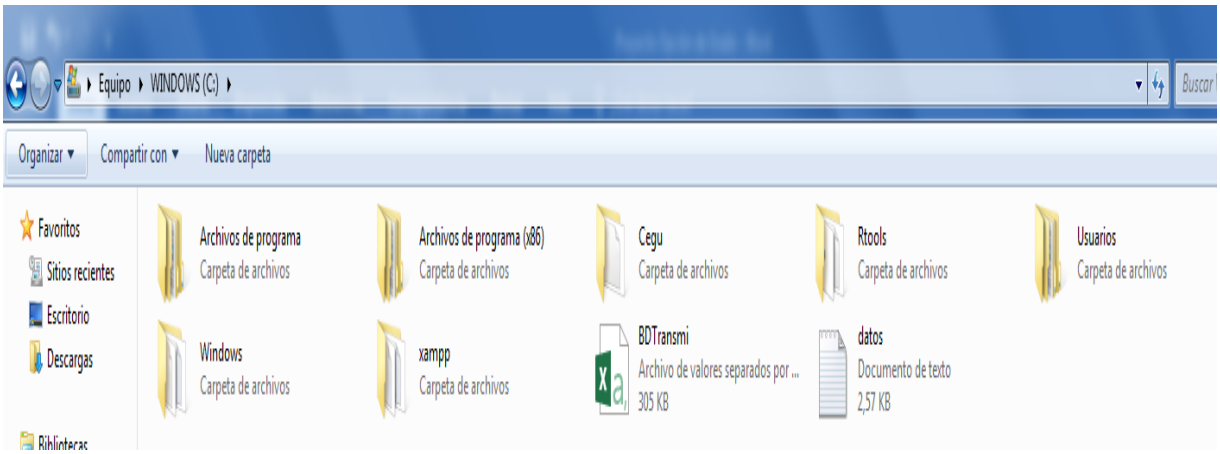


Figura 33. Exportación de Base de Datos. Fuente Autor.

6.8. Aplicación de técnicas de filtrado

Por medio de las siguientes sentencias se realiza el proceso de modelamiento de datos en la base de datos que se exporto en Excel

#convertimos la descripción de esos seguidores en un String.

texto2 <- toString(seguidores$description)

#El texto lo transformamos en una lista separada por espacios

texto_split2 = strsplit(texto2, split=" ")

- ***#Deshacemos esa lista y tenemos el data.frame***

texto_col2 = as.character(unlist(texto_split2))

texto_col2 = data.frame(toupper(texto_col2))

names(texto_col2) = c("V1")

- ***#Eliminamos algunos caracteres regulares***

texto_col2$V1 = gsub("([[:space:]])","",texto_col2$V1)

texto_col2$V1 = gsub("([[:digit:]])","",texto_col2$V1)

texto_col2$V1 = gsub("([[:punct:]])","",texto_col2$V1)

- ***#Creamos una variable longitud de la palabra***

texto_col2$largo = nchar(texto_col2$V1)

- ***#Quitamos palabras cortas***

texto_col2 = subset(texto_col2,largo>4 & largo<=10)

#crear dataframe

palabras_seguidores = data.frame(table(texto_col2$V1))

- ***#Ordenamos***

palabras_seguidores_f<-filter(palabras_seguidores, palabras_seguidores[["Freq"]]>3)

arrange(palabras_seguidores_f, desc(palabras_seguidores_f[["Freq"]]))

- ***#Hacemos la nube de palabras***

wordcloud(words = palabras_seguidores_f$Var1, freq = palabras_seguidores_f$Freq, min.freq = 1, max.words=200, random.order=FALSE, rot.per=0.35,colors=brewer.pal(8, "Dark2"))

asas.R* ×

Source on Save

```
107 nuevo_texto <- tolower(BDTransmi.CSV)
108
109 texto2 <- toString(busqueda$text)
110
111 texto_split2 = strsplit(texto2, split=" ")
112
113 texto_col2 = as.character(unlist(texto_split2))
114 texto_col2 = data.frame(toupper(texto_col2))
115 names(texto_col2) = c("V1")
116
117 texto_col2$V1 = gsub("([[:space:]])","",texto_col2$V1)
118 texto_col2$V1 = gsub("([[:digit:]])","",texto_col2$V1)
119 texto_col2$V1 = gsub("([[:punct:]])","",texto_col2$V1)
120
121
122 texto_col2$largo = nchar(texto_col2$V1)
123 texto_col2 = subset(texto_col2,largo>4 & largo<=10)
124
125 palabras_seguidores = data.frame(table(texto_col2$V1))
126
127 palabras_seguidores_f<-filter(palabras_seguidores, palabras_seguidores[["Freq"]]>3)
128 arrange(palabras_seguidores_f, desc(palabras_seguidores_f[["Freq"]]))
129
130
131 wordcloud(words = palabras_seguidores_f$Var1, freq = palabras_seguidores_f$Freq, min.freq = 1,
132 max.words=200, random.order=FALSE, rot.per=0.35,colors=brewer.pal(8, "Dark2"))
133
```

Figura 34. Técnica de Filtrado. Fuente Autor.

