Tipo de artículo: Artículo original

Desarrollo de un análisis de sentimientos sobre las tendencias de la red social twitter utilizando técnicas de analítica de datos

Development of an analysis of feelings about twitter social network trends using data analytical techniques

Adith Perez Orozco[1*] , **https://orcid.org/0000-0002-2149-1625**
Alvaro Oñate [2] , **https://orcid.org/0000-0002-6356-2161**
Jaime Ospino Navarro [3], **https://orcid.org/0000-0001-5419-3459**
Jose Molina Santiago [4] , **https://orcid.org/0000-0001-5708-2818**

[1]Departamento de Ingeniería de Sistemas- Universidad Popular del Cesar. adithperez@unicesar.edu.co
[2]Departamento de Ingeniería de Sistemas- Universidad Popular del Cesar. alvaroonate@unicesar.edu.co
[3]Departamento de Ingeniería de Sistemas- Universidad Popular del Cesar. jjesusospino@unicesar.edu.co
[4]Departamento de Ingeniería de Sistemas- Universidad Popular del Cesar. jrmolina@unicesar.edu.co

*** Autor para correspondencia: adithperez@unicesar.edu.co**

Resumen

Surge la necesidad de implementar técnicas de minería de opinión, para poder identificar el sentimiento con respecto a un tema específico. Muchas personas están desinformadas respecto a temas actuales, por tal razón se crea una herramienta capaz de medir y clasificar opiniones de los usuarios, acerca de un tema controversial. Como solución a dicha necesidad surge la idea de crear y experimentar distintos modelos descriptivos y predictivos, optando por encontrar el modelo analítico óptimo para desarrollar el proyecto. El seguimiento se enmarca en la metodología CRISP-DM, cumpliendo cada una de las fases y las actividades que la metodología provee, cada paso desarrollado de manera dinámica, se regresó a la fase anterior cuando se requirió. El resultado final pudo demostrar que los modelos implementados cumplieron con los requisitos de la evaluación y los resultados fueron óptimos ante las diferentes técnicas implementadas, el estudio de la información descargada fue clave para lograr una implementación limpia y con un margen de error estándar.

Palabras clave: Twitter; sentimiento; usuarios; fases

Abstract

The need arises to implement opinion mining techniques, in order to identify the sentiment regarding a specific topic. Many people are uninformed regarding current issues, for this reason a tool is created capable of measuring and classifying the different opinions of users, about a controversial issue that occurs in the country. As a solution to this need, the idea arises of creating and experimenting with different descriptive and predictive models, choosing to find the most optimal analytical model to develop the project. The monitoring of this project is part of the CRISP-DM methodology, fulfilling each of the phases and activities that the methodology provides, each step was developed dynamically, that is, it was returned to the previous phase when required. Ultimately, the final result of the project was able to demonstrate that the implemented models met the evaluation requirements and the results were optimal given the different techniques implemented, the study of the information downloaded was key to achieve a clean implementation and with a margin of standard error.

Keywords: *Twitter; feeling; users; phases*

Recibido: 20/12/2021
Aceptado: 18/03/2022
En línea: 01/06/2022

Introducción

Hoy día, hablar de análisis de sentimientos es un tema que no está del todo claro, puesto que, con el avanzar de las redes sociales este tipo de minería de opinión se ha convertido en una herramienta muy utilizada por las personas y el entorno empresarial, evolucionando la forma de pensar de los usuarios en un nuevo enfoque de analítica de texto que utiliza técnicas de clasificación para extraer tanto actitudes como polaridades de un texto, logrando transformarlo en conocimiento. (Montesinos García, 2014)

En este orden de ideas, se debe tener en claro el concepto de análisis de sentimientos, dado que, es una tendencia que ha generado una respuesta a patrones de incertidumbre presentados por las polaridades y opiniones expresadas por los usuarios, las cuales son generalizadas sobre algún tema específico, representando una nueva perspectiva de trabajo para muchas organizaciones, que implementan este tipo de técnicas de minería de datos para realizar análisis de texto y conocer la incidencia de las publicaciones que los usuarios realizan, en cuanto un producto, un gusto, un tema o un servicio.

Por otro lado, para situar y argumentar los temas que componen este trabajo de grado se ha estructurado su contenido en tres partes: la primera, la integran la formulación de la propuesta de grado, en la segunda parte se encuentran tanto la parte teórica como la metodología de desarrollo y el tipo de estudio; por último, se encuentra el desarrollo de las conclusiones.

En la formulación de la propuesta se analizaron varios puntos como son: el estado del arte, el planteamiento del problema, la justificación del proyecto, entre otros entes; los cuales hacen parte fundamental de todo el trabajo de grado. Asimismo, se realizó un análisis de las teorías de diversos autores, vistas desde un enfoque documental, formando una estructura sólida tanto en conceptos como en referencias; visualizando tanto el objeto de estudio como el problema que presentaba la identificación de tendencias bajo la realización de un análisis de sentimiento en la red social twitter, logrando así establecer una opinión, tanto positiva como negativa de las emociones descritas por los usuarios en redes sociales, determinando en el análisis realizado, la base sobre la que se sostiene la investigación y la viabilidad de la misma.

En el desarrollo de la teoría y la metodología utilizada en el trabajo de grado, se estableció el uso de la metodología Crisp-DM, argumentando la aplicación de modelos de minería de datos por los cuales se sustenta la investigación.

Se realizó una interpretación de los planteamientos encontrados en la investigación y las particularidades encontradas en el objeto de estudio, donde se resaltó la importancia del uso de la minería de opinión en el contexto de las redes sociales, generando tanto datos como reflexiones que servirán para actualizar el marco interpretativo. Por otra parte, por la naturaleza del objeto de estudio, las conclusiones se trabajaron bajo la construcción de modelos analíticos tanto descriptivos como predictivos en un orden dinámico, donde el desarrollo de las redes sociales va creciendo al ritmo de las tecnologías digitales, generando un impacto en las prácticas culturales de los usuarios día a día.

