Facultad de Ingeniería

INGENIERÍA INDUSTRIAL

Proyecto de Grado – Segundo Semestre 2019

[193001] Diseño de un modelo para el Análisis de Sentimientos sobre temas de interés de usuarios de Twitter en Colombia

Sofía Pérez Acevedo[a,c] , Luisa Fernanda Rivera Ramirez[a,c]

Juan Carlos García Díaz[b,c]

[a]*Estudiante de Ingeniería Industrial*
[b]*Profesor, Director del Trabajo de Grado, Departamento de Ingeniería Industrial*
[c]*Pontificia Universidad Javeriana, Bogotá, Colombia*

Resumen

Understanding people's opinion or posture is essential to understand the behavior of society regarding a topic. Currently, social media are data sources for opinion mining, sentiment analysis, trend topic detection, and even people profiling; due to that, these digital platforms are being used to publish their opinions on a variety of topics, discuss current problems, express feelings and emotions. These numerous applications require the implementation of machine learning algorithms and natural language processing (NLP) techniques. In this document, we focused on Twitter as a source of information to analyze Twitter user's sentiments of the opinions in Colombia on issues related to the COVID-19 pandemic. The Python programming language was used to develop those techniques.

First, a module for the extraction of data from Twitter was developed. It was used to collect the necessary information for 10 days. Then, a model to identify topic trends in the extracted collection of tweets, using a LDA Topic Modelling library was developed; the "Coherence Value" was used as performance measure. Topics were automatically assigned to tweets; however, a validation was manually made to corroborate that those tweets were really related to the topic; finally, a manual labeling of the topics was made.

A model for sentiment analysis (positive, negative, and neutral) of users on a specific topic was developed. A database constructed by manual tagging of a sample of tweets was used as training data to build such model; then, a comparison of the Support Vector Machine (SVM), Random Forest (RF), Tree Decision (TD), Logistic Regression (LR) and Naive Bayes (NB) algorithms was made in order to choose the best suited to a dataset of information extracted from Twitter. The Logistic Regression (LR) algorithm was finally selected, and the model was constructed using techniques such as data balancing, parameter adjustment, and cross-validation.

Three models for profiling of users based on demographic, gender, occupation, and age variables, were developed. For gender profiling, a first classification filter was made using the user's name; then, a Support Vector Machine (SVM) classification model for the tweets of the unidentified users was used; to develop such model, a collection of tweets from the 2016 PAN internet repository was used as training data. For the occupation model, a first filter was made using the user description; and, then a Support Vector Machine (SVM) classification model for the tweets of the unidentified users was used. A database built with the tweets of Twitter users whose occupation is known was used as training data. For the age model, a Support Vector Machine (SVM) classification model was developed; the 2016 PAN repository was used again as training data. Then, to bring the results to the user level, a voting scheme was used for each variable.

Finally, a database with a total of 560,223 user tweets was obtained; 189,492 of them were from Colombia. Using LDA we obtained four trending topics, including the pandemic of the COVID-19. The average accuracies of the classification models in the test phase were as follows, for the sentiment analysis: 0.82, for the gender model 0.78, for the occupation model: 0.58 and for the age model: 0.86. Then, in order to verify the training dataset, a cross validation was performed, the average accuracies of the classification models in the training phase using cross validation were as follows, for the sentiment analysis: 0.83 ± 0.06, for the gender model 0.76 ± 0.01, for the occupation model: 0.55 ± 0.01 and for the age model: 0.84 ± 0.01. After obtaining the profiling information, a clustering algorithm was implemented to allow segmenting Twitter users according to the characteristics identified in the previous models, and to determine the characteristics that define the users who have any opinion regarding the issue of the management of COVID-19 by the government, obtaining as a result 6 clusters describing these users.

We concluded that it is possible to create a sentiment analysis model on the opinions expressed by Twitter users in Colombia, about current topics of interest, and profile these users in terms of age, gender, and occupation. Also, in this study, we concluded that it is evident and necessary to evaluate a possible reorientation of the policies directed by the government to address the crisis originated from the COVID-19; since, in this research, a level of disapproval close to 70% of the users was found. One of the strategies proposed is to reinforce social networks as a means of disseminating these types of decisions that affect the entire country; the level of compliance will depend on their level of acceptance.

Palabras claves: Sentiment analysis, trending topics, Twitter, algorithms, *machine learning.*

1. Justificación y planteamiento del problema

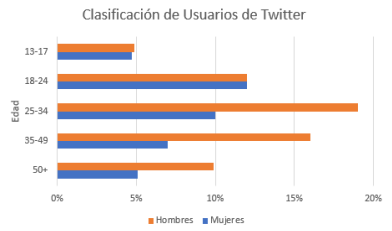
En los últimos años los avances tecnológicos han generado una sociedad dinámica que ha acelerado la forma en la que se percibe el mundo, presionando a los mercados a ser cada vez más flexibles y adaptativos ante los cambios frecuentes en las necesidades de la sociedad moderna. Estos avances han obligado a las organizaciones a hacerse más ágiles con el fin de crear estrategias competitivas que se adelanten a los patrones de comportamiento y perfilen los nichos de mercado al que están dirigiendo sus iniciativas. Estas nuevas dinámicas del mercado se han establecido con base en la creciente generación de datos desde diferentes fuentes, tales como, redes sociales, dispositivos móviles, sensores y, en términos generales todo lo que se puede clasificar dentro de las definiciones del Internet de las Cosas (Tabares & Hernández, 2014).

Una de las fuentes de información que más ha contribuido en la construcción de contenido para la sociedad moderna basada en el conocimiento, son los Servicios de Redes Sociales (SNS). Se reconoce que cualquier tipo de datos registrados por los usuarios en los SNS proporciona información crucial y el análisis de sus publicaciones conduce a la identificación de opciones y actitudes con respecto a la variedad de eventos. Entre el gran número de plataformas digitales disponibles, liderada por Facebook en número de usuarios y seguida de algunas como YouTube, Instagram y Twitter, solo se considerarán aquellas plataformas de medios sociales, donde los usuarios digitales puedan expresar sentimientos, pensamientos, criticas, puntos de vista, deseos, expectativas, entre otros. En este sentido las plataformas más populares que aplican para el estudio de las dinámicas de comportamiento son Facebook y Twitter.

Facebook es una red social que genera información en formato de texto, imágenes y videos. El tipo de datos que es posible obtener son, publicaciones, expresiones de “me gusta” y comentarios de tales publicaciones. A pesar de que es una fuente de datos popular y en la que los usuarios comparten opiniones, tiene restricciones tales como, acceso restringido a la información de los usuarios por políticas de privacidad y la posibilidad de estudiar únicamente las publicaciones de los perfiles pertenecientes a grupos o personas (Facebook Developers, 2020). Por otro lado, Twitter es una plataforma que permite el intercambio de textos cortos, máximo 140 caracteres, y algunos contenidos multimedia. La disponibilidad de la información es gratuita y se obtiene por medio de su Interfaz de Programación de Aplicaciones (API). La API de Twitter permite la administración de la información de cuentas, usuarios, tuits, mensajes directos y anuncios. Los permisos de acceso son supervisados por Twitter con el fin de garantizar siempre la protección de datos de los usuarios (Twitter Developers, 2020). A pesar de que no es la plataforma digital más popular, genera una gran cantidad de información que se puede utilizar eficientemente para la minería de opinión, tareas de análisis de sentimientos y perfilamiento de usuarios, a diferencia de Facebook que por su acceso restringido a la información no permite describir el comportamiento digital de los usuarios analizados.

Se reconoce que la información extraída de Twitter posee un alto nivel de confiabilidad debido a que el tipo de usuarios varía desde personas regulares hasta celebridades, personas de todas las edades como se muestra en la ilustración 1 y personas con numerosas opiniones relacionadas con diferentes temas de interés común, como, política, impacto de marca, salud, entre otros (Pack & Paroubek, 2010). Finalmente, según las características identificadas de las plataformas digitales Facebook y Twitter, se concluye que la fuente de datos que más se ajusta al estudio de patrones de comportamiento y perfilamiento de usuarios en la sociedad moderna es Twitter, por sus ventajas en términos del dinamismo de su contenido, accesibilidad y flexibilidad.

Ilustración 1. Clasificación de usuarios de Twitter por Edad y Género



Fuente: Elaboración Propia

Otro de los aspectos importantes que se deben considerar en la identificación de patrones de comportamiento y perfilamiento de usuarios, son las técnicas de procesamiento y análisis avanzado de la información obtenida.

Para el procesamiento de los datos extraídos de Twitter se hace necesaria la aplicación de algoritmos de Procesamiento del Lenguaje Natural (PLN) que alcancen masivamente el significado residente en el lenguaje humano. Por otro lado, para el análisis avanzado de los datos previamente procesados, se aplican técnicas de *Machine Learning*, una tecnología habilitante y automatizada que se encarga de extraer las nuevas variables y relacionar los datos con el objetivo de atribuir un verdadero valor agregado a la información de tal forma que facilite la toma de decisiones. En conjunto con la disponibilidad de información en la red social Twitter, los algoritmos de Procesamiento del Lenguaje Natural y las técnicas de *Machine Learning*, se podrían asegurar soluciones eficientes al estudio de las dinámicas de la sociedad moderna (Rivadeneira, 2018).

El estudio de las dinámicas de la sociedad moderna, en específico el análisis de sentimientos de usuarios de redes sociales, tiene consecuencias sociales y políticas. Las secciones de opinión de estas plataformas digitales tienen mucha información de políticas específicas, sobre temas de inmigración, transporte público, salud, entre otros. El análisis de las emociones durante situaciones de crisis es un asunto complicado, pues los eventos críticos se caracterizan por la experiencia de traspasar un umbral, ya sea individual o socialmente, a un estado nuevo pero desconocido, lo que produce incertidumbre y reacciones emocionales mixtas (Ruz, Henríquez & Mascareño, 2020). Durante el núcleo de la crisis, los estudios han argumentado que, al analizar cualitativamente la comunicación de Twitter, podemos obtener información relevante sobre las percepciones de las personas, la naturaleza del evento crítico y su visibilidad. Esta información es útil para mejorar la tecnología de respuesta a desastres y/o momentos de crisis (Ruz, Henríquez & Mascareño, 2020).

Existen herramientas de analítica de datos, como la creada por la Alianza CAOBA, con el fin de segmentar y perfilar a usuarios de Twitter conociendo sus características, pero hasta el momento no se ha creado una herramienta flexible que sea vigente en el tiempo y se enfoque en analizar los sentimientos de usuarios de Twitter sobre temas de interés (Vargas, Pomares, Alvarado, Quintero & Palacio, 2017). De acuerdo con las oportunidades identificadas en el campo de la Analítica de Datos para determinar patrones de comportamiento y perfilamiento en la sociedad, se plantea la pregunta de investigación, ¿Es posible la creación de un modelo de análisis de sentimiento sobre las opiniones expresadas por usuarios de Twitter en Colombia, sobre tópicos de interés actuales, y que permita perfilar dichos usuarios en términos de su edad, género y ocupación?

2. Antecedentes

Comprender la dimensión cualitativa de la comunicación de Twitter es particularmente útil en situaciones críticas, no solo con fines científicos o metodológicos. Los eventos críticos importantes, como los desastres naturales o las transiciones políticas son ciertamente inquietantes para las personas, las comunidades o las grandes regiones, especialmente durante el centro del evento (Ruz, Henríquez & Mascareño, 2020). Es por esto por lo que las personas tienden a interesarse por comunicar acerca de un mismo tópico al tiempo. La comunidad de usuarios que participan en las redes sociales tiende a compartir sobre intereses comunes al mismo tiempo, dando lugar a lo que se conoce como tendencias sociales. Una tendencia social refleja la voz de un gran número de usuarios que, por alguna razón, se vuelve popular en un momento específico. Por lo tanto, a través de las tendencias sociales, los usuarios sugieren que se está produciendo algún evento de gran interés (Zubiaga, Spina, Martinez & Fresno, 2015). El análisis sentimental es el proceso de derivar la información de calidad del texto. En otras palabras, es el proceso de derivar los datos estructurados de los datos no estructurados. Esto se utiliza para medir opiniones del cliente, comentarios, revisiones de productos (Siddharth, Darsini & Sujithra, 2018).