7. ANÁLISIS Y RESULTADOS

7.1. Análisis Lenguaje R

Al finalizar la ejecución del paso anterior, se generaron 116 grupos de palabras, las cuales en cada grupo se hace referencia al número de veces que ha sido utilizadas en los tweets.

Palabras	Numero	Palabras	Numero	Palabras	Numero
COLADOS	364	FRENTE	9	CAMBIAR	5
CONTRA	299	MINUTOS	9	COLOMBIA	5
SEGURIDAD	292	NOSOTROS	9	ESPERANDO	5
GUARDAS	290	RENOVAR	9	MANERA	5
ENTRE	289	CARACAS	8	OBRAEN	5
AUXILIARES	288	LLEVAMOS	8	PUEDE	5
COLCHECK	253	MEJORAR	8	TODOS	5
FALSO	244	PERSONAS	8	TRANCON	5
ASEGURA	242	RECUERDA	8	VIGILANTES	5
GASTADO	241	SANTAFE	8	AESTAHORA	4
PUSIMOS	80	VIGILANTE	8	ALGUIEN	4
CITYTV	79	BUSES	7	AVANZA	4
SOBRE	75	ENTIENDO	7	CIUDAD	4
MOMENTO	71	FORTALECER	7	ESTAFA	4
HABLAR	66	JUSTIFICA	7	ESTAMOS	4
PLANTEA	32	PARECE	7	EXCELENTE	4
PORQUE	28	POSTURA	7	FAVOR	4
AHORA	27	RESCATAR	7	GGERALDINE	4
RUTAS	25	TENER	7	GOBIERNOS	4
ALGUNAS	25	TRANSPORTE	7	HACER	4
DEMORAS	25	URGENCIA	7	MISMO	4
REPORTA	15	VEMOS	7	MOVILIDAD	4
BOGOTA	14	VIOLENCIA	7	OPONERSE	4
CUANDO	14	AHORRAR	6	PERSONAL	4
HOLLMAN	14	AMBIENTE	6	POPULAR	4
NEGROS	14	APORTAR	6	TIENE	4
PIERDA	14	BOTELLAS	6	UFUFUF	4
TIRABA	14	DESDE	6	UNACIONAL	4
NEGOCIO	13	DESPIDEN	6	VECES	4
CALLE	12	EMPAQUES	6		
METRO	11	ESTACIONES	6		
EMPLEADO	11	MIERDA	6		
PARECER	11	MUJER	6		
PETRO	11	PORTAL	6		
SISTEMA	11	PUEDES	6		
SUBALTERNO	10	RAZONES	6		
FINANZAS	10	RECARGASTE	6		
IDIOTA	10	REFIERE	6		
KINHO	10	SERVICIO	6		
MEDIOCRE	10	TARJETA	6		
MENTOR	10	TRONCALES	6		
QUEBRADO	10	TULLAVE	6		
SIQUIERA	10	BUENOS	5		

Figura 35. Palabras Más Utilizadas en Twitter. Fuente Autor.

Infraestructura	
Portales	9
Patios	11
Estaciones regulares	143
Corredores	12
Garajes	34
Paraderos	7171
Zonas operación	13
ciclo-parqueaderos	22
cupos parqueo	5260
conductores operativos	18571
Conductores Troncal	4978
Conductores Zonal	11449
Conductores Alimentación	2144
Puntos Recarga	5017

Figura 37. Datos Infraestructura. Fuente Autor.

Demanda Abordajes	
I-2013	3.271.463
I-2013	3.271.463
II-2013	6.735.003
III-2013	13.778.626
IV-2013	22.908.005
I-2014	36.732.814
II-2014	48.820.948
III-2014	64.472.453
IV-2014	75.862.467
I-2015	83.169.381
II-2015	96.756.365
III-2015	114.194.141
IV-2015	123.143.632
I - 2016	124.283.685
II - 2016	129.538.818
III - 2016	133.867.986
IV - 2016	125.178.121
I - 2017	124.801.691
II - 2017	116.782.095
III - 2017	124.001.500
IV - 2017	124.001.500
I - 2018	110.190.271
II - 2018	112.822.481
III - 2018	115.195.745
IV - 2018	111.960.974
I - 2019	110.225.313
II - 2019	35.877.256

Figura 39. Datos Abordaje. Fuente Autor.

Problematicas	
Vendedores ambulantes	47.800
Exceso de demanda	2.600.000

Figura 41. Datos Problemática. Fuente Autor.

Demanda Pasajeros	
Autopista Norte	99.188
Portal Américas	96.116
Cabecera Calle	82.750
Portal Suba	78.781
Portal Sur	70.553
Portal Eldorado	65.399
Cabecera Usme	54.795
Portal Tunal	51.280
Portal 20 de Julio	37.065
Total Portales	635.927

Figura 36. Datos Pasajeros. Fuente Autor.

Demanda Estaciones	
San Mateo	60.630
Banderas P. Central	48.344
Calle 100	43.385
Marly	41.026
Avenida Jiménez	38.180
Calle 72	38.094
Calle 63	34.242
Alcalá	32.626
Toberín	32.123
Calle 45	32.031
Total Top	400.681

Figura 38. Datos Estaciones. Fuente Autor.