Con el crecimiento que han presentado las redes sociales, es importante destacar el aporte que hacen a la vida cotidiana, así mismo resaltar la gran cantidad de información que nos brindan, tanto para comunicarnos como para estar informados en tiempo real sobre las principales tendencias. (Next_u, 2019)

En este contexto, se debe agregar que, con el uso de las redes sociales el entorno de la tecnología ha cambiado el diario a vivir a nivel mundial, con los amplios volúmenes de datos, sondeos de opinión expresados por diversidad de usuarios conectados, donde las noticias y tendencias mundiales están a la orden del día aumentando la necesidad de estar conectado a los portales comunicativos e informativos transformando la tecnología en un medio dinámico, ágil e informativo, utilizando la temática tanto de noticias como de opiniones de forma abierta y explicita para cualquier lector, lo que hace replantear la búsqueda de estrategias comunicativas para sobrellevar las situaciones presentadas en el entorno. (Next_u, 2019)

De lo antes expuesto, países como Colombia no escapan de una problemática similar, donde en la actualidad más de 15 millones de usuarios tienen una forma distinta de entender y analizar cada tema o tendencia que atraviesa el país, ya sea de manera virtual o tangible como en este caso serían las redes sociales o en su defecto los periódicos, revistas etc. Del mismo modo, los habitantes que se mantienen informados emiten sus conclusiones u opiniones de forma individual, acción que no conlleva de forma generosa a una toma de decisiones acertada y a la emisión resumida de información a los habitantes. (Colombia.com, 2018)

Así mismo, en la ciudad de Valledupar, por medio de encuestas y sondeos de opinión en redes sociales, las cuales se encuentran soportadas en el anexo I, se evidenció que el 44% de las personas presenta una alta desinformación acerca de temas que son tendencias importantes, y el 95,2% de las personas notaron la necesidad de tener una herramienta que les permita generar un resumen general de sentimientos respecto a una noticia específica, el 29,4%, 16,5%, 20%, y 34,1% de las personas han presentado alto índices de interés sobre información musical, política, deportiva, cultural y económica respectivamente, teniendo en cuenta, los principales contenidos que se emiten en redes sociales, como

los eventos que ocurren en nuestro país, donde el usuario final o el analista de la información, al momento de recibir el mensaje no conoce el impacto que este ha causado en la actualidad, dado que, la información generada en los portales comunicativos e informativos no es resumida a cuanto a argumentos de opinión se refiere ya que el contenido de las noticias es explícito y abierto al lector o televidente, por lo que se hace imprescindible replantear la búsqueda de estrategias o soluciones integrales para poder sobrellevar esta situación. (Colombia.com, 2018)

De no realizarse un análisis de sentimiento en la red social twitter no se podría obtener un análisis de la información suministrada para establecer una tendencia, tanto positiva como negativa de las emociones descritas por los usuarios en redes sociales.

En consecuencia, se pretende establecer estrategias de solución como son la adaptación de resúmenes de información de forma gráfica, emitiendo un grafo tanto positivo como negativo de acuerdo el tema que el usuario desee, con esto se obtendrían estadísticas para generar informes de actualidad sobre el manejo de temas o tendencias en redes, asimismo, otra estrategia de solución es usar todo el volumen de información que los usuarios de la red social twitter manejan, para descubrir patrones, resumir y transformar toda la información captada en tendencias e información de interés, las cuales son abiertas a cualquier público o usuario que tenga dominio de las redes sociales. (Mir Montserrat, 2015)

Materiales y métodos

A continuación, se muestran algunas de las técnicas que sirven para la visualización de la información recolectada después del procesamiento. Para el desarrollo de este proyecto se efectuó una investigación de tipo transaccional o transversal descriptiva, cuyo objetivo según el investigador Hernández Sampieri es "indagar la incidencia de las modalidades o niveles de una o más variables en una población", es decir, el estudio no se lleva a cabo con una manipulación de variables, sino en la observación de fenómenos tal y como son en su contexto sin ser intervenidos, para luego analizar los datos observados, con el fin de proporcionar una visión de alguna situación. (Hernández Sampieri, 2014).

Así mismo en este proyecto también se utiliza la investigación predictiva, dado que el fin de esta investigación es predecir la dirección futura de los acontecimientos indagados, la cual consiste en prever situaciones futuras, brindar información relevante y conocer posibles repercusiones, a partir del estudio de los eventos propuestos en el contexto en el cual está enmarcado y de la probabilidad de que esos eventos puedan presentarse. (Hurtado de Barrera, 2000)

La población en este proyecto es la sociedad colombiana, todas aquellas personas a las que les interesan los temas más relevantes de nuestro país, que requieren estar informados sobre las noticias más recientes y sobre todo las tendencias.

La muestra que tomamos fueron los tweets generados por los usuarios colombianos desde el mes de enero, hasta el mes de mayo del año 2020, que gracias a sus publicaciones y opiniones se llevó a cabo la investigación. (Díaz de León, 2016)

Para la recolección de información que garantice la correcta ejecución del proyecto, se hizo uso de una API, que permite a los desarrolladores crear aplicaciones (software), que se integren con Twitter y de esta manera, mediante la implementación de un framework, hacer peticiones al punto de conexión, para extraer las publicaciones que los usuarios deciden compartir de forma pública. (Twitter, Centro de ayuda, 2020)

La metodología utilizada para este proyecto es la CRISP-DM (Cross-Industry Standard Process for Data Mining), la cual es la más idónea para el desarrollo de proyectos de minería de datos, se explica cómo usa serie de pasos en forma jerárquica, compuesta por cuatro niveles de abstracción, es decir de lo global a lo concreto: Fase, tarea general, tarea específica e instancia de proceso. (Chapman, y otros, 2000)

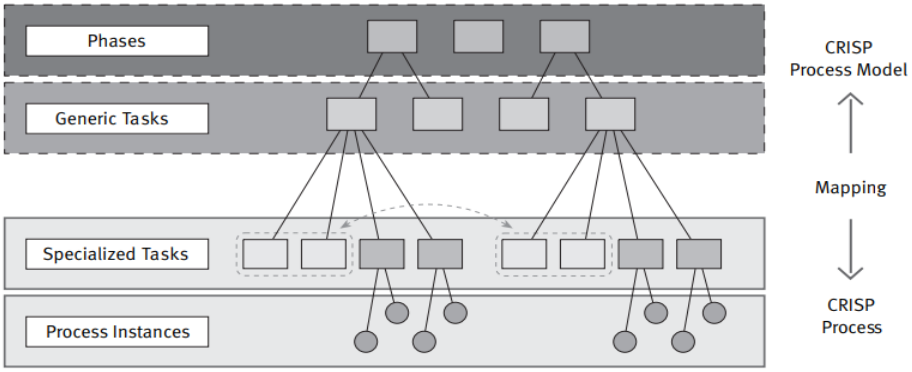

Figura 1. Niveles de la metodología CRISP

Para empezar, el primer nivel está estructurado con varias fases, que tienen a su vez en un segundo nivel unas tareas genéricas. Este nivel dos recibe el nombre de genérico porque busca ser lo más amplio posible para abarcar todos los casos de minería de datos por haber. Así mismo, en el segundo nivel las tareas genéricas tienen que estar lo más completas y estables posibles, lo cual significa que puedan cubrir el proceso de creación del análisis, que se adapten a nuevos cambios y los entornos donde sea aplicable el análisis, esto quiere decir que el modelo debe ser flexible a modificaciones y adaptable para posibles nuevos análisis.