Así mismo, en los últimos años las redes sociales se han convertido en una plataforma prometedora para el intercambio de opiniones, por lo que es una fuente ideal para obtener opiniones, críticas y sugerencias de usuarios sobre productos, personas y eventos. Los sitios web de *micro blogs* son una de las fuentes más importantes de información variada. Esto se debe al hecho de que todas las personas publican sus opiniones sobre una variedad de temas, discuten problemas actuales, se quejan y expresan un sentimiento por los productos que usan en la vida diaria (Siddharth, Darsini, & Sujithra, 2018). La plataforma de Twitter puede incluso influir indirectamente en la configuración tradicional de la agenda de los medios de comunicación, especialmente en eventos críticos, ya que los periodistas recopilan información de tuits y retuitean valiosos mensajes compartidos por los usuarios (Ruz, Henríquez & Mascareño, 2020).

El uso de redes sociales para la identificación de temas se ha vuelto esencial cuando se trata de la detección de eventos, especialmente cuando los eventos impactan a la sociedad (Suri & Roy, 2017). Estos autores, evaluaron

dos modelos de detección de tópicos, *Latent Dirichlet Allocation* (LDA) y *Non-Negative Matrix Factorization* (NMF), obteniendo que ambos algoritmos funcionan bien en la detección de temas de secuencias de texto, siendo los resultados de LDA más semánticamente interpretables. De la misma manera, Ostrowski en el 2015 explora el modelado de temas considerando las técnicas de Asignación de *Dirichlet* Latente, evaluando la técnica desde la perspectiva de la clasificación, así como la identificación de temas notables, ya que se aplica a una colección filtrada de mensajes de Twitter, observando que esta técnica funcionó bien como modelo sin supervisión, demostró potencial como una técnica complementaria que podría servir como un medio para apoyar la derivación de información sobre tendencias (Ostrowski, 2015).

En el año 2018, se realizó un estudio con el objetivo de explorar cómo se pueden utilizar las técnicas de análisis de texto para profundizar en algunos de los datos en una serie de publicaciones que se centran en las diferentes tendencias de los tuits (Siddharth, Darsini, & Sujithra, 2018). Los autores proponen el uso de dos modelos para el análisis sentimental, el clasificador bayesiano ingenuo y las redes neuronales. Los resultados mostraron que los clasificadores de aprendizaje automático propuestos son eficientes, sin embargo, los modelos de análisis de sentimientos existentes se pueden mejorar aún más con un mayor conocimiento semántico y de sentido común.

En el 2020, Ruz, Henríquez y Mascareño, hacen una comparación de 5 diferentes modelos para el análisis de sentimientos (uno es una variante del modelo TAN) y evaluaron su desempeño con dos conjuntos de datos de Twitter de dos eventos críticos diferentes, llegando a la conclusión de que el uso de Twitter tiene muchas potencialidades y aplicaciones antes, durante y después de eventos críticos. La precisión de la clasificación del contenido semántico de los tuits es muy importante para reducir los riesgos de desinformación en esas situaciones (Ruz, Henríquez & Mascareño, 2020). Los resultados mostraron, con respecto a la precisión de los modelos que, el SVM fue el único clasificador que obtuvo más del 80% de precisión en ambos conjuntos de datos, el clasificador de árbol bayesiano (TAN) obtuvo resultados competitivos cuando había datos suficientes para respaldar la estructura de árbol, también se concluyó que TAN y BF TAN ofrecen información cualitativa interesante para comprender histórica y socialmente las características principales de la dinámica del evento.

El desarrollo de la tecnología y la información ha hecho que las redes sociales se conviertan en la herramienta de comunicación más popular, según el Ministerio de las TIC, es una de las redes más usadas por los usuarios en Colombia, con alrededor de 6 millones de personas vinculadas (Ministerio de Tecnologías de la Información y las Comunicaciones, 2018)., por lo que las empresas se ven bastante atraídas por estas, en especial Twitter, ya que es una red social ideal para obtener opiniones de lo que ocurre en tiempo real. En 2020, Syahputra, Basyar y Tamba hicieron un estudio relacionado con la opinión pública sobre los servicios prestados por Go-Jek Indonesia. Los autores hacen uso del algoritmo SVM, una técnica para hacer predicciones, tanto en el caso de clasificación como de regresión. Los resultados muestran que el algoritmo *Support Vector Machine* puede usarse para clasificar sentimientos de texto para 3 clases usando el método *Multiclass One Vs Rest*.

Caracterizar a las personas es importante para conocer y analizar el porqué de su postura y tomar mejores decisiones. En este sentido, cuanto mejor se conozca a la sociedad, más posibilidad hay de mantenerla satisfecha. La capacidad de clasificar los atributos de los usuarios latentes, incluidos el género, la edad, el origen regional y la orientación política únicamente del lenguaje de usuario de Twitter o contenido similar altamente informal, tiene aplicaciones importantes en publicidad, personalización y recomendación (Rao, Yarowsky, Shreevats & Gupta, 2014)., Rao, Yarowsky, Shreevats & Gupta, hacen una investigación de tres algoritmos de clasificación basados en SVM, teniendo como resultado que la mejor combinación es un modelo de agrupación y la técnica de SVM, ya que se obtuvieron los siguientes valores de precisión, Género 72.33%, Edad 74.11% y Orientación Política 80.19%. Por otro lado, En el 2016 Al-Ghadir y Azmi, realizan un estudio con el fin de predecir la variable de género para la colección de usuarios obtenida de Twitter haciendo uso de dos técnicas de clasificación, *Support Vector Machine* (SVM) y vecino más cercano (1-NN) y experimentaron con diferentes tamaños para la lista de características. Los resultados obtenidos mostraron que la técnica 1-NN tiene un mejor rendimiento (93.16%), comparado con la técnica SVM (87.33%), cuando se están analizando pocos datos. Mientras que en el caso en el que sea un gran volumen de información el SVM ofrece un mejor rendimiento.

En el *anexo 1* se presenta un resumen de los principales artículos – revisados dentro de esta investigación – relacionados con el tema de temas tendencia, análisis de sentimientos y perfilamiento. En cada uno de estos, se identifica temas tendencia en redes sociales o se hace un análisis de sentimientos en datos de Twitter lo que nos permite analizar el uso de las distintas metodologías usadas.

Se propone para este proyecto hacer el análisis de sentimientos de los temas de interés actuales de las personas en Colombia, tomando como base los usuarios de Twitter en el país, y las opiniones que ellos expresan sobre dichos temas. Como herramientas analíticas se usarán el algoritmo *Latent Dirichlet Allocation* (LDA) para la detección de tópicos tendencia, un algoritmo apropiado para análisis de sentimientos de los usuarios con respecto a los temas identificados; por último, se realizará el perfilamiento de estos usuarios según su postura, teniendo en cuenta las variables demográficas de edad, género y ocupación, haciendo uso del algoritmo *Support Vector Machine* (SVM) ya que es el que arroja los mejores resultados para estas variables según las referencias de Rao, Yarowsky, Shreevats y Gupta (2014) y Al-Ghadir y Azmi, (2016).

3. Objetivos

Construir un modelo de análisis de sentimientos sobre temas de interés actuales, de las personas en Colombia, usando datos de la red social Twitter

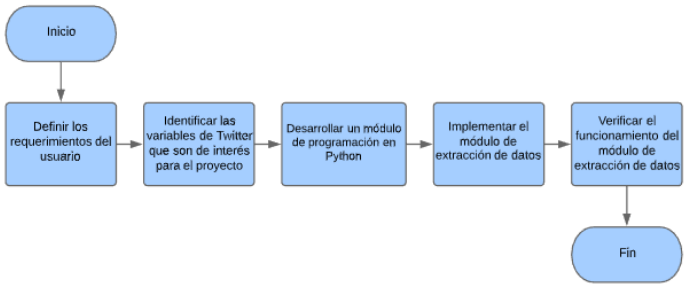
- Diseñar un módulo para la obtención de datos de Twitter
- Desarrollar un modelo para la detección de tópicos de interés y análisis de sentimientos sobre dichos tópicos, tomando como base las opiniones de los usuarios de Twitter en Colombia
- Desarrollar un modelo para el perfilamiento de los usuarios de Twitter según los sentimientos de sus opiniones sobre tópicos de interés.
- Analizar el sentimiento de las opiniones de los usuarios de Twitter en Colombia sobre temas relacionados con la pandemia del COVID-19, durante un periodo de tiempo específico.

4. Cuerpo del documento

4.1. Diseñar un módulo para la obtención de datos de Twitter

El componente de extracción de datos representa la primera etapa fundamental para el análisis de sentimientos sobre temas de interés y perfilamiento de los usuarios en Colombia. El diseño de este módulo permite capturar los datos necesarios para el desarrollo de los demás objetivos. La metodología seguida para el diseño del módulo es la siguiente:

Ilustración 2. Diagrama de flujo de la metodología para el diseño del módulo para la obtención de datos de Twitter



Fuente: Elaboración propia

- Definir los requerimientos del usuario: Antes de la programación del módulo es necesario reconocer el tipo de usuarios al que va dirigido, que para este caso son personas y entidades públicas o privadas que estén interesadas en estudiar la dinámica de las redes sociales, desde el tipo de personas que la usan, temas en tendencia hasta opiniones y/o sentimientos.

- Identificar las variables de Twitter que son de interés para el proyecto: Según la información extraída a través de la API de Twitter (Twitter Developer, 2020), tanto a nivel de tuit como de usuario y teniendo en cuenta el objetivo del proyecto, se analizaron los campos disponibles en la estructura de datos de un tuit definiendo los siguientes campos como variables de interés:

Tabla 1. Variables de interés para el proyecto

Variable	Tipo de dato	Descripción
created_at	String	Fecha y hora de creación del tuit
text	String	Texto del tuit
user	JSON	Datos del usuario que realizó el tuit
place	JSON	Datos del lugar donde se realizó el tuit

Fuente: Elaboración propia

- Desarrollar un módulo de programación en Python: Una vez se identificaron las variables para el desarrollo de los objetivos, se procedió a desarrollar el módulo en Python que extrae los tuits a través del servidor de información de Twitter que responde a la solicitud de la API y los almacena en una base de datos en tiempo real.

    1. Acceso a la API Pública de Twitter: Para acceder al servidor de información de Twitter, es necesario solicitar permisos en el desarrollador de la red social para la creación de una aplicación que permita el acceso a la información. Al crear esta aplicación, Twitter genera unas credenciales de acceso aleatorias para acceder a los servicios de la API creada. (Twitter Developer, 2020)

    2. Acceso a los datos de Twitter

    2.1. Realizar la autenticación de las credenciales: Para poder conectarse con la API de Twitter se utiliza una interfaz que permita el envió continuo de las credenciales a la aplicación web, para este caso se seleccionó *OAuth.*

    2.2. Extraer los datos de Twitter: El desarrollador de Twitter ofrece dos tipos de API Públicas, *API Rest* y *API Streaming* (Twitter Developer, 2020), que permiten extraer información en diferentes momentos del tiempo. En este caso, se seleccionó el tipo de flujo de datos de la *API Streaming* ya que permite obtener los tuits en tiempo real. Para ejecutar el módulo de extracción de datos desde *API Streaming*, la aplicación solicita usar alguno de los tres filtros disponibles, que son palabras claves, ubicación del tuit e identificación de usuarios. Teniendo en cuenta la solicitud del filtro y que los datos requeridos son para Colombia se decide depurar la información por ubicación del tuit, especificando las coordenadas rectangulares del país: (-81.728111, -4.2304, -66.869827, 13.39029).

    3. Almacenamiento de datos: Teniendo en cuenta que los tuits recuperados del módulo tienen una estructura en formato JSON se reconoce la necesidad de crear una base de datos no relacional que acepte esta estructura. De acuerdo con esto, se selecciona MongoDB para almacenar los tuits. El proceso de almacenar los tuits se ejecuta en el mismo módulo desde Python por medio de una interfaz de MongoDB utilizando la librería "*MongoClient*", con el fin de almacenar los datos en tiempo real y de forma directa a la base de datos.