Numero de Rutas	
Buses Troncales Servicios	98
Buses Alimentadores Servicios	107
Complementarios Servicios	283

Figura 40. Datos Rutas. Fuente Autor.

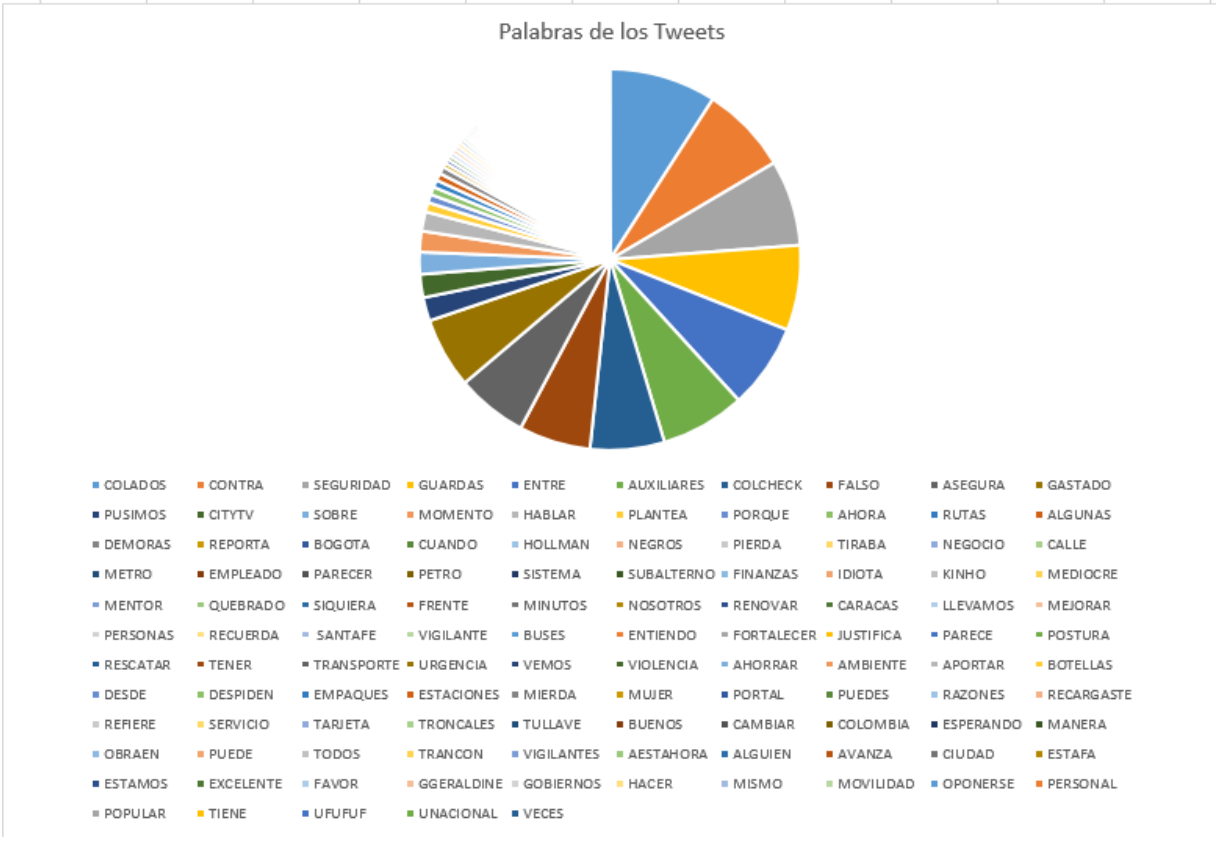


Figura 42. Grafica Palabras. Fuente Autor.

Figura 43. Wordcloud. Fuente Autor

7.2. Análisis Tableau

Se toman los mismos datos extraídos de Twitter mediante lenguaje R y se analizan en Tableau.

En la Figura 44 se evidencia que el número de sentimientos negativos sobrepasa de 20 diferentes emociones.