También, en el tercer nivel, se hallan las tareas especializadas, es aquí donde se describe la manera correcta de hacer las tareas genéricas en situaciones puntuales. Por ejemplo, en el nivel dos puede existir una tarea llamada limpiar datos, en este nivel tres se describe como esta tarea encaja en diferentes situaciones, tales como limpiar los caracteres especiales, limpiar los valores de tipo numérico, o si se trata de un tipo de modelado diferente, descriptivo o agrupamiento. Por último, en el cuatro se encuentra la instancia de proceso, que es el resultado de una minería de datos, donde quedan registradas las acciones y las decisiones tomadas en el transcurso del desarrollo. Organizándose de acuerdo a las tareas descritas en niveles previos, representando lo sucedido en específico, en vez de lo que sucede en general.

Fases de crisp-dm: El ciclo de vida de un proyecto de minería de datos con CRISP-DM, consiste en 6 fases, como se observa en la figura 2. La secuencia en la que avanzan las fases es flexible, es decir que se puede avanzar y retroceder de una fase a otra. El producto de una fase describe que fase o tarea específica debe desarrollarse. Las flechas indican las dependencias, el orden y la secuencia que se debe seguir.

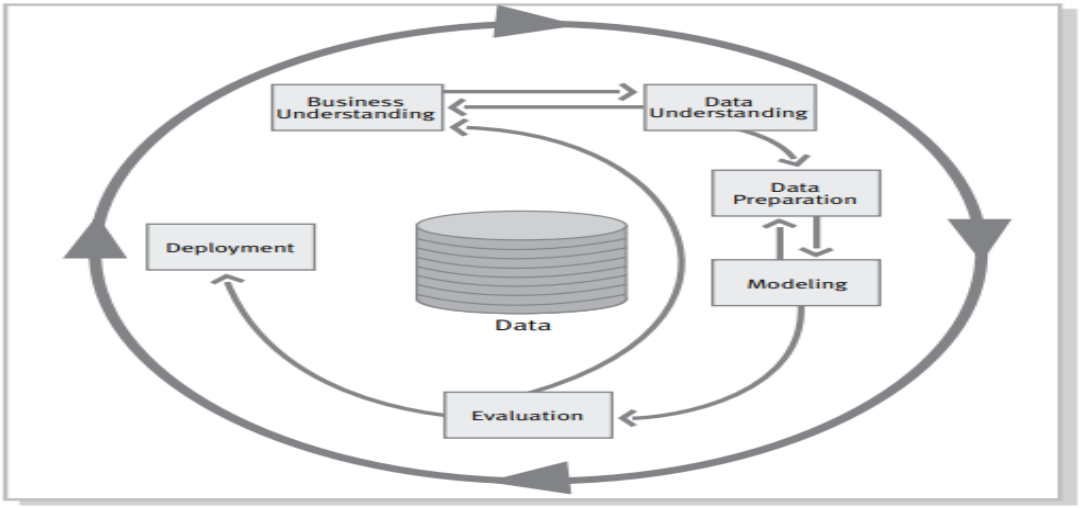

Figura 2. Fases del modelo de referencia CRISP-DM

El círculo más grande en la figura 2, representa la esencia del ciclo de minería de datos como tal. A continuación, un breve resumen de cada fase:

Entendimiento del negocio.

Esta primera fase se enfoca en conocer las necesidades y los objetivos del proyecto desde el punto de vista del negocio, posteriormente transformar ese conocimiento en una descripción de un problema de minería de datos y un plan inicial diseñado para cumplir los objetivos.

Comprensión de los datos

La comprensión de los datos inicia con la obtención de los datos y seguidamente se realizan procedimientos que posibiliten familiarizarse con los datos, hallar fallas en la calidad, encontrar, tendencias, descubrir las primeras características en los datos y/o detectar patrones ocultos que permitan formular hipótesis.

Preparación de los datos

En esta fase se desarrollan todas las actividades necesarias para tener como resultado en conjunto de datos preparado (dichos datos serán ingresados a la herramienta de modelado) partiendo de los datos iniciales. Las tareas de esta fase, probablemente sean realizadas varias veces y no se tiene un orden establecido de realización. Las tareas pueden ser, la selección de registros, atributos o tablas, también la limpieza y transformación de los datos para luego pasar por las herramientas que modelan.

Modelado

En esta fase se escogen y se utilizan diferentes técnicas de modelado, por medio de las cuales se configuran los parámetros, para obtener valores óptimos. Sin embargo, existen varias técnicas que se pueden utilizar para resolver un mismo problema de minería de datos. Algunas de las técnicas requieren un tipo de dato en específico o que los datos estén en una forma para que sea posible su tratamiento, es por esto que es necesario a menudo retroceder a la fase de preparación de los datos.

Evaluación

En esta etapa se tiene un modelo escogido, que se espera que se optimo y precisa desde una perspectiva de análisis de datos.

Antes de continuar con el lanzamiento del modelado final, es primordial evaluarlo y revisarlo detenidamente, chequear cada uno de los pasos que requiere el modelo, para confirmar que cumpla con los objetivos comerciales. Un objetivo principal es definir si hay algún problema del negocio que no se haya tomada en cuanta. Finalmente, en esta fase se toma la decisión sobre el uso del resultado de la minería de datos.

Despliegue

La evaluación del modelo por lo general no es el final del proyecto, incluso si el fin del modelo es incrementar el conocimiento de los datos, el conocimiento deberá organizarse de manera que al presentarlo sea entendible por el cliente y sea adecuado para su uso. Lo que implica la implementación de modelos en tiempo real dentro de los procesos de toma de decisiones de una empresa. Por ejemplo, personalizar una página web o la obtención de bases de datos de mercadeo. Depende de los requisitos, esta fase puede resultar tan fácil como generar un informe para el cliente o tan complicado como la realización repetida del proceso de minería de datos en la empresa. En la mayoría de

los casos es el cliente quien lleva el proceso de implementación, que por lo general es el analista quien lo realiza, pero es importante que el cliente sepa que acciones realizar para ejecutar eficazmente los modelos creados.