- Implementar el módulo de extracción de datos: Para la implementación del módulo se determinó que el tiempo de ejecución serían 10 días, en los cuales se debía asegurar la conexión continua a internet y unos requerimientos computacionales que permitieran el almacenamiento en tiempo real de la información. Con este fin se seleccionó un equipo que poseía las siguientes características: RAM = 8GM, VCPU= 4 y Disco = 80GB. Una vez se aseguraron los recursos que no afectaran el funcionamiento del módulo se inició su ejecución el día 3 de mayo de 2020 y terminó el 13 de mayo de 2020. Este periodo de ejecución se determinó con base en la situación que vivía el país en ese momento por el COVID-19 y en la cantidad de datos que se podían capturar.

- Verificar el funcionamiento del módulo de extracción de datos: Una vez se terminó la fase de implementación se verificó que todos los componentes del módulo funcionaron correctamente y cumplieran con los requisitos, teniendo como resultado final la base de datos con los tuits de 10 días de usuarios en Colombia y en algunos países limítrofes.

Como resultado final del ***Módulo para la obtención de datos de Twitter*** se obtuvo un total de 560,223 tuits de usuarios en Colombia y países limítrofes. Como el caso de estudio de este proyecto es únicamente en Colombia,

se redujo la base de datos inicial excluyendo aquellos usuarios que estuvieran fuera del país, obteniendo en total 189,492 tuits. En la ilustración 3(a), se muestra la estructura que tiene cada uno de los tuits extraídos y almacenados en MongoDB:

Ilustración 3. Estructura de los tuits en MongoDB

(a) (b)

Fuente: MongoDB

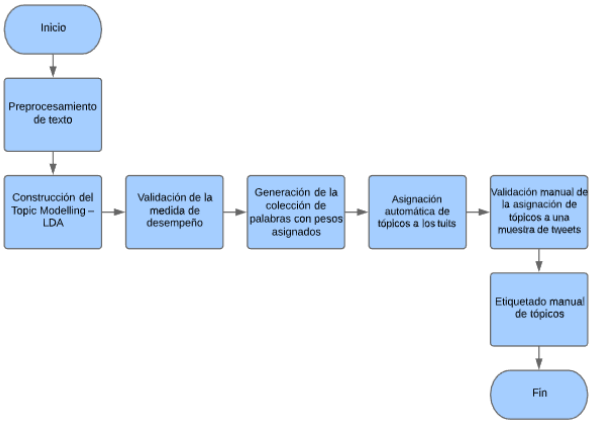
De acuerdo con la estructura del tuit en la ilustración 3(a), los campos “user”, “place” y “entities” son expandibles debido a su formato tipo JSON, en la ilustración 3(b), se muestra un ejemplo para la estructura de este tipo de datos.

4.2. Desarrollar un modelo para la detección de tópicos de interés y análisis de sentimientos sobre dichos tópicos, tomando como base las opiniones de los usuarios de Twitter en Colombia.

4.2.1. Modelo para la detección de tópicos de interés en Colombia

Con este modelo desarrollado en Python se pretende detectar los tópicos de interés en Colombia durante un periodo de tiempo específico, teniendo como datos de entrada los tuits de los usuarios extraídos en el módulo para la obtención de datos de Twitter diseñado en el primer objetivo. A continuación, se presenta la metodología seguida para el diseño de este modelo:

Ilustración 4. Diagrama de flujo de la metodología para el diseño del modelo para la detección de tópicos de interés en Colombia



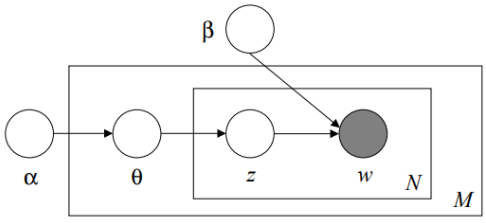
Fuente: Elaboración propia

- Preprocesamiento de texto: Se realizó el preprocesamiento de texto a los tuits, con el fin de convertir una cadena de texto larga en un formato intermedio que le permita ser tratada computacionalmente en el procesamiento de lenguaje natural. El preprocesamiento se realizó con base en la estructura del tuit (No estructurado) y sus características especiales. La metodología para el preprocesamiento fue la siguiente:

    1. Eliminación de ruido: Es un proceso de normalización de texto para las características especiales que puede presentar un tuit:
        1.1. Eliminación de usuarios y *retweets*: Eliminación de los usuarios que se encuentran *arrobados* en el tuit, además de los "RT" que aparece cuando el texto es un *retweet.*
        1.2. Eliminación de *links*: Eliminación de los enlaces provenientes de internet.
        1.3. Eliminación de signos de puntuación: Eliminación de los signos de puntuación innecesarios y/o que no aportan nada al significado del texto.
    2. *Tokenización*: Permite delimitar las palabras de un tuit para reconocer en los siguientes pasos del preprocesamiento si son significativas para el modelo que se está desarrollando.
    3. Eliminación de *stopwords*: Después del proceso de *tokenización* se pueden eliminar palabras que no tienen una semántica específica, y que no contribuyen al significado del texto, por ejemplo, "el", "la", "es", etc.
    4. *Stemming*: Este paso reduce las palabras a su raíz o base, estas raíces son la parte invariable de las palabras, por ejemplo, las palabras: canta, cantamos, cantáis y cantan, se convierten en 'cant-'

- Construcción del *Topic Modelling – LDA*: En esta fase los tuits de los usuarios de Colombia se pasaron por el algoritmo *Latent Dirichlet Allocation (LDA)*, el cual funciona bajo la suposición de que la forma en que se genera un documento, o en este caso un tuit, es eligiendo un conjunto de temas y una serie de palabras para cada uno de ellos. En este sentido, el LDA identifica los temas por medio de un conjunto de palabras que lo describen (Honing Data Science, 2020). El LDA es un modelo bayesiano jerárquico de tres niveles, dentro del cual cada elemento de un corpus se modela como una mezcla finita sobre un conjunto subyacente de temas. Luego, cada tema se modela como una mezcla infinita sobre un conjunto subyacente de probabilidades para cada tema. El modelo representa los documentos, que para este caso son tuits, como una mezcla de temas que contienen palabras con una probabilidad de ocurrencia, de tal forma que, a partir de una colección de documentos y un número fijo de temas, el modelo probabilístico aprende la representación de un tema en cada documento y sus palabras asociadas por medio de un procedimiento iterativo (Suri & Roy, 2017)., estas réplicas se muestran en la representación gráfica del modelo, donde la placa exterior son los documentos y la interior es la elección repetitiva de palabras y temas dentro de un documento, en cada una de estas iteraciones se tienen en cuenta los siguientes parámetros:

Ilustración 5. Representación gráfica del modelo LDA



Fuente: Blei, Ng & Jordan, 2003

La placa exterior (M): Representa los documentos.
La placa interior (N): Representa la elección repetida de temas y palabras dentro de un documento.
$\propto$: Es la distribución de los temas por documento.
θ: Es la distribución de temas para el documento m.
w: Palabra especifica n.
z: Es el tema de la palabra n en el documento m.
β: Es la distribución de las palabras por tema.

El procedimiento para el LDA fue el siguiente:

1. Crear modelos *bigram*: Los modelos *bigram* se crean a partir del texto preprocesado con el fin de crear agrupaciones de dos palabras que aparecen con frecuencia juntas en el texto y tienen un significado.
2. *Bag of words:* Este modelo se crea con el objetivo de representar numéricamente los tuits en función de las palabras *tokenizadas* por cada una de ellas obteniendo así una "bolsa de palabras". El modelo crea una única identificación para cada palabra y el corpus resultante es un mapeo del ID que le asigna el modelo a una palabra y la frecuencia de esta en toda la colección de tuits, por ejemplo, (0,2) significa que la palabra con ID igual a 0 esta 2 veces en toda la "bolsa de palabras"

3. Construir el modelo: Una vez que se tuvo todo lo necesario para ejecutar el modelo se necesita determinar la cantidad de tópicos que el modelo debe identificar en la colección de tuits, teniendo en cuenta una medida de desempeño que determine si se está escogiendo una agrupación con tópicos semánticamente coherentes.

- Validación de la medida de desempeño

1. Medida de Desempeño – *Coherence Value*: En esta fase se usó la medida de desempeño conocida como *Coherence Value*, la cual determina el nivel coherencia midiendo el grado de similitud semántica entre el conjunto de palabras que más puntúan dentro de cada tema. (Kapadia, 2019). El *Coherence Value* tiene diferentes formas de medición, pero la más adecuada por la estructura de los datos es la *umass*, esta medida de coherencia se basa en el recuento de coocurrencia entre los documentos del corpus original. (Keith, Philip, Andrzejewski & Buttler, 2012).

2. Validación Medida de Desempeño: Para validar el número óptimo de tópicos se construyen varios modelos LDA con diferente número de tópicos y se toma aquel que devuelva el valor de coherencia más cercano a 0 debido a la forma de medición del *umass* y el que según la interpretación humana de los tópicos tenga más coherencia. (Hammoe, 2018).

- Colección de palabras con pesos asignados: Una vez que se aplica el modelado de tópicos LDA al conjunto de tuits, se puede ver las palabras que componen cada tema junto con su contribución porcentual al tema.

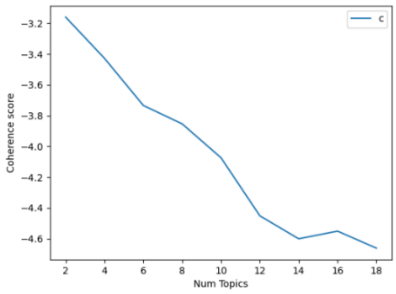
- Asignación automática de tópicos a los tuits: Ya que se tiene el conjunto de temas de la base de datos y las palabras que lo conforman, el modelo asigna automáticamente alguno de los tópicos identificados a cada uno de los tuits, esto con el fin de crear una base de datos con los tuits correspondientes a cada tópico.

- Validación manual de la asignación de tópicos a una muestra de tuits: Para validar la confiabilidad de la información en una población de 189,492 tuits, se calculó una muestra con un nivel de confianza de 95%, un margen de error de 5% y una probabilidad de éxito y fracaso del 50% a partir de la fórmula de muestreo para poblaciones finitas (Torres & Paz, 2015), obteniendo como resultado que se debían verificar mínimo 384 tuits de muestra. A partir de esto, se tomó una muestra mayor para tener un mejor alcance teniendo en cuenta que son varios tópicos los que se están clasificando, se verificaron 600 tuits con el fin de validar que la asignación de tópicos si tienen relación y están correctamente asignados, como resultado se obtuvo que el 64,5% de los tuits si tienen relación con el tópico asignado, lo que se considera una medida aceptable para el modelo.

- Etiquetado manual de tópicos: Debido a que el LDA arroja un conjunto de palabras relacionadas con el tema como se muestra en la tabla 2, es necesario hacer un etiquetado manual del tópico a cada conjunto de palabras.

Como resultado final el ***Modelo para la detección de tópicos de interés en Colombia*** generó 4 temas con un conjunto de 10 palabras para cada uno de los temas como se muestra en la tabla 2 y un *Coherence Value* de -3,58. La selección del número de tópicos se realiza siguiendo la metodología de validación de la medida de desempeño ya mencionada, para lo cual se graficó el comportamiento del valor de coherencia con respecto a la variación en la cantidad de tópicos.