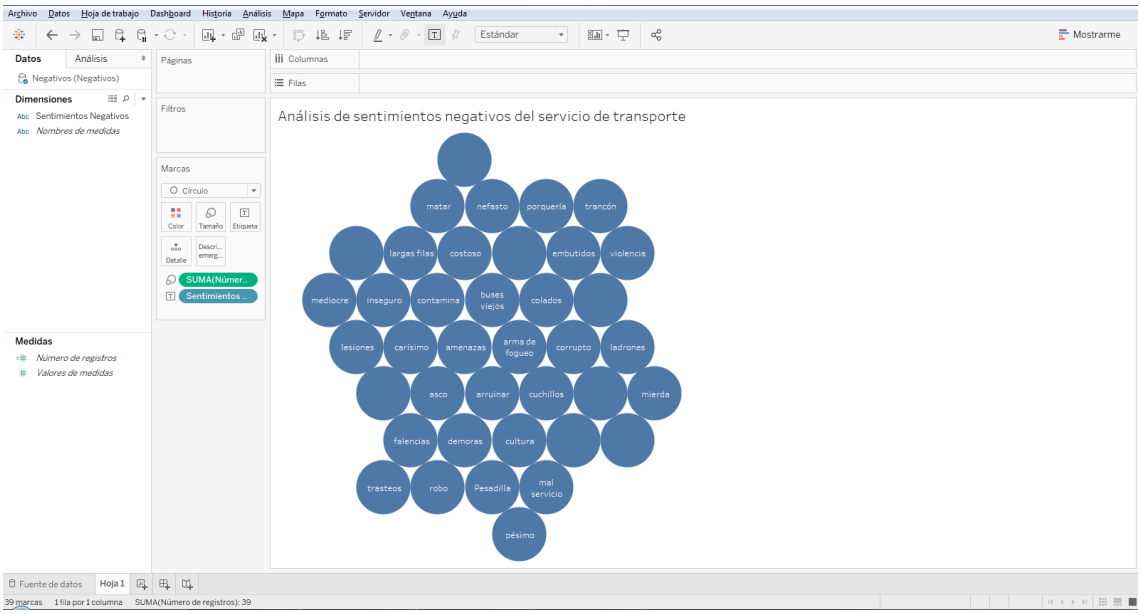


Figura 44. Grafica Sentimientos Negativos. Fuente Autor.

En la figura 45 se evidencia que son muy pocas las emociones a favor del sistema.

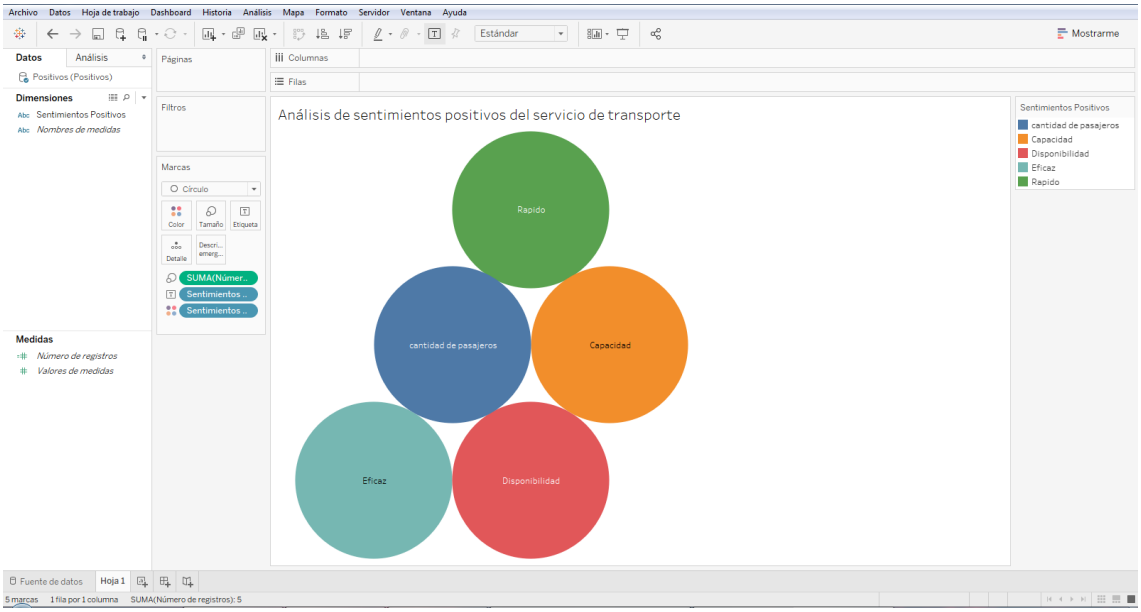


Figura 45. Grafica Sentimientos Positivos. Fuente Autor.

Figura 46 indica que la gran problemática del sistema de transporte es enfoca en la demanda del transporte, esto se debe a que los buses que se encuentran en funcionamiento no dan abasto para la cantidad de habitantes.