La figura 3, muestra un cuadro con las fases y cada una de las tareas genéricas con sus respectivas salidas.

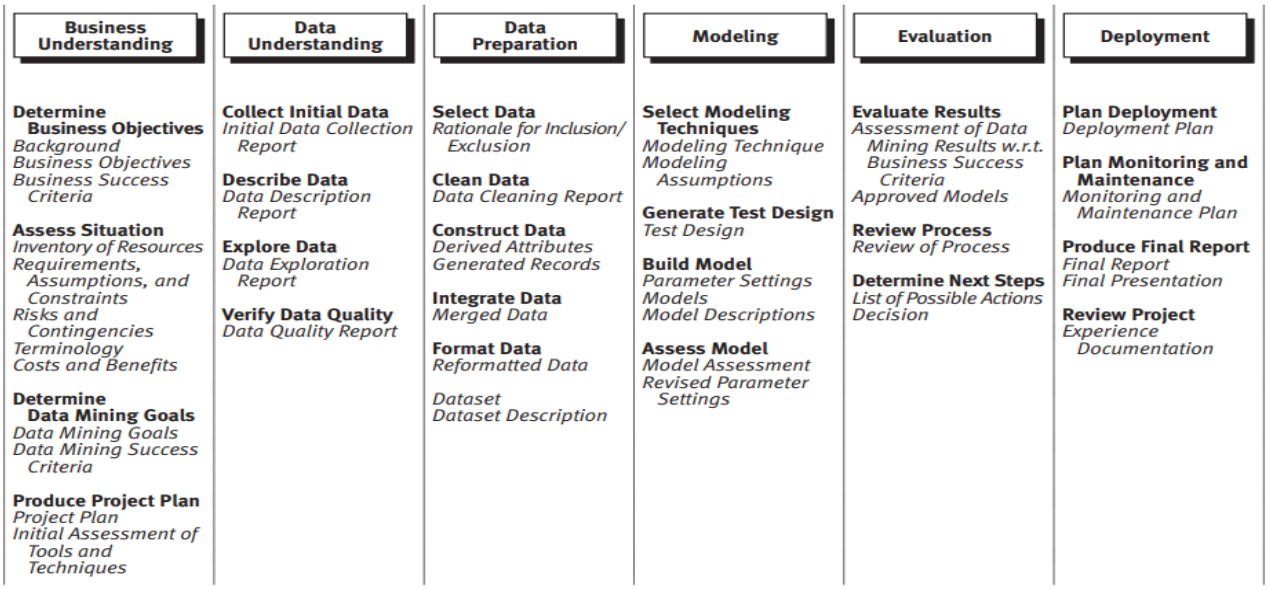

Figura 3. Tareas genéricas y salidas del modelo de referencia CRISP-DM.

El proyecto se desarrolló con los datos de Twitter, siendo esta la fuente de los datos a ser analizada, estos datos son las opiniones de los usuarios acerca de cualquier tema, los cuales son resumidos para generar valor informativo a empresas como periódicos, radios y portales informativos en la web.

Twitter cuenta con 393 millones de usuarios en todo el mundo, según (Statista, Statista, 2020), para crear una cuenta, se necesita un correo electrónico, un numero de celular y datos personales básicos. Permite publicar mensajes o noticias con un máximo de 240 caracteres, que puede contener emoticones, videos e imágenes llamados tuits, que se visualizan en la página principal del usuario, además permite republicar o compartir un tuit de cualquier usuario en el perfil del usuario que lo realice, llamado retuit. Inicialmente la herramienta a utilizar para el desarrollo del proyecto es RStudio, que es un entorno de desarrollo integrado (IDE), desarrollado para facilitar la utilización del lenguaje de programación R, que permite la utilización de modelos estadísticos, desde regresión lineal hasta redes neuronales, modelación descriptiva, predictiva, visualización de gráficas. Implementado por empresas e industrias que requieren un tratamiento de su información para sacar un beneficio económico.

Así mismo, las técnicas a utilizar en la realización de este proyecto son: Agrupamiento y clasificación.

Agrupamiento: Es una técnica de aprendizaje no supervisado, cuyo objetivo principal es crear clusters de datos, basados en variables similares entre estos, a fin de identificar datos atípicos, este proceso se realiza con el fin de llegar a una descripción sintética; Esta descripción sintética, se obtiene sustituyendo la descripción de cada uno de los elementos del grupo, por la de un único representante característico del mismo.

Clasificación: Es una técnica de aprendizaje supervisado, que permite la aplicación de modelos, expertos en agrupar datos en conjuntos, con la finalidad de predecir el conjunto al que pertenece un objeto desconocido. Se basa fundamentalmente en la partición del conjunto de datos en dos subconjuntos más pequeños, que luego se emplearan con los siguientes objetivos: entrenamiento y test. El subconjunto de datos para el entrenamiento es utilizado para estimar los parámetros del modelo y el subconjunto de datos de test se emplea para comprobar el comportamiento del modelo estimado

Recolección inicial de los datos: La obtención de los datos, es el factor más importante para el cumplimiento de los objetivos del proyecto, conviene subrayar que twitter dispone de una serie de servicios, que permiten a los desarrolladores o investigadores establecer un punto de conexión con su api, a través de la herramienta RStudio, se obtuvieron los tweets en tiempo real, dando cumplimiento al objetivo número uno. (Twitter, Centro de ayuda Twitter, 2020)

De igual importancia, para la obtención de los datos, es necesario cumplir con los requerimientos de autenticación que dispone twitter, tales como los tokens y las llaves de conexión, estos factores permiten el acceso a los servicios del api y se obtienen en el registro de usuario Developer.