Ilustración 6. Gráfica Coherence Value para LDA



Fuente: Elaboración propia

Finalmente se clasifica de forma manual cada conjunto de palabras y se determina que la colección de tuits extraída con el módulo de Twitter trata los siguientes temas:

Tabla 2. Conjunto de palabras LDA

Tópico 1: Contenido multimedia en cuarentena	0,013 “quier”	0,011 “acab”	0,011 “si”	0,010 “fot”	0,009 “gan”
	0,009 “hace”	0,007 “vide”	0,007 “jajaj”	0,007 “da”	0,006 “com”
Tópico 2: Día de la madre y amor	0,041 “día”	0,020 “madr”	0,019 “feliz”	0,018 “graci”	0,017 “hoy”
	0,016 “dios”	0,015 “vid”	0,015 “mam”	0,014 “tod”	0,014 “amor”
Tópico 3: Actividades en cuarentena	0,021 “si”	0,018 “hac”	0,012 “pued”	0,011 “pas”	0,011 “sol”
	0,010 “asi”	0,009 “quier”	0,009 “ser”	0,00 “tan”	0,008 “sal”
Tópico 4: Manejo del COVID-19 por el gobierno	0,010 “cov”	0,009 “colombi”	0,006 “cas”	0,006 “pag”	0,006 “may”
	0,005 “trabaj”	0,005 “nuev”	0,005 “gobiern”	0,005 “deb”	0,005 “país”

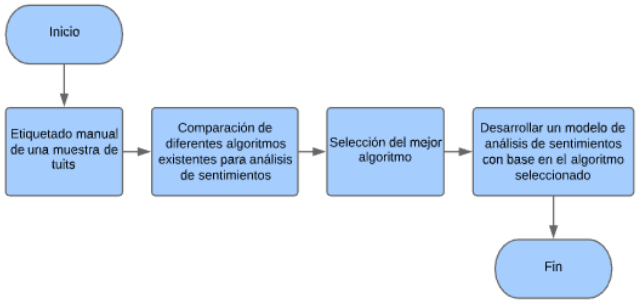
Fuente: Python

A partir de los resultados del modelo se seleccionó un tópico de interés para el desarrollo y análisis de los demás objetivos, en este caso fue el tópico 4, el cual tiene relación con el manejo del COVID-19 por el gobierno.

4.2.2. Modelo para el análisis de sentimientos sobre tópicos de interés en Colombia

El análisis de sentimientos es un campo de investigación dentro del procesamiento de lenguaje natural que trata de extraer de manera automática y mediante técnicas computacionales, información subjetiva expresada en el texto de un documento dado y acerca de un determinado tema (Sobrino, 2018). En este apartado, se pretende seleccionar y mostrar el proceso de entrenamiento de un algoritmo de aprendizaje automático en Python, para realizar análisis de sentimientos sobre el tema del manejo del COVID-19 por el gobierno, con una definición de tres clases de sentimiento, “Positivo”, “Negativo” y “Neutro”. A continuación, se presenta la metodología trabajada para la selección y diseño de este modelo:

Ilustración 7. Diagrama de flujo de la metodología para el diseño del modelo para el análisis de sentimientos



Fuente: Elaboración propia

- Etiquetado manual de una muestra de tuits: Las técnicas para el análisis de sentimientos se basan en aprendizaje automático supervisado. Estas técnicas necesitan datos previamente etiquetados con el fin de entrenar un modelo que será usado para clasificar nuevos textos. Para obtener estos datos de entrenamiento, se etiquetaron manualmente una muestra de 500 tuits en “positivos”, “negativos” y “neutros” con los cuales se entrenó el modelo y según los resultados obtenidos se decidió agregar otra muestra de 500 tuits etiquetados. Luego del etiquetado y una limpieza de los tuits que no aportan suficiente información se entrenó de nuevo el modelo. Con estos nuevos datos se obtuvieron mejores resultados, por lo cual se decide que este será el set de entrenamiento, con el que el modelo “aprenderá”.
- Comparación de diferentes algoritmos existentes para análisis de sentimientos: Con base en los modelos más usados en las referencias investigadas, se decide comparar los resultados que arrojan estos modelos. Se compararán los modelos de *Support Vector Machine (SVM), Random Forest (RF), Tree Decision (TD), Logistic Regression (LR)* y *Naive Bayes (NB)* con el fin de escoger el que mejor se adaptara a una base de datos de información extraída de Twitter.

Para la comparación de los modelos se usan diferentes bases de datos de entrenamiento además de los ya etiquetados manualmente, con el fin de comparar los resultados en diferentes escenarios. Las tres bases de datos

con las que se compararan cada uno de los modelos son: base de datos propia (etiquetado manualmente), base de datos tomada de GitHub y base de datos combinada (etiquetado manual y GitHub). La base de datos propia hace referencia a una base de datos etiquetada manualmente con un total de 718 tuits, la base de datos tomada de internet hace referencia a una base de datos de ejemplo tomada de una publicación de GitHub con un total de 3954 tuits, y la base de datos combinada hace referencia a la combinación de las dos bases de datos anteriores con un total de 4672 tuits.

Para cada una de estas bases de datos de entrenamiento se evaluaron dos medidas de desempeño, la exactitud (*accuracy*) y la exhaustividad (*recall*). De acuerdo con la matriz de confusión:

Tabla 3. Matriz de confusión

		Predicción	
		Cp	Cn
Clase Real	Cp	TP:True positive	FN: False negative
	Cn	FP: False positive	TN: True negative

Fuente: Elaboración propia

La exactitud se calcula como el número de todas las predicciones verdaderas dividido en el número total del conjunto de datos, es decir, este mide la proporción de tuits que predice el modelo correctamente sobre todos los tuits a los que les asignó una predicción.

$$Accuracy = {}^{TP+TN}/_{TP+TN+FP+FN} \qquad \textit{Ecuación 1 (Kuhn, 2019)}$$

La exhaustividad, también conocido como *sensitivity*, se calcula como el número de predicciones positivas verdaderas – para cada categoría – dividido en el número total de positivos, es decir, este mide la proporción de tuits que el modelo predice correctamente sobre todos los tuits a los que les asignó una predicción, para cada categoría.

$$Recall\ (Sensitivity) = {}^{TP}/_{TP+FN} \qquad \textit{Ecuación 2 (Kuhn, 2019)}$$

Teniendo estas definiciones en cuenta, de aquí en adelante se hará referencia solamente a exactitud y exhaustividad

A continuación, los resultados para la comparación de los diferentes algoritmos:

Tabla 4. Primera comparación de algoritmos para el análisis de sentimientos

Modelo	**Origen base de datos**	**Exhaustividad**			**Exactitud**
		Negative	**Neutral**	**Positive**	
SVM	Propia	0,67	0,69	0,13	0,51
	Internet	0,96	0,50	0,65	0,85
	Combinada	0,93	0,12	0,55	0,76
Random Forest	Propia	0,19	0,96	0,10	0,53
	Internet	0,95	0,19	0,40	0,73
	Combinada	0,99	0,11	0,35	0,74
Tree Decision	Propia	0,38	0,78	0,4	0,57
	Internet	0,85	0,38	0,39	0,67
	Combinada	0,86	0,37	0,45	0,71
Logistic Regression	Propia	0,38	0,87	0,00	0,53
	Internet	0,98	0,06	0,39	0,73
	Combinada	0.99	0,00	0,41	0,76
Naive Bayes	Propia	0,50	0,74	0,00	0,51
	Internet	0,99	0,06	0,13	0,67
	Combinada	1,00	0,05	0,16	0,69

Fuente: Elaboración propia

Teniendo en cuenta los resultados obtenidos se descartaron los dos modelos con el promedio de exactitud para las tres bases de datos más bajo, TD (0,65) y NB (0,62) y se continuó haciendo pruebas para los otros modelos. Esta vez, se tuvo en cuenta el desbalance que había entre las clases de cada una de las bases de datos y se realizó de forma manual la eliminación de tuits para las clases con más tuits etiquetados. De acuerdo con este ajuste se obtuvieron los siguientes resultados:

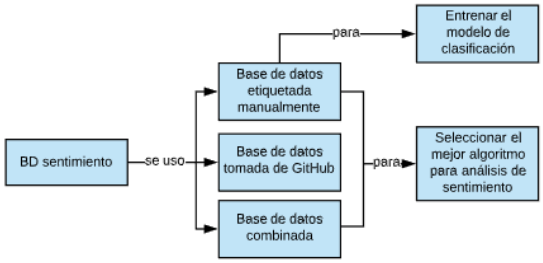
Tabla 5. Segunda comparación de algoritmos para el análisis de sentimientos

Modelo	**Origen base de datos**	**Exhaustividad**			**Exactitud**
		Negative	**Neutral**	**Positive**	
SVM	Propia	0,76	0,86	0,30	0,71
	Internet	0,73	0,95	0,79	0,80
	Combinada	0,86	0,83	0,70	0,79
Random Forest	Propia	0,41	0,95	0,20	0,60
	Internet	0,87	0.96	0,57	0,79
	Combinada	0,89	0,69	0,65	0,75
Logistic Regression	Propia	0,76	0,90	0,20	0,71
	Internet	0,73	0,98	0,75	0,81
	Combinada	0,86	0,91	0,70	0,81

Fuente: Elaboración propia

- Selección del mejor algoritmo: Con base en los resultados de las pruebas realizadas, se seleccionó el algoritmo que arrojó la mejor exactitud promedio para las tres bases de datos, es decir, *Logistic Regression* con un promedio de 0,77.
- Desarrollar un modelo de análisis de sentimientos con base en el modelo seleccionado: Para el desarrollo del modelo se define la base de datos propia como la más apropiada para el entrenamiento, entre las tres usadas para la selección del modelo, ya que esta cuenta con información a nivel de usuario. Esta base de datos se compone de un total de 718 tuits pertenecientes a 589 usuarios, con un promedio de alrededor 1 tuit por usuario con una desviación de 0,73. Además, cada tuit tiene en promedio alrededor de 17 palabras con una desviación de 5,20. A continuación, se muestra el diagrama de las bases de datos usadas:

Ilustración 8. Diagrama bases de datos modelo de Análisis de Sentimiento



Fuente: Elaboración propia

1. Preprocesamiento de datos

1.1. Convertir cadena de texto en minúsculas: Este paso convierte el texto del tuit en minúscula con el fin de poner todas las palabras en un mismo formato y que el proceso continúe de manera uniforme.
1.2. Eliminación de ruido: Es un proceso de normalización de texto para las características especiales que puede presentar un tuit:
  1.2.1 Eliminación de usuarios y *retweets*: Eliminación de los usuarios que se encuentran *arrobados* en el tuit, además de los "RT" que aparece cuando el texto es un *retweet*.
  1.2.2. Eliminación de *links*: Eliminación de los enlaces provenientes de internet.
  1.2.3. Eliminación de signos de puntuación: Eliminación de los signos de puntuación innecesarios y/o que no aportan nada al significado del texto.
1.3. *Tokenización*: Permite delimitar las palabras de un tuit para reconocer en los siguientes pasos del preprocesamiento si son significativas para el modelo que se está desarrollando.

1.4. Eliminación de *stopwords*: Después del proceso de *tokenización* se pueden eliminar palabras que no tienen una semántica específica, y que no contribuyen al significado del texto, por ejemplo, “el”, “la”, “es”, etc.
1.5. Lematización: Este paso relaciona una palabra derivada con su forma canónica o lema, por ejemplo, canto, cantamos, cantan, son relacionadas todas con la palabra cantar.

2. Vectorización: Las palabras no son cosas que las computadoras entienden naturalmente. Al codificarlas en forma numérica, podemos aplicar reglas matemáticas y hacerles operaciones matriciales (Heidenreich, 2018). Dicho esto, se realizó una vectorización a los tuits para transformar un corpus de texto en un vector de recuento de términos, para lo que se tuvo en cuenta la función *TF-IDF* la cual identifica las palabras que aparecen con más frecuencia en un tuit y con menos frecuencia en toda la colección de tuits para así darles más importancia ya que son más útiles para la clasificación (Bafna, Pramod & Vaidya, 2016).

3. Construcción del modelo de clasificación: La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica, en función de las variables independientes o predictoras. Es un caso particular de los modelos lineales generalizados que concierne al análisis de datos binarios, donde la función de enlace es la función *logit* o logística (Robles-Velasco, Cortés, Muñuzuri & Onieva, 2019).