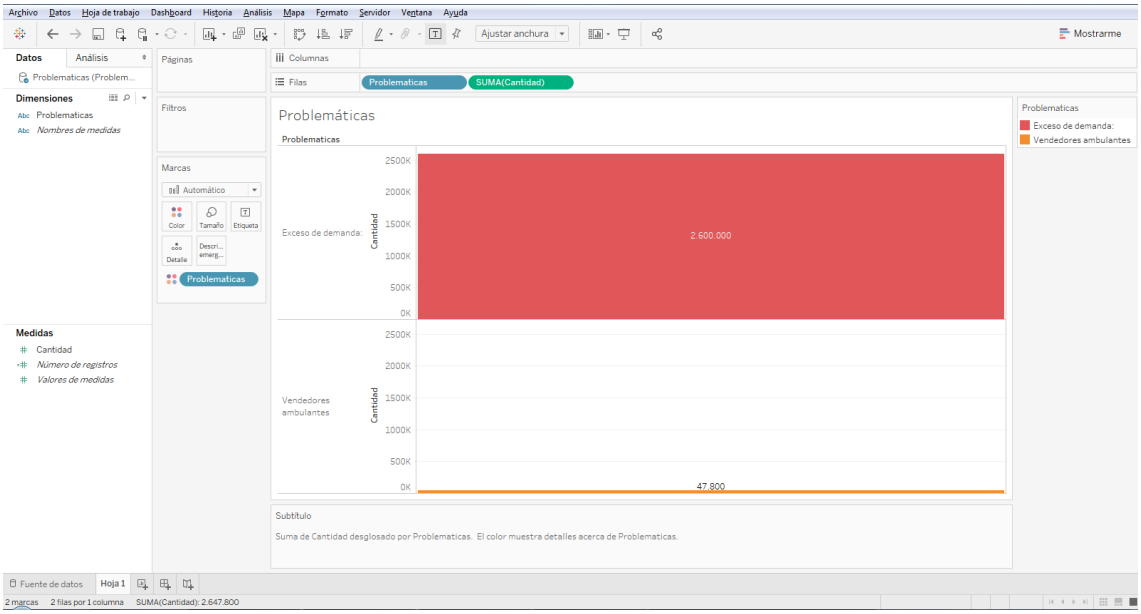


Figura 46. Grafica Problemáticas. Fuente Autor.

Figura 47 muestra que la mayoría de conductores están en la línea zonal, esta es de los servicios SITP.

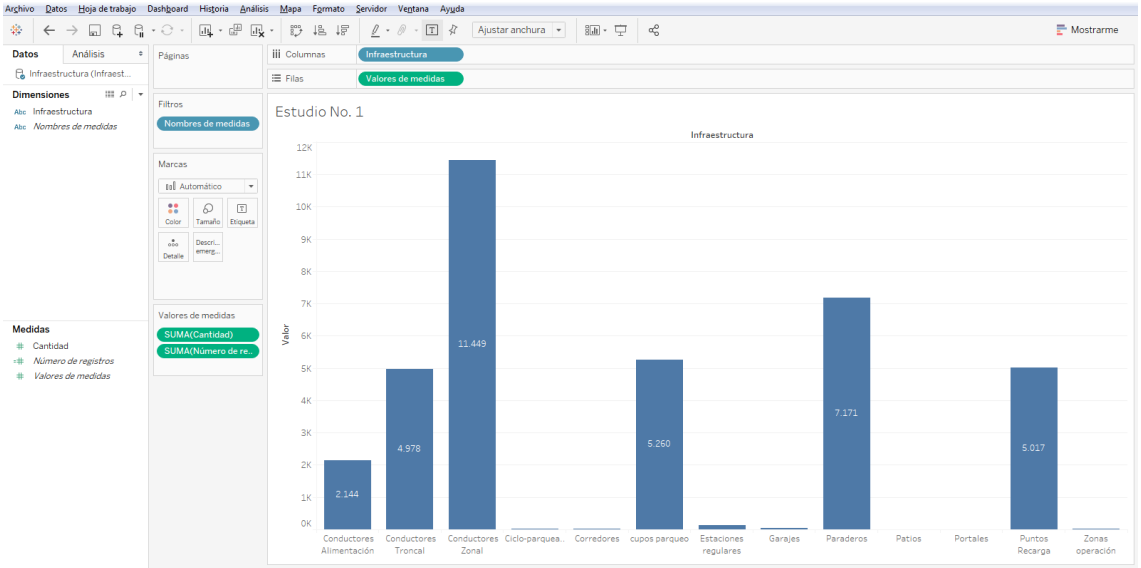


Figura 47. Grafica Infraestructura. Fuente Autor.

Figura 48 muestra que la mayor línea de servicios pertenece a servicios complementarios y la menor es en servicios troncales.