Por consiguiente, una vez obtenida la conexión directa con el api, se identifican los temas tendencias en Colombia, además es importante conocer el código identificador del país para realizar la búsqueda y descarga en twitter del tema elegido, en este caso el tema será el número uno de las tendencias en Colombia del día 10/11/20: #colombiaEsprovida

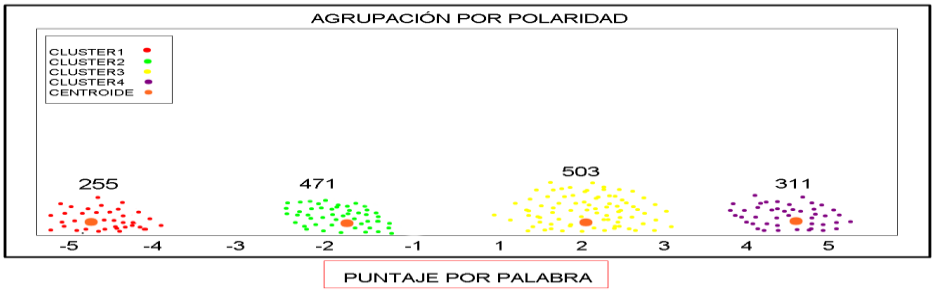

Figura 4. Tareas genéricas y salidas del modelo de referencia CRISP-DM.

Descripción de los datos: Una vez descargados los datos, son almacenados en un dataframe compuesto por filas y columnas, en el que cada fila corresponde a un objeto de los datos y cada columna a una variable o característica del conjunto de datos. Dado el planteamiento anterior, se procede a describir cada uno de los atributos que componen a la información descargada, en este caso los tweets obtenidos.

Tabla 1. Descripción del dataframe

Total de datos requeridos	3000 (parámetro dinámico)
Total de datos descargados	1540
Numero de filas	1540
Numero de columnas	15
filas con valores N/A	1
Total de atributos numéricos	9
Total de atributos de texto	3
Total de atributos de fecha	1
Total de atributos de imagen	1
Total de atributos alfanuméricos	1
Total de atributos	15
Sample algorithm	Sample performance
A	Bad
B	Excellent
C	Unstable
D	Unforgettable

Tabla 2. Descripción atributo

Atributo	Descripción	Tipo de dato
1. User_id	Identificador del usuario que realizó la publicación.	Numérico
2. Status_id	Identificador de la publicación.	Numérico
3. Creat_at	Fecha en que se creó la publicación.	Fecha
4. Screen_name	Nombre del usuario que originó la publicación.	Texto
5. Text	Almacena el contenido de la publicación realizada.	Texto
6. Source	Desde que dispositivo se originó la publicación.	Texto
7. Display_text_width	Total de caracteres que componen al tweet.	Numérico, limite 1000
8. Favorite count	Total de Tweets que le han gustado al usuario durante la vida de la cuenta.	Numérico
9. Retweet count	Cantidad de Retweets que ha obtenido un tweet en especifico	Numérico
10. Location	Lugar donde se origina el tweet	Numérico
11. Follower count	Total de seguidores de la cuenta en la que se originó el tweet	Numérico

twitter		
12. Friends_count	Total de personas que el usuario sigue.	Numérico
13. Listed_count	Total de listas creadas por el usuario para subdividr el tiempo de lectura de los tweets.	Numérico
14. Status_url	Almacena el link que contiene la dirección especifica del tweet publicado.	Alfanumérico con hipervínculos
15. File image url	imagen del usuario que origina el tweet	Archivo

La exploración de los datos, es una actividad que consiste en representar las principales características de la información obtenida mediante, técnicas de consulta, visualización, generación de informes y análisis estadísticos simples, con el objetivo de destacar las relaciones y propiedades entre atributos.

De acuerdo a lo anterior, se procede a escoger los atributos clave del dataframe, es decir, que generan valor a la investigación, por otro lado, los atributos que no son tenidos en cuenta, es porque contienen imágenes o hipervínculos que no aportan nada al estudio. Los atributos seleccionados para la exploración son los siguientes.

Tabla 3. Atributos seleccionados para la exploración de los datos

Atributos seleccionados
1. Text
2. Source
3. Display_text_width
4. Screen_name
5. Retweet count
6. Location
7. Follower count twitter

Resultados y discusión

En esta fase se determina si los modelos cumplen con el objetivo para el que fueron creados y decidir si los modelos son los óptimos para la realización del proyecto o son deficientes. A continuación, se muestran los objetivos del negocio y los criterios de éxito de la minería de datos propuestos en la fase de entendimiento del negocio según la metodología CRISP-DM.

Objetivos de la mineria de datos.

- Construir un conjunto de datos en tiempo real para determinar las tendencias e influencia en distintos temas controversiales en la red social Twitter.

- Construir y evaluar un modelo analítico descriptivo para identificar las opiniones de los usuarios en la red social Twitter.
- Construir y evaluar un modelo analítico predictivo para clasificar las opiniones de los usuarios en la red social Twitter
- Desarrollar un framework para la visualización de los tweets en tiempo real que resuman las opiniones de los usuarios en la red social Twitter

Criterios de éxito de la mineria de datos

- El grado de fiabilidad de los modelos es de 70%.

Por medio del algoritmo de agrupamiento (clustering), se logró cumplir con el objetivo dos de la minería de datos y con el primer criterio de éxito, puesto que se agruparon con éxito todos los tweets descargados, dejando los siguientes resultados.

- Se agruparon los tweets en 4 categorías.

Ahora bien, con relación al objetivo número tres de la minería de datos, se realizó la clasificación de los tweets de manera exitosa. El proceso de clasificación se llevó a cabo con una técnica de clasificación bayesiana, Naive Bayes. De modo que, dicha clasificación se logró con ayuda de un diccionario personalizado en español, construido a partir de palabras más utilizadas en el vocabulario y en la jerga colombiana, permitiendo así obtener resultados más coherentes y aceptables durante el experimento. Se obtuvo el siguiente resultado.