   - Inicialmente se decide probar el modelo dividiendo la base de datos de entrenamiento en un set de prueba con el 30% del total de tuits y un set de entrenamiento con el 70% restante. Cuando ya se ha realizado la repartición aleatoria de los datos, se entrena el modelo y se evalúa su rendimiento con el set de prueba, basándose en las medidas de desempeño de exactitud y exhaustividad.
   - Teniendo en cuenta los resultados anteriores se reconoce que la cantidad de tuits etiquetados en la base de datos de entrenamiento para cada una de las clases del modelo generan un sesgo en la clasificación. Para evitar que este tipo de sesgo altere los resultados se decide implementar una herramienta en *machine learning* que consiste en balancear los datos de entrenamiento para todas las clases. Para este balance se tuvo en cuenta el método *oversampling*, el cual consiste en obtener más datos de las clases minoritarias de los que originalmente había, dejando intacta la cantidad de elementos de la clase mayoritaria (Núñez, Angulo & Abril, 2012).
   - Adicional a los cambios en el tamaño de los datos de entrenamiento se decide evaluar cambios en la parametrización del modelo. De acuerdo con los parámetros del modelo de regresión logística, el que se puede calibrar para obtener un mejor rendimiento es el “costo”, el cual controla el número y la severidad de las violaciones del margen que se tolera en el proceso de ajuste, en donde los valores más pequeños especifican una regularización más fuerte.
     En este sentido, primero se decide iterar sobre rangos bajos del costo (1,5) y se evalúan las medidas de desempeño sobre los datos de entrenamiento y de prueba, teniendo en cuenta que dependiendo de su valor ya sea bajo o alto se puede generar respectivamente un subajuste o un sobreajuste del modelo, para evitar estos posibles sesgos se decide iterar nuevamente sobre rangos de costo más altos (6 ,20). A partir del mejor resultado de exactitud que se reporte para cada rango, se toman estos dos valores del costo y se verifica la consistencia del modelo con validación cruzada (explicada en el numeral 6), tomando finalmente el valor del costo que presente un menor sesgo para la exactitud.

4. Evaluación de las medidas de desempeño: Para conocer si el modelo funciona correctamente, al igual que en la selección del algoritmo, se tomaron como medidas de desempeño la exactitud y la exhaustividad.

5. Validación cruzada: Como forma de comprobar la consistencia en el comportamiento de los resultados que arroja el modelo, se realizó una validación cruzada sobre el set de entrenamiento, que tiene como propósito ejecutar el modelo con diferentes divisiones de datos de entrenamiento y de prueba, generando así diferentes resultados, uno para cada división, lo que nos ayuda a comprobar que el modelo se está entrenando correctamente y que no hay sesgos.

Para la construcción del ***Modelo para el análisis de sentimientos sobre tópicos de interés en Colombia*** se obtuvieron diferentes resultados debido a los ajustes que se realizaron con el fin de obtener los mejores, el primer resultado que se obtuvo fue para un modelo inicialmente planteado en el que no se tuvieron en cuenta los parámetros de la Regresión Logística, ni el desbalance en el que se encontraban los datos de entrenamiento para cada una de las clases, en este caso los resultados quedaban sesgados hacia las clases mayoritarias como

se evidencia en la tabla 6 que para la clase "Neutro" reportaba el valor de exhaustividad más alto y la exactitud para todo el modelo resultante fue de 0,55.

Al reportar los primeros resultados se decide hacer un primer ajuste en el balance de datos de entrenamiento tomando como referencia la categoría con más datos de entrenamiento, en este caso la de tuits "Neutros", siguiendo la metodología expuesta inicialmente para este tipo de ajuste. Además del balance de los datos se realiza un segundo ajuste en la parametrización del modelo que, para el caso de la Regresión Logística, se determina que se debe iterar sobre el parámetro del costo, el cual con una exactitud del 0,82 determina que el mejor valor para este parámetro en el modelo es $costo = 4$, mejorando los resultados en un 49% sobre el valor inicial. Para validar esta medida de desempeño se implementó la metodología de validación cruzada con 10 iteraciones, la cual comprueba que el comportamiento del modelo es estable y se mantiene con un promedio de exactitud del 0,83 y una desviación de 0,06.

Tabla 6. Resultados modelo análisis de sentimientos para los tuits

Clase	Antes de parametrización y balance		Después de parametrización y balance		Promedio Exactitud Validación Cruzada
	Exhaustividad	Exactitud	Exhaustividad	Exactitud	
Positivo	0,12	0,55	0,83	0,82	0,83 ± 0,06
Negativo	0,16		0,85		
Neutro	0,98		0,78		

Fuente: Elaboración propia

4.3. Desarrollar un modelo para el perfilamiento de los usuarios de Twitter según los sentimientos de sus opiniones sobre tópicos de interés.

4.3.1. Modelo para la clasificación de Género

Para la clasificación de la variable "Género" se establecieron como clases del modelo: Femenino y Masculino, basándose en un primer filtro de clasificación por el nombre del usuario y posteriormente un modelo de clasificación SVM para aquellos tuits de los usuarios no identificados con el primer filtro.

4.3.1.1. Detección del Género por filtro: Con el fin de mejorar los resultados del modelo y tener una mayor precisión en la clasificación de los usuarios se identificó que es posible detectar el género al que pertenece por medio de su nombre. Los datos extraídos de Twitter contienen una variable *name* como una de las características de la cuenta del usuario. Antes de realizar la clasificación por el nombre del usuario se identifica que las cuentas pueden pertenecer tanto a personas como a instituciones públicas o privadas que no entran en el estudio y por lo tanto es necesario identificarlas, para lo cual se creó una nueva variable "Tipo de Cuenta" con los atributos "Persona" y "Empresa". De acuerdo con esto la metodología que se utilizó fue la siguiente:

- Filtro Tipo de Cuenta: Para este filtro se creó una lista "CuentasEmpresas" con el nombre de algunas instituciones públicas y privadas de toda clase, y una lista de palabras del que hacen uso este tipo de cuentas "DescripcionEmpresas". La primera lista se cruza con la variable *screen_name* incluida en la información del tuit con el fin de encontrar coincidencias y asignarle al usuario en la variable "Tipo de Cuenta" el atributo de "Empresa". La segunda lista se cruza con la variable *description* incluida en la información del tuit con el fin de identificar si la cuenta usa alguna de esas palabras y así asignarle el atributo de "Empresa". Si la variable "Tipo de Cuenta" no tiene asignado "Empresa" se le asigna el atributo de "Persona".
- Filtro Género: De acuerdo con la asignación final de la variable "Tipo de Cuenta" se toman aquellas cuentas que tienen el atributo de "Personas" y se procede con el filtro que detecta el género de los usuarios por su nombre. Para el filtro se crea una lista "Nombres" con los nombres de hombres y mujeres. La lista se cruza con la variable *name* y si hay coincidencias, la variable "Género" toma el atributo de acuerdo con el género que el nombre tiene en la lista.

4.3.1.2. Modelo de clasificación SVM: Luego de pasar por el filtro, los usuarios que no son reconocidos pasan a ser clasificados por el modelo de clasificación SVM, el cual se construyó a partir de la siguiente metodología:

- Obtención de la base de datos de entrenamiento: La base de datos de entrenamiento se obtuvo de una colección de tuits de un repositorio de internet, conocido como PAN y se tomaron los datos de perfilamiento de usuarios de Twitter del año 2016 (PAN, 2016). Esta base de datos se compone de un total de 9186 tuits pertenecientes a 98 usuarios, con un promedio de alrededor 94 tuits por usuario con una desviación de 11,27. Además, cada tuit tiene en promedio alrededor de 7 palabras con una desviación de 3,59. A continuación, se muestra el diagrama de la base de datos usada:

Ilustración 9. Diagrama base de datos modelo de clasificación Género

Fuente: Elaboración propia

1. Preprocesamiento de datos: Esta metodología es la misma aplicada en el modelo de análisis de sentimientos.

1.1. Convertir cadena de texto en minúsculas
1.2. Eliminación de ruido: Es un proceso de normalización de texto para las características especiales que puede presentar un tuit.
  1.2.1. Eliminación de usuarios y *retweets*
  1.2.2. Eliminación de *links*
  1.2.3. Eliminación de signos de puntuación
1.3. *Tokenización*
1.4. Eliminación de *stopwords*
1.5. Lematización

2. Vectorización: Esta metodología es la misma aplicada en el modelo de análisis de sentimientos.

3. Construcción del modelo de clasificación: El SVM es un algoritmo de aprendizaje supervisado que, con base en un conjunto de datos inicial, entrena al modelo para la clasificación binaria o múltiple. El SVM es un algoritmo basado en el principio estructurado de minimización de riesgos. Las variables explicativas se mapean a través de estructuras no lineales en un espacio de alta dimensión y luego, se genera un hiperplano que separa de manera óptima ambas clases. Este hiperplano tiene como objetivo minimizar los errores de clasificación mientras maximiza los márgenes o la suma de distancias desde el hiperplano hasta las muestras de entrenamiento más cercanas de cada clase (Robles-Velasco, Cortés, Muñuzuri & Onieva, 2019).

   - Inicialmente se decide probar el modelo dividiendo la base de datos de entrenamiento en un set de prueba con el 30% del total de tuits y un set de entrenamiento con el 70% restante. Cuando ya se ha realizado la repartición aleatoria de los datos, se entrena el modelo y se evalúa su rendimiento con el set de prueba, basándose en las medidas de desempeño de exactitud y exhaustividad.
   - Teniendo en cuenta los resultados anteriores se reconoce que la cantidad de tuits etiquetados en la base de datos de entrenamiento para cada una de las clases del modelo generan un sesgo en la clasificación. Para evitar que este tipo de sesgo altere los resultados se decide implementar una herramienta en *machine learning* que consiste en balancear los datos de entrenamiento para todas las clases. Para este balance se tuvo en cuenta el método *oversampling*, el cual consiste en obtener más datos de las clases minoritarias de los que originalmente había, dejando intacta la cantidad de elementos de la clase mayoritaria (Núñez, Angulo & Abril, 2012).
     Adicional a los cambios en el tamaño de los datos de entrenamiento se decide evaluar cambios en la parametrización del modelo. De acuerdo con los parámetros del modelo SVM, los que se pueden calibrar para obtener un mejor rendimiento son el "costo", "*kernel*" y "gamma". El "costo" es un parámetro que controla el número y la severidad de las violaciones del margen que se tolera en el proceso de ajuste, primero se decide iterar sobre rangos bajos del costo (1,5) y se evalúan las medidas de desempeño sobre los datos de entrenamiento y de prueba, teniendo en cuenta que dependiendo de su valor ya sea bajo o alto se puede generar respectivamente un subajuste o un sobreajuste del modelo, para evitar estos posibles sesgos se decide iterar nuevamente sobre rangos de costo más altos (6 ,20).

A partir del mejor resultado de exactitud que se reporte para cada rango, se toman estos dos valores del costo y se verifica la consistencia del modelo con validación cruzada tomando finalmente el valor del costo que presente un menor sesgo para la exactitud.
El parámetro "*kernel*" representa el comportamiento de los datos, para el caso de estudio se calibraron dos tipos, el lineal y el RBF (*Gaussian Kernel*). El parámetro "gamma" se debe tener en cuenta cuando se considera la función *kernel* de RBF, y es el que mide la distancia entre las observaciones que separan los subespacios del SVM (Betancourt, 2005)., este parámetro puede tomar los valores "*scale*", "*auto*" o un número mayor a 0, para el caso de estudio se calibraron los tipos "*scale*" y "*auto*", sugeridos por la librería, con el fin de restarle complejidad y porque estos dos parámetros son formulas definidas a partir de la distribución de los datos en el espacio y su nivel de dispersión (Sklearn, 2020).

4. Evaluación de las medidas de desempeño: Estas medidas de desempeño son las mismas empleadas en el análisis de sentimientos.

5. Validación cruzada: Esta metodología es la misma aplicada en el modelo de análisis de sentimientos.

6. Perfilamiento de usuarios: El modelo construido tiene como resultado la predicción de la clase para un set de tuits, sin embargo, dada la naturaleza de los datos de Twitter, se identifica que un usuario pudo registrar más de un tuit durante el periodo de tiempo en el que se ejecutó el módulo de extracción. En este sentido y teniendo en cuenta que el objetivo es perfilar usuarios, se necesita la clasificación a nivel individual dentro de cada una de las clases del modelo. Para clasificar al usuario en una de las categorías se utilizó un esquema de votación, donde se tiene en cuenta la predicción que el modelo hizo para cada uno de sus tuits y se realiza un conteo del número de ocurrencias de cada una de las clases dentro de estos, asignando finalmente la clase que mayor número de ocurrencias tuviera. Según la tabla 6 en la que se ejemplifica el esquema de votación para una de las variables del estudio, al usuario 1 se le asigna la clase (25-34) años la cual es predicha por el modelo en mayoría (tres tuits) sobre el total de los tuits del usuario (cinco tuits).