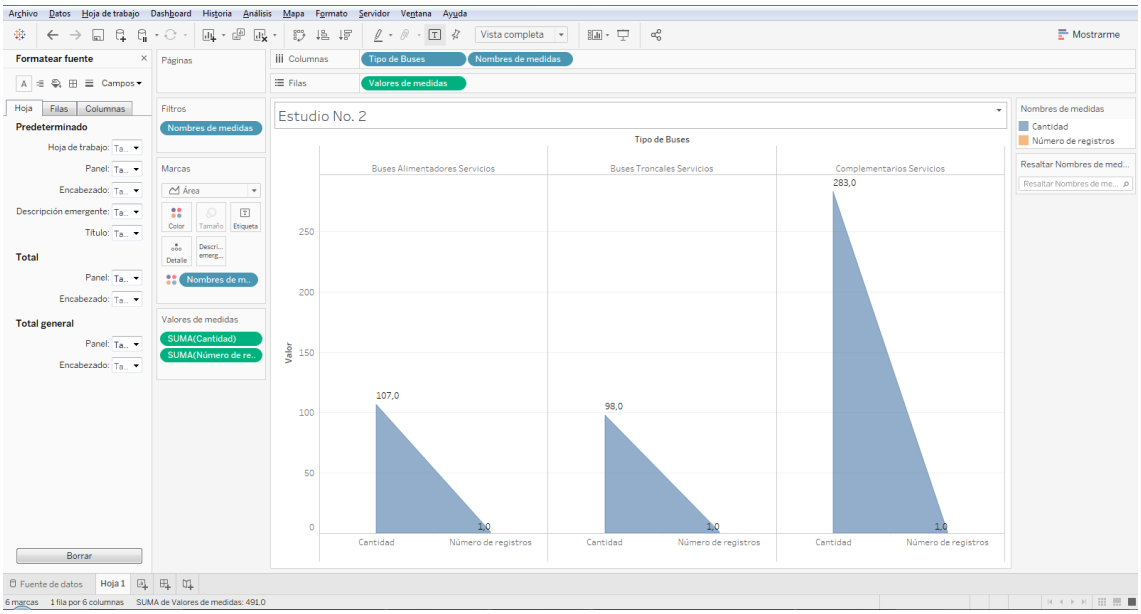


Figura 48. Grafica Numero de Buses. Fuente Autor.

Grafica 49 muestra que el portal con más demanda es el portal del norte y el que menos demanda tiene portal 20 de julio.

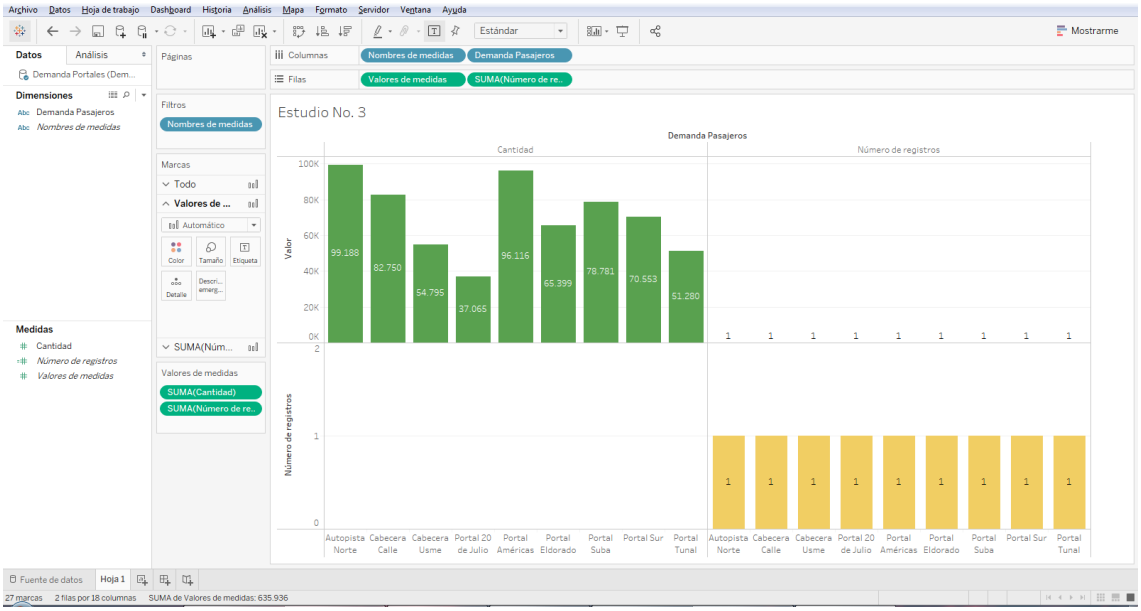


Figura 49. Grafica Pasajeros. Fuente Autor.

Figura 50 indica las 10 estaciones con más demanda de pasajeros, ubicando la estación san mateo como la primera y la estación 45 como la última.

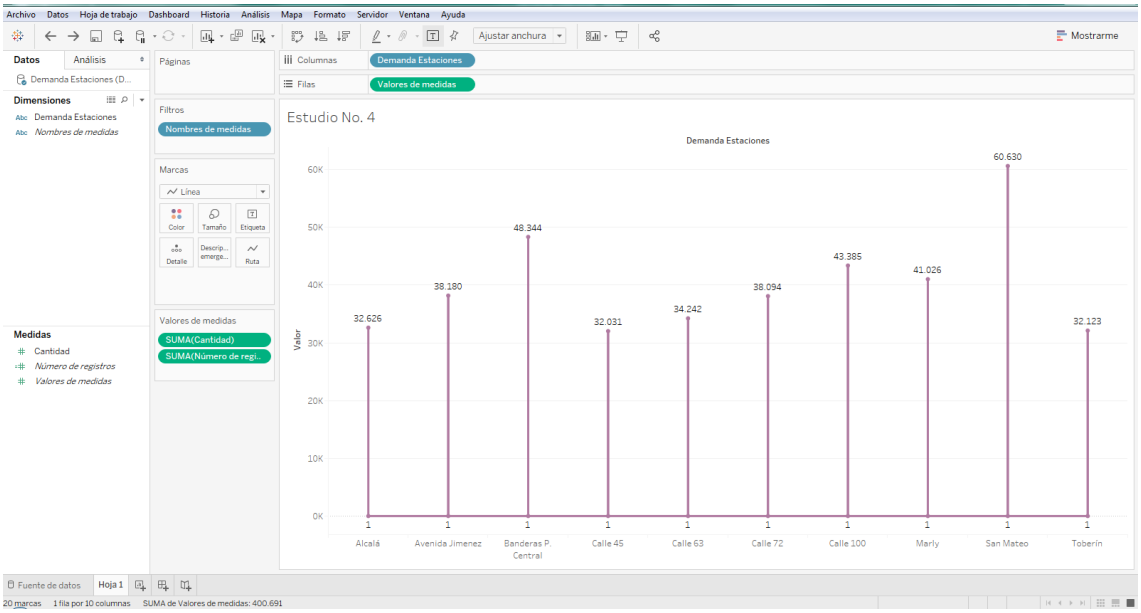


Figura 50. Grafica Estaciones. Fuente Autor.