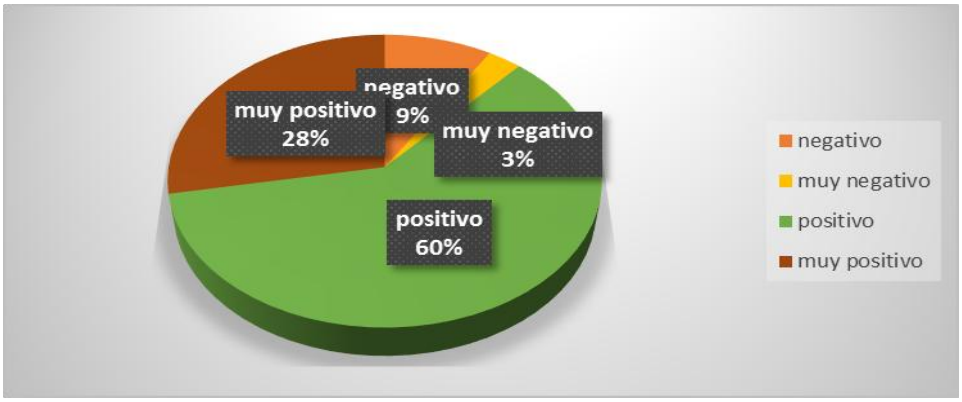

Figura 5. Criterio de éxito de polaridad de la tendencia

De acuerdo con la figura 5, que expresa, que el 28% de la población realizo publicaciones muy positivas al tema tendencia #MingaSomosTodos, el 9% de la población realizo publicaciones negativas, el 3% realizo publicaciones

muy negativas y el 60% realizo publicaciones muy positivas. Concluyendo así que el tema generó un impacto positivo en nuestro país. De igual manera con todos los temas tendencia.

Despliegue

En esta etapa, es necesario determinar una estrategia para poner en funcionamiento el modelo, es decir, la creación de un framework que permita al usuario final una herramienta capaz de recopilar todos los datos y convertirlos en información valiosa, para empresas o personas interesadas en el sentimiento que genera cierto tema en la red social Twitter, también se debe determinar los pasos necesarios y cómo realizar estos pasos, para lograr un despliegue continuo y así obtener como resultado un fácil mantenimiento del framework, el siguiente proceso, se realiza con el fin de planificar y desarrollar de mejor forma la carga de archivos y modelos, en esta etapa, también se conocen los procesos y las herramientas que tendrán intercesión en el proceso de despliegue, para llevar una mejor planificación y desarrollo de la subida de archivos y modelos. A continuación, se muestra el proceso de creación del framework.

Creación del framework con la herramienta shiny y r studio

En esta etapa del proyecto, haciendo cumplimiento al objetivo número 4, se dará inicio a la estandarización de los diferentes procesos que componen al proyecto de análisis de sentimientos, es importante conocer y tener en cuenta, que el sistema está compuesto por 3 capas que cumplen diferentes funciones desde el punto de vista de la arquitectura de aplicaciones de análisis.

Para la creación interna del framework, se hará uso de la sintaxis de R, ya que, por medio de esta codificación, se programarán las diferentes funciones y servicios que la aplicación va a servir. El objetivo final del framework a crear, es unificar y estandarizar los procesos de análisis de sentimiento, para que los futuros desarrolladores o la comunidad de R y Shiny puedan tener un patrón a seguir y a cumplir para el diseño de este tipo de aplicaciones.

Para el desarrollo del framework se hizo uso de las siguientes herramientas:

SHINY

La ventaja de Shiny, es que permite que el usuario experimente la interactividad, permitiéndole manipular los datos sin tener que manipular el código fuente de la aplicación, ya que todo es manipulado desde la vista, por medio de parámetros en tiempo real. Las aplicaciones web creadas con Shiny pueden estar enfocadas a numerosos ámbitos: investigación, profesional o, por supuesto, la docencia. Estas aplicaciones pueden abrirse desde el propio ordenador, una Tablet o incluso el móvil, garantizando así la experiencia final del usuario con la interfaz interactiva de la aplicación. (RStudio, 2018)

De modo que, para el proyecto presente se hará uso de shiny, la cual facilita las diferentes funciones y servicios para la creación del framework, que ayudará al usuario final a visualizar el análisis de sentimiento de los tweets tendencias según de su interés.

De manera importante, también servirá de modelo o plantilla para la creación de nuevas aplicaciones. La creación del framework lo que busca es estandarizar la programación en este tipo de proyectos de análisis de texto o de tweets, la ventaja principal de este proceso es la reutilización y el estándar en la codificación.

El framework a crear lleva por nombre ShinySentiment la cual se encuentra en su versión v.1 Alfa.

En efecto, el framework consta de 3 capas y está en unas de las arquitecturas de desarrollo más reconocidas a nivel de aplicaciones (Pavón Mestras, 2008), la modelo, vista, controlador.

En el modelo, se ejecuta todo el procesamiento lógico de las entidades a analizar, en este caso los tweets descargados, también se ejecutan las diferentes funciones que hace posible el agrupamiento y la clasificación de los tweets, por medios del score relacionado a cada palabra que compone al tweet, también en esta capa se declaran los métodos, variables y constantes del sistema.

En el controlador se ejecutan todos los servicios de la aplicación, el servicio más importante del controlador, es el establecer el puente de comunicación entre la vista y el modelo, toda petición que se realice en la app web y móvil debe pasar por el controlador, agregando así una capa de seguridad intermedia entre los modelos y las peticiones que realiza el usuario.

En la vista se presenta toda la experiencia de usuario, las interfaces de usuario, las diferentes gráficas y nubes de palabra, también la representación de los tweets descargados. Se puede decir que es la capa donde inicia todo el proceso, ya que el usuario realiza la petición, dicha petición pasa por el controlador, el controlador acciona el servicio, y el servicio acciona la lógica en el modelo, devolviendo así los datos de manera gráfica al usuario. (Pavón Mestras, 2008)

Conclusiones

La redes sociales se han convertido en una herramienta actualmente muy poderosa en el mundo de la información y conexión entre las personas, las diferentes funcionalidades que tiene cada una de las redes sociales como Facebook, Twitter, Instagram, han servido para que las personas, empresas y medios de comunicación hagan uso de la información en tiempo real, generando así nuevos métodos de ventas, seguimiento, estudio de mercadeo, análisis, comentarios, hasta las elecciones locales, nacionales de los dirigentes de cada ciudad o país se han visto involucradas en dichos procesos de la información por medio de las redes sociales.

En primer lugar, se desarrolló una aplicación fascinante, ante la vista del usuario, sobre la cual se pueden consultar las tendencias de Twitter. Los tweets recolectados se clasifican y se presentan resumidos en gráficas, de manera que el usuario puede observar los diferentes resultados de polaridad obtenidos. Esto permite determinar la opinión de los usuarios colombianos a cerca de un tema tendencia expuesto.