Tabla 7. Ejemplo esquema de votación

Usuario	Tuits	Edad predicha	Edad Esquema Votación
Usuario 1	Tuit 1	(25-34) años	(25-34) años
	Tuit 2	(25-34) años	
	Tuit 3	(25-34) años	
	Tuit 4	(18-24) años	
	Tuit 5	(35-49) años	

Fuente: Elaboración propia

Además de las clases del modelo, el esquema de votación puede etiquetar al usuario como "Sin clasificar" para aquellos en los que se presente el caso en que etiqueto varias clases en la misma proporción, es decir no existe una clase dominante entre los tuits del usuario.

Para la construcción del ***Modelo para la clasificación de Género*** se obtuvieron diferentes resultados debido a los ajustes que se realizaron con el fin de obtener los mejores, el primer resultado que se obtuvo fue para un modelo inicialmente planteado en el que no se tuvieron en cuenta los parámetros del SVM, ni el desbalance en el que se encontraban los datos de entrenamiento para cada una de las clases. Es importante mencionar que en este caso los resultados no quedan sesgados hacia la clase mayoritaria, en decir, "Femeninos" debido a que la diferencia con los tuits "Masculinos" es muy pequeña, por lo cual no existe un mayor impacto de esta medida. El valor de la exactitud para todo el modelo resultante fue de 0,66.

Al reportar los primeros resultados y a pesar de que no hay un mayor sesgo entre las clases se decide hacer un primer ajuste en el balance de datos, manteniéndolo como una buena práctica en el modelamiento, tomando como referencia la categoría con más datos de entrenamiento, en este caso la de tuits "Femeninos", siguiendo la metodología expuesta inicialmente para este tipo de ajuste. Además del balance de los datos se realiza un segundo ajuste en la parametrización del modelo que para el caso de SVM, se determina que se debe iterar para calibrar sobre el parámetro del "costo", "*kernel*" y "gamma", los cuales con una exactitud del 0,78 determinan que los mejores valores son $costo = 2$, $kernel = 'rbf'$, $gamma = 'scale'$, mejorando los resultados en un 18% sobre el valor inicial. Para validar esta medida de desempeño se implementó la metodología de validación

cruzada con 10 iteraciones, la cual comprueba que el comportamiento del modelo es estable y se mantiene con un promedio de exactitud del 0,76 y una desviación de 0,01.

Tabla 8. Resultados modelo SVM género para los tuits

Clase	Antes de parametrización y balance		Después de parametrización y balance		Promedio Exactitud Validación Cruzada
	Exhaustividad	Exactitud	Exhaustividad	Exactitud	
Femenino	0,64	0,66	0,84	0,78	0,76± 0,01
Masculino	0,68		0,71		

Fuente: Elaboración propia

4.3.2. Modelo para la clasificación de Ocupación

Para la clasificación de la variable "Ocupación" se establecieron como clases del modelo: Artes, Ciencias Económicas y Administrativas (CEA), Ciencias Sociales y Jurídicas (CSJ), Educación, Ingeniería y Ciencias, Salud y Otros, basándose en un primer filtro de clasificación por la descripción y posteriormente un modelo de clasificación SVM para aquellos tuits de los usuarios no identificados con el primer filtro.

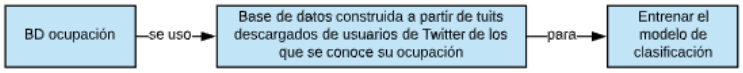
4.3.2.1. Detección de la Ocupación por filtro: Los datos extraídos de los usuarios de Twitter tienen un campo denominado *description* en donde se encuentra una pequeña descripción del usuario y muchas veces la ocupación de este, por lo que se realizó un primer filtro en el que se busca la ocupación exacta del usuario. Teniendo en cuenta que existen múltiples ocupaciones se definieron 7 categorías para la clasificación de la ocupación final. De acuerdo con esto la metodología que se utilizó fue la siguiente:

- Filtro Ocupación: Solo se tuvo en cuenta aquellas cuentas que tienen el atributo de "Personas" de acuerdo con la variable "Tipo de Cuenta" identificada en la clasificación de género y se procede al filtro que detecta la ocupación por la descripción del usuario. Para el filtro se creó una lista "Ocupaciones" con un total de 881 ocupaciones. La lista se cruza con la variable *description* y si hay coincidencia la variable "Ocupación" toma el atributo de acuerdo con la categoría a la que pertenece la ocupación encontrada en la lista.

4.3.2.2. Modelo de clasificación SVM: Luego de pasar por el primer filtro, los usuarios que no cuentan con una ocupación en su descripción pasan a ser clasificados por el modelo de clasificación SVM, el cual se construyó a partir de la siguiente metodología:

- Obtención de la Base de Datos de Entrenamiento: Para este tipo de modelo y de clasificación no se encuentran datos públicos en internet, por lo que se construyó una base de datos con los tuits de usuarios de Twitter de los que gracias al primer filtro se conoce su ocupación. Esta base de datos se compone de un total de 30446 tuits pertenecientes a 207 usuarios, con un promedio de alrededor 147 tuits por usuario con una desviación de 13,83. Además, cada tuit tiene en promedio alrededor de 7 palabras con una desviación de 4,01. A continuación, se muestra el diagrama de la base de datos usada:

Ilustración 10. Diagrama base de datos modelo de clasificación Ocupación



Fuente: Elaboración propia

1. Preprocesamiento de datos: Esta metodología es la misma aplicada en el modelo para la clasificación de género.

2. Vectorización: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

3. Construcción del modelo de clasificación: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

4. Evaluación de las medidas de desempeño: Estas medidas de desempeño son las mismas empleadas en el análisis de sentimientos.

5. Validación Cruzada: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

6. Perfilamiento de usuarios: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

Para la construcción del ***Modelo para la clasificación de Ocupación*** se obtuvieron diferentes resultados debido a los ajustes que se realizaron con el fin de obtener los mejores, el primer resultado que se obtuvo fue para un modelo inicialmente planteado en el que no se tuvieron en cuenta los parámetros del SVM, ni el desbalance en el que se encontraban los datos de entrenamiento para cada una de las clases. Es importante mencionar que en este caso los resultados no quedan sesgados hacia la clase mayoritaria, es decir, "CEA" debido a que la diferencia con los tuits de las demás clases es muy pequeña, por lo cual no existe un mayor impacto de esta medida. El valor de la exactitud para todo el modelo resultante fue de 0,32.

Al reportar los primeros resultados y a pesar de que no hay un mayor sesgo entre las clases se decide hacer un primer ajuste en el balance de datos, manteniéndolo como una buena práctica en el modelamiento, tomando como referencia la categoría con más datos de entrenamiento, en este caso la de tuits "CEA", siguiendo la metodología expuesta inicialmente para este tipo de ajuste. Además del balanceo de los datos se realiza un segundo ajuste en la parametrización del modelo que para el caso de SVM, se determina que se debe iterar para calibrar sobre el parámetro del "costo", "*kernel*" y "gamma", los cuales con una exactitud del 0,58 determinan que los mejores valores son $costo = 5$, $kernel = 'rbf'$, $gamma = 'scale'$, mejorando los resultados en un 81% sobre el valor inicial. Para validar esta medida de desempeño se implementó la metodología de validación cruzada con 10 iteraciones, el cual comprueba que el comportamiento del modelo es estable y se mantiene con un promedio de exactitud del 0,55 y una desviación de 0,01.

Tabla 9. Resultados modelo SVM ocupación para los tuits

Clase	Antes de parametrización y balanceo		Después de parametrización y balanceo		Promedio Exactitud Validación Cruzada
	Exhaustividad	Exactitud	Exhaustividad	Exactitud	
Artes	0,33	0,32	0,67	0,58	0,55 ± 0,01
Ciencias económicas y admón.	0,35		0,55		
Ciencias sociales y jurídicas	0,45		0,55		
Educación	0,24		0,53		
Ingeniería y ciencias	0,34		0,57		
Salud	0,27		0,56		
Otros	0,30		0,61		

Fuente: Elaboración propia

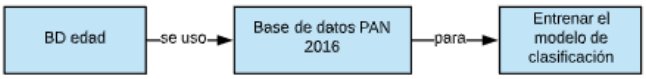
4.3.3. Modelo para la clasificación de Edad

Para la clasificación de la variable "Edad" se establecieron como clases del modelo: (18-24), (25-34), (35-49) y (50-XX), basándose en un modelo de clasificación SVM.

4.3.3.1. Modelo de clasificación SVM: Los usuarios es este modelo son clasificados por el modelo de clasificación SVM, el cual se construyó a partir de la siguiente metodología:

- Obtención de la Base de Datos de Entrenamiento: La base de datos de entrenamiento se obtuvo de una colección de tuits de un repositorio de internet, conocido como PAN y se tomaron los datos de perfilamiento de usuarios de Twitter del año 2016 (PAN, 2016). Esta base de datos se compone de un total de 9186 tuits pertenecientes a 98 usuarios, con un promedio de alrededor 94 tuits por usuario con una desviación de 11,27. Además, cada tuit tiene en promedio alrededor de 7 palabras con una desviación de 3,59. A continuación, se muestra el diagrama de la base de datos usada:

Ilustración 11. Diagrama base de datos modelo de clasificación Edad



Fuente: Elaboración propia

1. Preprocesamiento de datos: Esta metodología es la misma aplicada en el modelo para la clasificación de género.

2. Vectorización: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

3. Construcción del modelo de clasificación: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

4. Evaluación de las medidas de desempeño: Estas medidas de desempeño son las mismas empleadas en el análisis de sentimientos.

5. Validación Cruzada: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

6. Perfilamiento de usuarios: Esta metodología es la misma aplicada en el modelo para la clasificación del género.

Para la construcción del ***Modelo para la clasificación de Edad*** se obtuvieron diferentes resultados debido a los ajustes que se realizaron con el fin de obtener mejores resultados, el primer resultado que se obtuvo fue para un modelo inicialmente planteado en el que no se tuvieron en cuenta los parámetros del SVM, ni el desbalance en el que se encontraban los datos de entrenamiento para cada una de las clases, en este caso los resultados quedaban sesgados hacia las clases mayoritarias como se evidencia en la tabla 10 que para la clase "(25-34)" reportaba el valor de exhaustividad más alto y la exactitud para todo el modelo resultante fue de 0,54.

Al reportar los primeros resultados se decide hacer un primer ajuste en el balance de datos tomando como referencia la categoría con más datos de entrenamiento, en este caso la de tuits "(25-34)", siguiendo la metodología expuesta inicialmente para este tipo de ajuste. Además del balance de los datos se realiza un segundo ajuste en la parametrización del modelo que para el caso de SVM, se determina que se debe iterar para calibrar sobre el parámetro del "costo", "kernel" y "gamma", los cuales con una exactitud del 0,86 determinan que los mejores valores son $costo = 4$, $kernel =' rbf'$, $gamma =' scale'$, mejorando los resultados en un 60% sobre el valor inicial. Para validar esta medida de desempeño implementamos la metodología de validación cruzada con 10 iteraciones, la cual comprueba que el comportamiento del modelo es estable y se mantiene con un promedio de exactitud del 0,84 y una desviación de 0,01.

Tabla 10. Resultados modelo SVM edad para los tuits

Clase	Antes de parametrización y balanceo		Después de parametrización y balanceo		Promedio Exactitud Validación Cruzada
	Exhaustividad	Exactitud	Exhaustividad	Exactitud	
(18-24)	0,38	0,54	0,83	0,86	0,84 ± 0,01
(25-34)	0,84		0,84		
(35-49)	0,27		0,82		
(50-XX)	0,19		0,97		

Fuente: Elaboración propia

4.4. Analizar el impacto del sentimiento de las opiniones sobre temas relacionados con la pandemia del COVID-19, en la sociedad.