Figura 51 se enfoca en el histórico de los transbordos que se realizan al interior del sistema identificando que en el año 2016 era muy usual por lo que las rutas no estaban bien estructuradas como lo están en el año presente 2019.

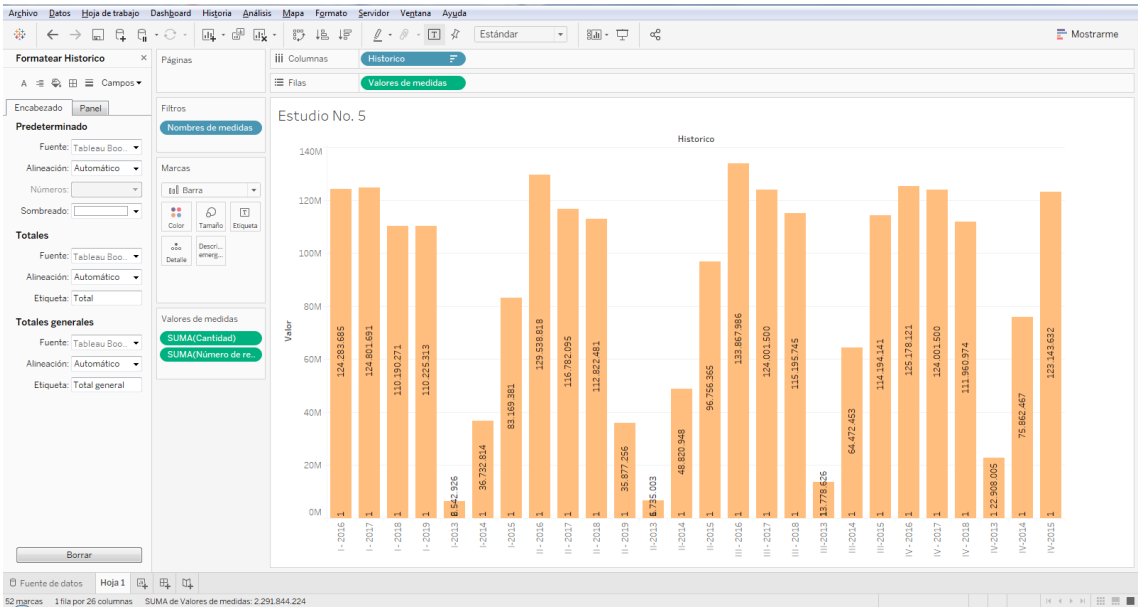


Figura 51. Grafica Histórico Abordajes. Fuente Autor.

Figura 52 muestra el número de palabras más twitteadas, mostrando que el tema de los colados es el más frecuente en el sistema.

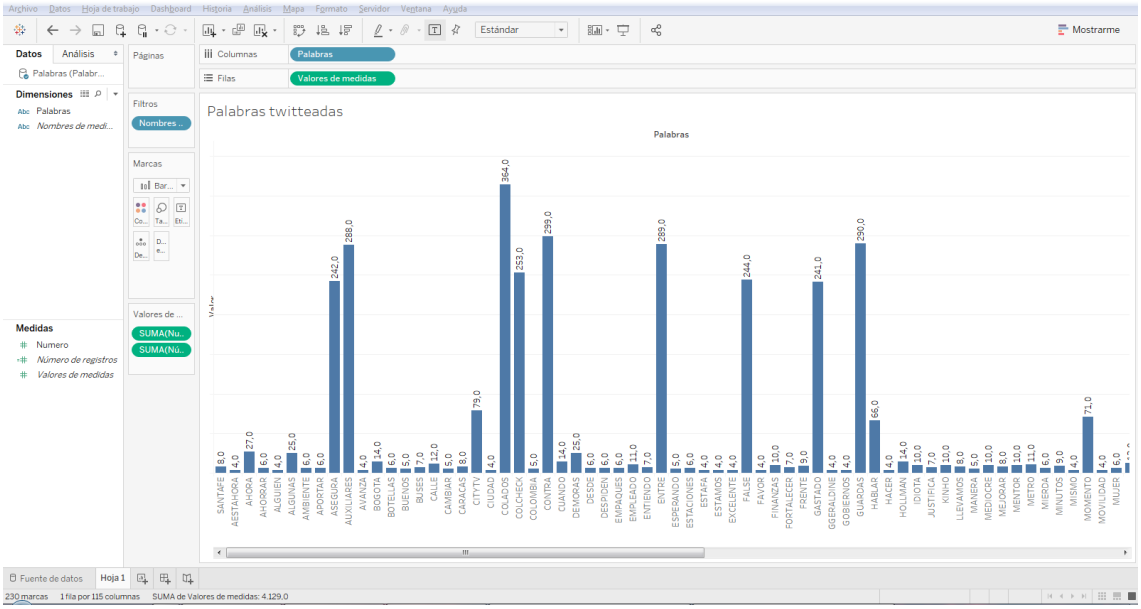


Figura 52. Grafica Palabras Más Comunes-1. Fuente Autor.