En definitiva, con el desarrollo de este proyecto se pudo demostrar que Twitter es una herramienta de gran utilidad, e impacto en la sociedad ya que en esta red social, se publican los temas más importantes de una ciudad, país o de un objeto en estudio actual, las opiniones de los diferentes usuarios han creado una nueva tendencia en el uso de la información, ya que en los comentarios que cada persona realiza, se puede evidenciar el gusto o la opinión que tiene acerca de un tema que esté sucediendo actualmente en el país, es de vital importancia conocer las diferentes opiniones de manera precisa y resumida por medio métodos estadísticos, la información que estos usuarios proveen acerca de un tema en estudio o por aprobación de una ciudad o de un país, en este caso Colombia.

Por otro lado, la investigación se realizó con el fin de aportar a la sociedad, una herramienta para que las personas puedan conocer los diferentes temas tendencias que atraviesa el país. A pesar de la problemática de acceso a la información que atraviesa el mundo actual de las redes sociales, se logró establecer una comunicación exitosa con los servidores de Twitter. Por ende la aplicación creada es una herramienta, que tiende o promueve una nueva manera de informarse, ya que en ella se conoce el impacto que tiene dicho tema tendencia, por medio de análisis semánticos, cálculos estadísticos, aplicando modelos descriptivos y predictivos, precisando grandes volúmenes de datos, agrupando y clasificando las opiniones de los usuarios, con algoritmos entrenados y óptimos para el proceso de agrupación, en este caso K-MEANS Y técnicas como as vector para la partición y estudio descriptivo de la información descargada del api de twitter.

Englobando, para la realización del objetivo de la creación del framework, se utilizó la arquitectura modelo, vista, controlador, esta arquitectura es unas de la más recomendadas en el desarrollo de aplicaciones interactivas y de tipo web, ya que la información viaja en el controlador de manera segura a los servicios y métodos estadísticos para los diferentes procesos de agrupación y renderización a la vista del usuario, para este proceso se utilizaron diferentes herramientas, para que el usuario hiciera uso de la aplicación y obtuviera así los datos en tiempo real. Por otra parte también es importante conocer el comportamiento interno de los procesos, cómo se evalúan y se monitorean para que no haya un error en el tiempo de uso de la aplicación y así el usuario no quede descontento del aplicativo y de sus funcionalidades, para este ítem se hizo uso de la herramienta shiny.io, que nos provee un sin número de funcionalidades gratis para el seguimiento de nuestras aplicaciones y el volumen de peticiones realizadas en el día, en el mes y en el primer semestre del año en curso.

Eventualmente, la creación del diccionario para los diferentes grupos, basados en los puntajes semánticos, es un ítem de vital importancia ya que los algoritmos que vienen por defectos en R o librerías de terceros, están implementados con una agrupación totalmente diferente al idioma español, la agrupación que se realizó en el proyecto se hizo con un diccionario propio y netamente local con palabras de léxico colombiano para lograr puntaje semántico real en la agrupación y mejorar estadísticamente los resultados en monitoreo del modelo que lidera la investigación, en este caso el descriptivo por medio de algoritmos de agrupación.

Por último, se puede agregar que la implementación de modelos descriptivos y predictivos pueden ser de gran utilidad para el desarrollo de nuevas herramientas que promuevan el buen uso de la información de uno o varios temas que sean de carácter tendencia en el país.

Conflictos de intereses

Ninguno.

Contribución de los autores

1. Conceptualización: Alvaro Oñate, Jaime Ospino Navarro, Jose Molina Santiago
2. Curación de datos: Jaime Ospino Navarro, Jose Molina Santiago
3. Análisis formal: Adith Perez Orozco, Alvaro Oñate, Jaime Ospino Navarro, Jose Molina Santiago
4. Adquisición de fondos: Alvaro Oñate
5. Investigación: Adith Perez Orozco, Alvaro Oñate, Jaime Ospino Navarro, Jose Molina Santiago
6. Metodología: Adith Perez Orozco, Alvaro Oñate, Jaime Ospino Navarro, Jose Molina Santiago
7. Administración del proyecto: Jaime Ospino Navarro, Jose Molina Santiago
8. Recursos: Adith Perez Orozco, Alvaro Oñate, Jaime Ospino Navarro, Jose Molina Santiago
9. Software: Jaime Ospino Navarro, Jose Molina Santiago
10. Supervisión: Alvaro Oñate
11. Validación: Adith Perez Orozco
12. Visualización: Jaime Ospino Navarro, Jose Molina Santiago
13. Redacción – borrador original: Adith Perez Orozco, Alvaro Oñate, Jaime Ospino Navarro, Jose Molina Santiago
14. Redacción – revisión y edición: Adith Perez Orozco

Financiamiento

Universidad Popular del Cesar

Referencias

ARCMAP. ArcGis for Desktop. [En línea]. 2016 [Consultado el: 2020 de 11 de 18. Disponible en: https://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-analyst-toolbox/how-dendrogram-works.htm

CARRAPINI, F. S. Sistemas Basados en Reglas. [En línea]. 2019. [Consultado el: 28 de 04 de 2020. Disponible en: http://www.cs.us.es/~fsancho/?e=103

CARRILLO, A., URUEÑA, J., FORERO, J., & CAICEDO, L. Análisis Del Sentimiento Político Mediante La Aplicación De Herramientas De Minería De Datos A Través Del Uso De Redes Sociales. Repositorio Pontifica Universidad Javeriana De Bogota, 2015.

CHAPMAN, P., CLINTON, J., et. Al. [En línea]. 2000. Step-by-step data mining guide. Computer Science. [Consultado el: 28 de 04 de 2020. Disponible en: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

COLOMBIA.COM. [En línea]. 2018. ¿Qué tanto usan los colombianos Twitter? [Consultado el: 28 de 04 de 2020. Disponible en: https://www.colombia.com/tecnologia/aplicaciones/que-tanto-usan-los-colombianos-twitter-209225

CONTRERAS, L., ROSALES, K. Analysis of the Behavior of Customers in the Social Networks Using Data Mining Techniques. IEEE xplore. 2016

DÍAZ DE LEÓN. N Población y muestra. Respositorio institucional de la Universidad Autonoma del Estado de Mexico. 2016

DZUL, M. [En línea]. 2010. Aplicación basica de metodos cientificos. Diseño no experimental. Sistema virtual de la universidad autonoma del estado de Hidalgo. [Consultado el: 28 de 04 de 2020. Disponible en: https://www.uaeh.edu.mx/docencia/VI_Presentaciones/licenciatura_en_mercadotecnia/fundamentos_de_metodologia_investigacion/PRES38.pdf

ESCORTELL PÉREZ, M. A. El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter. Repositorio de la Universidad Politecnica de Valencia. 2017

FEYNMAN, R. [En línea] 2018 Naive Bayes Classifier. Towardsdatascience. [Consultado el: 30 de 04 de 2020. Disponible en: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

GO, A., BHAYANI, R., HUANG, L.Twitter Sentiment Classification using Distant Supervision. Repositorio de la Universidad de Stanford. 2010

HERNANDEZ ORALLO, J. Introducción a la mineria de datos. Madrid: Pearson Prentice Hall. 2004. 656, p.