Los modelos anteriormente descritos permiten que, a partir de los tuits recolectados del módulo de extracción de Twitter se pueda perfilar a los usuarios que tengan alguna opinión acerca del manejo que le ha dado el gobierno colombiano al COVID-19. De la base de datos recolectada con 189,492 tuits de 30,163 usuarios se detectó que el 19,3% de tuits pertenecientes a 12,913 usuarios corresponden a este tema. La clasificación de tuits por tópico se muestra en la ilustración 13(a), siendo el tópico 1: Contenido multimedia en cuarentena,

tópico 2: Día de la madre y amor, tópico 3: Actividades en cuarentena, tópico 4: Manejo del COVID-19 por el gobierno y sin clasificar: tweets que no pertenecen a ninguno de los tópicos anteriores ya que la ponderación por cada uno de los temas fue la misma.

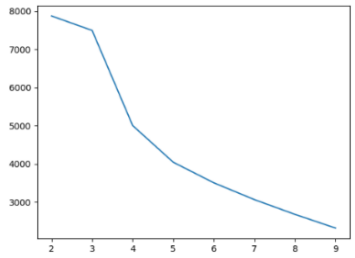
A partir de la construcción de los modelos para análisis de sentimientos y perfilamiento se decide implementar un algoritmo de *clustering* que permita segmentar a los usuarios de Twitter de acuerdo con las características identificadas en los modelos, con el fin de determinar las características que definen a aquellos usuarios que tienen alguna opinión con respecto al tema del manejo del COVID-19 por el gobierno. La metodología que se siguió para la construcción del modelo fue la siguiente:

1. Definición de las variables de interés: Para realizar la agrupación de usuarios en *clusters* se identifica que los resultados del modelo de análisis de sentimientos se encuentran clasificados a nivel de tuit, pero es necesario identificar cuál es su tipo de sentimiento dominante con respecto al tema seleccionado. Para llevar a nivel de usuario los resultados del modelo a nivel de tuit, se utiliza el mismo esquema de votación que se aplicó para perfilar usuarios según los resultados de los modelos de clasificación de género, ocupación y edad. En este sentido, se reconoce que el esquema de votación para perfilamiento y análisis de sentimientos permite que se analicen de forma individual a los usuarios.

2. Determinación del algoritmo: Con base en el tipo de variables con las que se perfilaron los usuarios, y debido a su naturaleza categórica se decide implementar una versión del algoritmo *K-means* que se adapta a este tipo de variables. El algoritmo seleccionado es *K-modes*, el cual se diferencias de *K-means* en la medida de similitud usada para la actualización de la posición de los centroides, ya que reemplaza el uso de promedios por el de modas y utiliza un método basado en frecuencias para actualizarla. Los parámetros de este tipo de algoritmo varían dependiendo las características de los datos, el primero es el valor de "n_init" el cual mide la aleatoriedad del modelo, este parámetro se modifica con el fin de que se estabilice el modelo. El segundo es el "init", el cual define el método de inicialización de los centroides teniendo en cuenta la distribución de los datos en el espacio. El último parámetro es el "n_clusters", el número de segmentos se define con base en la minimización de un costo objetivo (López, 2007).

3. Calibración del modelo: Con el fin de obtener la mejor segmentación de datos posible, basado en una medida de costo que evalúa el desempeño del modelo, se calibran los parámetros "init" y "n_clusters" a partir de escenarios con los siguientes valores de "n_init" ($n_{init} = 10$, $n_{init} = 20$, $n_{init} = 30$, $n_{init} = 40$, $n\_init = 50$) y replicando 5 veces cada escenario, con el fin de encontrar el parámetro que estabiliza el modelo. De acuerdo a los escenarios planteados y teniendo en cuenta que a partir de $n_{init} = 50$ el modelo se estabiliza, se itera sobre los dos posibles valores de "init" ('*Ciao*', '*Huang*') y se concluye que el valor que minimiza el costo es $init =' Ciao'$ independientemente del número de segmentos que se determinen. A partir de la definición de estos dos parámetros se grafica el comportamiento del costo vs el número de segmentos propuestos en el rango de (2,10):

Ilustración 12. Tendencia de los costos vs número de segmentos



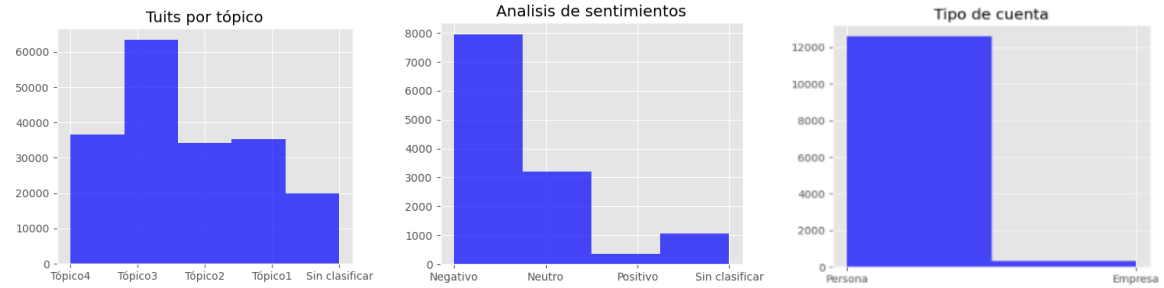
Fuente: Python

A partir de la identificación de los tuits y los usuarios que tienen alguna opinión acerca del manejo que le ha dado el gobierno colombiano al COVID-19, se aplica el modelo de análisis de sentimientos que ya ha sido entrenado. Teniendo en cuenta los resultados del modelo y que los usuarios tienen una opinión acerca del tema

seleccionado se realiza un esquema de votación con el fin de que se pueda analizar el tipo de sentimientos que tienen los usuarios en Twitter. Los resultados del esquema de votación demuestran que el 66,85% de personas piensan negativamente sobre este tema, el 23,10% piensan neutralmente, el 1,50% piensan positivamente y el 8,56% no fue clasificado. El histograma del análisis de sentimientos se muestra en la ilustración 13(b).

Cuando se identifica el tipo de sentimiento que tiene el usuario con respecto al tema, se inicia el proceso de perfilamiento, con el filtro que clasifica las cuentas de los usuarios en "Persona" o "Empresa" dependiendo de las características de su cuenta. La ilustración 13(c) muestra los resultados, de acuerdo con la base de datos de 36,668 tuits de 12,913 usuarios, en donde se tiene que el 97,3% de usuarios son personas y el 2,7% son empresa.

Ilustración 13. Histogramas para la clasificación por tópico, análisis de sentimientos y tipo de cuenta

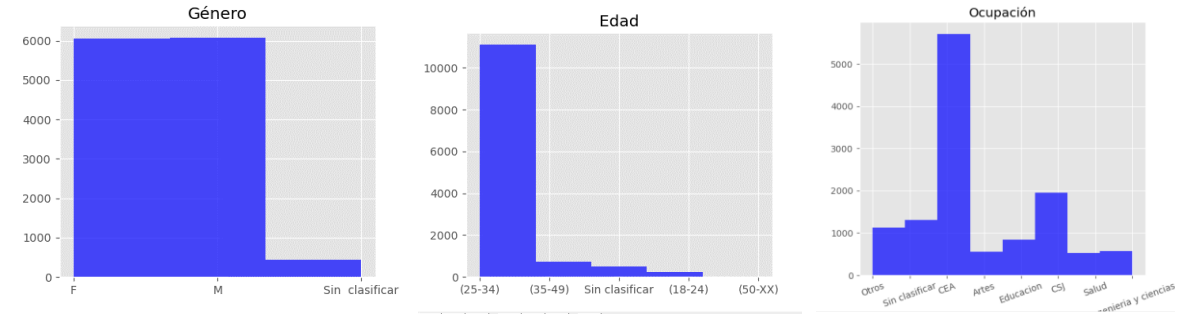


(a) (b) (c)

Fuente: Python

A partir de la clasificación de las cuentas por "Tipo Cuenta", se procede a la aplicación de los modelos de perfilamiento usando una base de datos únicamente de cuentas clasificadas como "Persona". A partir de los resultados de los modelos de perfilamiento a nivel de tuit se clasifican con el esquema de votación a los usuarios dentro de una de las clases, los resultados se muestran en la ilustración 14 con histogramas para cada una de las variables donde se evidencia que para cada una de ellas existe una clase dominante, para el género es la clase "Masculino", para la edad "(25-34)" años y para la ocupación "Ciencias Económicas y Administrativas (CEA)".

Ilustración 14. Histogramas variables de perfilamiento Género, Edad y Ocupación

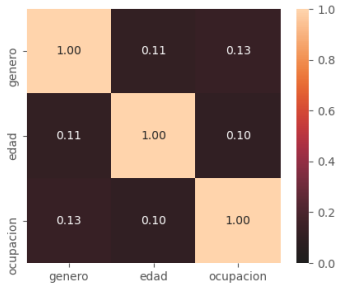


(a) (b) (c)

Fuente: Python

De acuerdo con los resultados obtenidos, se graficó la correlación que existe entre las variables demográficas estudiadas para cada uno de los usuarios, esto con el fin de analizar la dependencia entre estas para la implementación de un algoritmo de *clustering*. Para esto se hizo uso de la librería *dython*, la cual permite trazar la correlación y asociación de variables categóricas, haciendo uso de la medida de asociación "V de Cramer". Como se evidencia en la siguiente ilustración 15, la correlación entre pares de variables no es significativa ya que el valor más alto es de 0,13 lo que nos indica que las variables demográficas son independientes y se pueden usar en el algoritmo de *clustering*.

Ilustración 15. Correlación entre variables demográficas

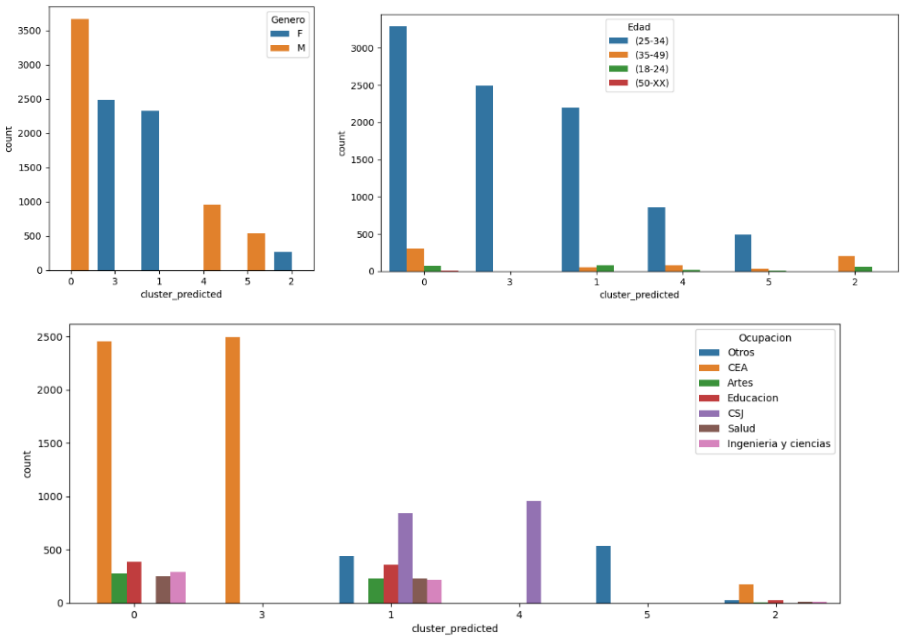


Fuente: Python

Finalmente, se decide agrupar a los usuarios en 6 segmentos con el fin de obtener mejores resultados en términos de interpretación para cada uno de los grupos y porque como se observa en la gráfica de la ilustración 12 a partir de este valor para el número de segmentos la variación en los costos se vuelve constante y no se evidencian cambios significativos. En ese punto el valor del costo es $costo = 3504$.

Cuando se han definido los parámetros se corre el algoritmo y se grafica el comportamiento de cada una de las variables de perfilamiento dentro de cada segmento.