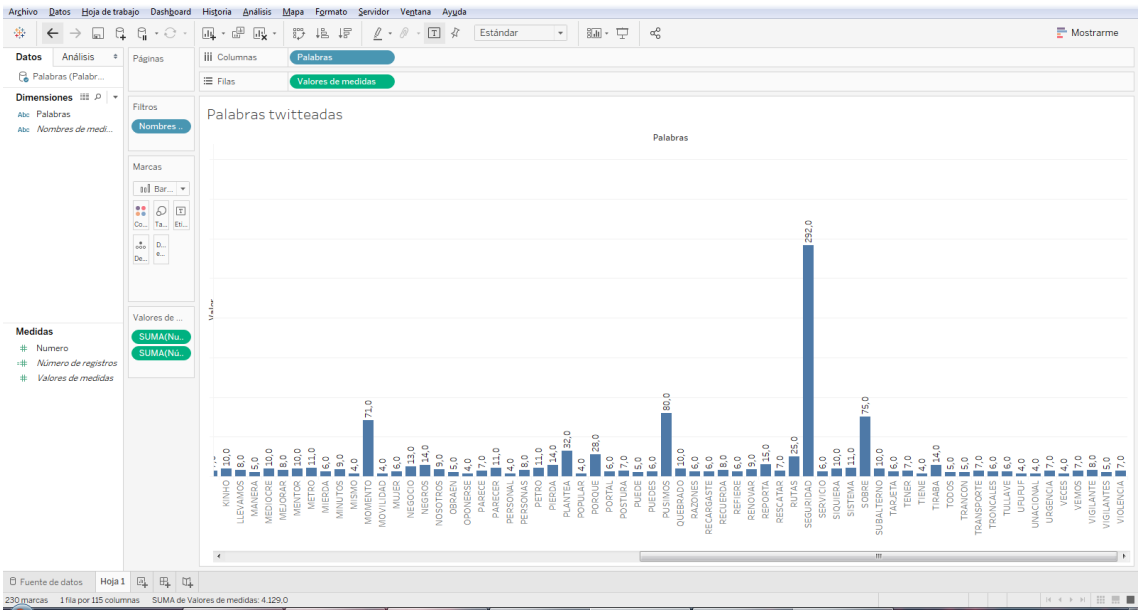


Figura 53. Grafica Palabras Más Comunes-2. Fuente Autor.

8. CONCLUSIONES

Big Data e Business Intelligent es la nueva era de las tecnologías de la información, los datos literalmente son poder sobre cualquier campo en el que se aplique, depende de ellos conocer, identificar, analizar y tomar decisiones con base a ellos, como se evidencio en este estudio de la problemática que presenta el transporte masivo de TransMilenio, se evidencia que la mayoría de comentarios y/o emociones son negativas y de insatisfacción, los causantes de estas inconformidades son los entes que rigen el sistema, las causas son variadas y llegar a soluciones por medio de este tipo de estudios son de gran viabilidad, debido a que la información es contundente a la hora de examinar una situación.

9. REFERENCIAS

[1]. https://blogs.solidq.com/es/big-data/que-es-mapreduce/

[2]. https://www.edureka.co/blog/apache-hadoop-hdfs-architecture/

[3]. https://economipedia.com/definiciones/mineria-de-datos.html

[4]. https://www.sinnexus.com/business_intelligence/datamining.aspx

[5]. https://es.wikipedia.org/wiki/Microblogging

[6]. https://www.juancmejia.com/marketing-digital/estadisticas-de-redes-sociales-usuarios-de-facebook-instagram-linkedin-twitter-whatsapp-y-otros-infografia/

[7]. http://www.hisocial.com/esp/blog/cuantos-usuarios-tiene-facebook

[8].https://mott.marketing/origen-historia-e-informacion-completa-sobre-la-red-social-twitter/

[9]. https://hipertextual.com/archivo/2011/03/historia-twitter/

[10]. https://www.datanalytics.com/libro_r/programacion-en-r.html

[11]. https://help.tableau.com/current/pro/desktop/es-es/stories.htm

[12]. https://www.brandwatch.com/es/blog/analisis-de-sentimiento/

[13]. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[14]. https://www.juancmejia.com/redes-sociales/social-big-data-definicion-e-importancia-de-big-data-en-redes-sociales/

[15].https://www.trecebits.com/2019/01/25/6-herramientas-de-analitica-para-redes-sociales-que-deberias-conocer/

[16].https://www.infotecarios.com/mineria-opinion-una-tecnica-analisis-informacion-linea/#.XXRNoy5Kj4Y