HERNÁNDEZ SAMPIERI, R. Metodología de la investigación. Ciudad de México: McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V. 2014. 634,p.

HURTADO DE BARRERA, J. Metodologia de la investigacion Holistica. Caracas: Fundación SYPAL. 2000. 666,p.

KUSHAL, D., LAWRENCE, S., PENNOCK, D. [En línea]. 2003. Mining the Peanut Gallery: Opinion Extraction and

Semantic Classification of Product Reviews. [Consultado el: 28 de 04 de 2020. Disponible en: https://www.kushaldave.com/p451-dave.pdf

LEÓN GUZMÁN, E. [En línea]. 2019. Métricas para la validación de Clustering. Universidad Nacional de Colombia, facultad de ingeniería. [Consultado el: 28 de 04 de 2020. Disponible en: https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf

LÓPEZ, C. P. Minería de datos: técnicas y herramientas. Google books, 1-2. 2007

MEDINA, F., GÓMEZ, C. [En línea]. 2014. Funcionalidades de la minería de datos. Dialnet. [Consultado el: 30 de 04 de 2020. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=5432252

MENESES, M., MARTÍN DEL CAMPO, A., RUEDA, H. [En línea]. 2018. #TrumpenMéxico. Acción conectiva transnacional en Twitter y la disputa por el muro fronterizo. Dialnet. [Consultado el: 30 de 04 de 2020. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=6353337

MESAS JÁVEGA, R. M. Análisis de tendencias y marcas deportivas a través de twitter. Repositorio Universidad Autonoma de Madrid. 2015

MIR MONTSERRAT, D. Analitica de datos en twitter. Repositorio Universidad Autonoma de Barcelona. 2015

MONTESINOS, L. Analisis de sentimientos y predicción de eventos en twitter. Repositorio Unversidad de Chuke. 2014

MONTOYA, D. Análisis de sentimientos a través de Twitter. Repositorio de la universdad internacional de la Rioja. 2019

Next_u.[En línea]. 2019 Redes sociales: ventajas y desventajas. [Consultado el: 01 de 05 de 2020. Disponible en: https://www.nextu.com/blog/redes-sociales-ventajas-y-desventajas/

OROZCO, B. [En línea] 2017. Gruplac GISICO. [Consultado el: 30 de 04 de 2020. Disponible en: https://scienti.minciencias.gov.co/gruplac/jsp/visualiza/visualizagr.jsp?nro=00000000002099

PAK, A., & PAROUBEK, P. [En línea] 2010 Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Semantic Scholar. [Consultado el: 30 de 04 de 2020. Disponible en: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

PANG, B., & LEE, L. Opinion Mining and Sentiment Analysis. Editor y Prensa internacional "Now". 2008

PAVÓN, J. [En línea] 2008 Estructura de las aplicaciones orientadas a objetos el patrón Modelo- Vista- Controlador. Respositorio de la Universidad Complutense Madrid. [Consultado el: 30 de 04 de 2020. Disponible en: https://www.fdi.ucm.es/profesor/jpavon/poo/2.14.mvc.pdf

PIÑON , L. Minería de datos aplicada a Twitter y análisis de sentimientos mediante algoritmos de inteligencia artificial. Respositorio de la Universidad de Jaume I, España. 2018

RSTUDIO. [En línea] 2018. Shiny. [Consultado el: 30 de 04 de 2020. Disponible en: https://shiny.rstudio.com/

RODRIGUEZ ALDAPE, F. M. Cuantificación del interés de un usuario en un tema mediante mineria de texto y análisis de sentimiento. Repositorio de la Universidad Autonoma de Nuevo León. 2013

RUIZ RAMÍREZ, Á. Mineria de datos en redes sociales y pymes. Repositorio institucional de trabajos académicos de la Universidad de Jaén. 2018

RUIZ, D., DELGADO, Y. Prototipo de herramienta de software que permite realizar minería de opinión en español utilizando un motor de bases de datos no relacional. Repositorio de la universidad distrital Francisco Jose De Caldas. 2018

SÁNCHEZ, P., MARTÍN, M., BLANCO, H. Del Data-Driven Al Data-Feeling: Análisis De Sentimiento En Tiempo Real De Mensajes En Español Sobre Divulgación Científica Usando Técnicas De Aprendizaje Automático. Portal De Revistas De La Universidad Del Rosario. 2019

SELVA, J. Desarrollo de un sistema de análisis de sentimiento sobre Twitter. Universidad Politecnica de Valencia, 1-2. 2015

SOBRINO, J. Análisis de sentimientos en Twitter. Repositorio instotucional de la Universidad Abierta de Catalunya. 2018

SOLARTE, G., SOTO, J. [En línea] 2011. Arboles de decisiones en el diagnóstico de enfermedades cardiovasculares. Redalyc. [Consultado el: 30 de 04 de 2020. Disponible en: https://www.redalyc.org/pdf/849/84922625018.pdf

STATISTA. [En línea] (2020). La empresa detrás del exitoso producto. [Consultado el: 30 de 04 de 2020. Disponible en: https://es.statista.com/acercadenosotros/

TIMOTHY, P., JURKA, COLLINGWOOD, L., AMBER, E. et al. [En línea] 2020. CRAN R Project. [Consultado el: 30 de 04 de 2020. Disponible en: https://cran.r-project.org/web/packages/RTextTools/index.html

TORRES, L. Análisis de sentimientos sobre el posconflicto colombiano utilizando herramientas de mineria de texto. Repositorio Escuela Colombiana de Ingenieria Julio Garavito. Colombia, 2015.

VALCÁRCEL, V. [En línea] 2004. Data Mining Y El Descubrimiento Del Conocimiento. 83. . [Consultado el: 30 de 04 de 2020. Disponible en: https://www.redalyc.org/pdf/816/81670213.pdf