Ilustración 16. Segmentación de usuarios



Fuente: Python

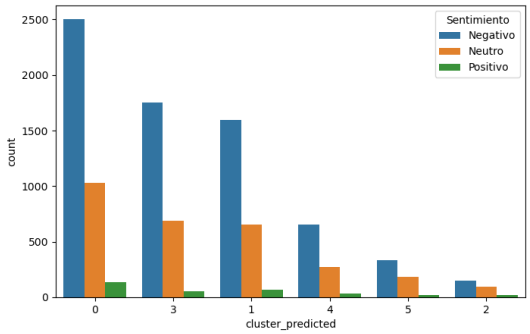
A partir de la creación de los segmentos de usuarios se describe cada uno de ellos con las siguientes características:

- Segmento 0: Este segmento está conformado en su totalidad por hombres, la mayoría entre los 25 y 34 años y pertenecientes a la categoría de ocupación Ciencias Económicas y Administrativas (CEA).
- Segmento 1: Este segmento está conformado en su totalidad por mujeres, la mayoría entre los 25 y 34 años y pertenecientes a la categoría de ocupación Ciencias Sociales y Jurídicas (CSJ).
- Segmento 2: Este segmento está conformado en su totalidad por mujeres, la mayoría entre los 35 y 49 años y pertenecientes a la categoría de ocupación Ciencias Económicas y Administrativas (CEA).
- Segmento 3: Este segmento está conformado en su totalidad por mujeres entre los 25 y 34 años y pertenecientes a la categoría de ocupación Ciencias Económicas y Administrativas (CEA).

- Segmento 4: Este segmento está conformado en su totalidad por hombres, en su mayoría entre los 25 y 34 años pertenecientes a la categoría de ocupación Ciencias Sociales y Jurídicas (CSJ).
- Segmento 5: Este segmento está conformado en su totalidad por hombres, en su mayoría entre los 25 y 34 años pertenecientes en su totalidad a la categoría de ocupación Otros.

Para todos los segmentos de usuarios se reconoció el tipo de sentimiento que se registra para cada uno de los segmentos con respecto al tema del manejo del COVID-19 por el gobierno.

Ilustración 17. Sentimientos por segmento



Fuente: Python

De acuerdo con los resultados de la segmentación vs el tipo de sentimiento en cada una de ellas, se puede reconocer que los segmentos 0, 3 y 1 tienen una tendencia pesimista más marcada que los otros segmentos ya que su proporción de sentimientos negativos es significativamente mayor sobre los positivos y los neutros casi en un 100%. Sobre los otros segmentos el 4, 6 y 2 se puede concluir que a pesar de que su proporción de negativos es mayor que los positivos y neutros, su tendencia no es tan marcada hacia el pesimismo como los otros segmentos. Finalmente, según estos resultados se puede concluir que los usuarios de la red social Twitter tienen una predominancia hacia sentimientos negativos acerca del manejo que le está dando el gobierno al COVID-19, manifestando así un alto grado de inconformidad.

Es importante que el gobierno evalué este tipo de comportamientos en la población ya que le permiten medir la percepción y el nivel de aceptabilidad que tienen sus políticas públicas. Como lo mencionan Ceron y Negri, esta información puede ayudar a los responsables políticos a calificar las alternativas políticas disponibles de acuerdo con las preferencias de los ciudadanos para monitorear los comportamientos, opiniones y percepciones de estos (Ceron & Negri, n.d). En este caso es evidente que es necesario evaluar una posible reorientación de estas políticas dirigidas a atender la crisis que se ha originado por el COVID-19, ya que se evidencia un nivel de desaprobación cercano al 70% de los usuarios evaluados en esta investigación.

Las diferentes interpretaciones de datos gubernamentales en redes sociales, pueden afectar enormemente la formulación de políticas y la relación de confianza entre gobierno-ciudadano, por esto es importante que se realicen este tipo de investigaciones que no solo ayudan a medir los niveles de aceptabilidad, sino que también permiten que el gobierno reconozca el perfil de las personas que las usan y así mejorar la toma de decisiones enfocadas en determinados sectores de la población (Chen, Franks & Evans, 2016). En este sentido es importante reconocer que la opinión de los segmentos de usuarios que mayor participación tienen en las redes sociales- mujeres y hombres entre los 25 y los 34 años que su ocupación está clasificada entre “CEA” y “CSJ”- sean tenidos en cuenta como representación de las necesidades de la mayoría en el país que muestran un alto grado de inconformidad con las medidas tomadas hasta ahora en relación con el COVID-19. También es importante que se canalicen estas medidas a través de fuertes campañas de difusión en redes sociales, que les transmita confianza a estos usuarios, ya que son medidas que afectan a todo el país, y que de su nivel de aceptación depende su nivel de cumplimiento.

5. Componente de Diseño en ingeniería

5.1. Declaración de Diseño: Diseño de un modelo para el análisis de sentimientos de los intereses actuales de las personas en Colombia, además del diseño de un módulo computacional para la obtención de datos de la red social Twitter.

5.2. Requerimientos de desempeño: A partir del diseño se espera tener información actualizada que permita conocer los intereses de las personas y su comportamiento, todo a partir de fuentes de datos ágiles y económicas, como son las redes sociales, sin tener que implementar formatos tradicionales como encuestas, debido a su alta inversión de recursos

5.3. Restricciones:

- Limitaciones en el uso de la API de Twitter: En el diseño del módulo para la obtención de datos de Twitter se hará uso de una interfaz de programación de aplicaciones (API) que permite acceder de forma gratuita a datos de usuarios. La aplicación al ser gratuita tiene restricciones de uso, como volumen y periodicidad de la información.
- Usuarios de Twitter: A pesar de las altas cifras de uso de la red social Twitter, según el Informe Digital 2019 realizado por las firmas *Hootsuite y we are social*, los usuarios que más hacen uso de la plataforma se encuentran entre los 18 y los 49 años y son personas en su mayoría con niveles de ingresos medio-alto. La información con la que se construirá el módulo no representa en su totalidad a la población general.
- Estructura de los tuits: Twitter es una red social de opinión que permite que los usuarios se expresen en tuits de 140 caracteres, lo cual puede generar inconvenientes porque los usuarios deben abreviar sus opiniones.

5.4. Cumplimiento del estándar: Para el desarrollo de los procesos de diseño del proyecto, se tomará como base la metodología *CRISP-DM* 2000, una guía utilizada en proyectos de minería de datos que se encuentra dividida en cuatro niveles de abstracción organizados de forma jerárquica en tareas que van desde el nivel más general hasta el más específico: fases, tareas generales, tareas especializadas e instancias de proceso. A su vez la metodología organiza el desarrollo de un proyecto de minería de datos en una serie de seis fases: comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación (IBM, 2012). En el *anexo 2* se enmarca todo lo realizado, dentro de las etapas de *CRISP-DM*.

6. Limitaciones, conclusiones y recomendaciones

La cantidad de datos que se pueden extraer de un tuit, incluyendo los de tipo *JSON* nos brindan una gran cantidad de información para analizar, lo que hace que el alcance en proyectos de minería de datos sea significativamente alto. Además, al ser Twitter una red social donde los usuarios pueden expresar sus sentimientos, pensamientos, criticas, puntos de vista, deseos, expectativas, entre otros., es ideal para entender como diferentes escenarios están influyendo en el público. Teniendo esto en cuenta, podemos concluir que esta red social nos proporciona un flujo de información rápido, actualizado y en gran cantidad para poder analizar y concluir sobre la postura de los usuarios no solo en cuanto a un evento especifico, sino también en cuanto a las opiniones que se pueden tener acerca de un producto o servicio, lo que hace que pueda ser usada por diferentes entidades y con diferentes fines.

De acuerdo con los resultados obtenidos en los modelos de clasificación, se identificó que un análisis detallado de los parámetros y características del modelo nos permite obtener medidas de desempeño aceptables, esto basado en que los resultados finales fueron superiores a los referenciados en la investigación inicial. Las herramientas que en este caso permiten estas mejoras, se basaron en el balance de los datos y la calibración de los parámetros, que se pudieron verificar con una validación cruzada que asegura un comportamiento estable del modelo en un rango determinado de exactitud. También es importante para el análisis de este tipo de variables reconocer otras herramientas que ayuden a mejorar la predicción, como los filtros para las variables de género y ocupación, basados en la información compartida por los usuarios en redes sociales.

Teniendo en cuenta las características de los datos extraídos y los resultados obtenidos en los modelos de clasificación a lo largo del proyecto, se puede concluir que es posible la creación de un modelo de análisis de sentimiento sobre las opiniones expresadas por usuarios de Twitter en Colombia, sobre tópicos de interés actuales, como el manejo de la pandemia del COVID-19 por el gobierno y que permita perfilar dichos usuarios en términos de su edad, género y ocupación.

El algoritmo de *clustering k-modes* permite segmentar a usuarios de Twitter que tienen alguna opinión con respecto al tema del manejo del COVID-19 por el gobierno de acuerdo con las características identificadas a partir de los modelos de perfilamiento. Para obtener los mejores resultados de segmentación es importante calibrar los parámetros del algoritmo, los de mayor impacto en los resultados son el número de clusters y el método de inicialización de los centroides.

La segmentación y análisis de resultados requiere identificar las características demográficas de cada usuario, y su tipo de sentimiento dominante. Para cumplir con este requerimiento, se utilizó un esquema de votación con el fin de obtener los resultados en términos del usuario a partir de la predicción que el modelo hizo para cada uno de sus tuits. El resultado de este típico esquema de votación permite perfilar de manera adecuada a los usuarios ya que la clase asignada es la más representativa.

En este proyecto se trabajaron diferentes herramientas flexibles de *Machine Learning* y *Data Analytics* que permiten que instituciones tanto públicas como privadas tomen decisiones respecto a un tema en específico conociendo desde varias perspectivas a los usuarios y sus tendencias de comportamiento, basándose en información de las redes sociales que son un nuevo foco de información sin explorar. En este caso la investigación se enfocó en evaluar el tipo de sentimientos que estaban manifestando las personas con respecto a las políticas que ha implementado el gobierno de Colombia con respecto a la emergencia sanitaria por el COVID-19 y es importante reconocer que a diferencia de los análisis tradicionales que suelen hacer este tipo de entidades basados en encuestas, este estudio permite que conozcan en tiempo real la opinión de los ciudadanos, proporcionando un flujo de información rápido y económico que les va a permitir reaccionar de manera ágil y pertinente ante los cambios.

El LDA es un modelo probabilístico que identifica patrones latentes a partir de observaciones y por lo tanto la inferencia de sus parámetros depende del número y la consistencia de los datos. En este sentido los tuits al ser textos cortos pueden no aportar la suficiente cantidad de información para que el modelo identifique esos patrones en los que se basa para crear el conjunto de palabras. De modo que a partir de lo observado en los resultados se recomienda que se evalúen otros modelos que tengan una mayor adaptabilidad para a este tipo de texto con el fin de mejorar el nivel de coherencia entre el conjunto de palabras que representa a cada tópico.

Se recomienda la creación de un aplicativo, en donde los interesados puedan extraer los datos y almacenarlos automáticamente, reentrenar, calibrar y ejecutar los modelos de manera sencilla y visualizar los resultados en un ambiente interactivo, esto con el fin de que la información sea de entendimiento general tratando de responder a un público interdisciplinar.

Se recomienda para futuros estudios, considerar el análisis del tiempo y las coordenadas en las que se crea el tuit, con el fin de conocer si estas variables afectan el sentimiento de las opiniones de los usuarios con respecto a un tema de interés. Tener en cuenta si según el día y la hora en la que se crean los tuits ocurrió algún evento importante que genera una reacción en los usuarios o si según la ubicación de los usuarios en Colombia, por ejemplo, para cada ciudad, predomina un sentimiento, podría brindar un mayor alcance del estudio.

Referencias

Abdallah, E., Alzghoul, J., & Alzghool, M. (2020). Age and Gender prediction in Open Domain Text. Procedia Computer Science, 170, 563–570.

Aletras, N., & Chamberlain, B. P. (2018). Predicting twitter user socioeconomic attributes with network and language information. HT 2018 - Proceedings of the 29th ACM Conference on Hypertext and Social Media, (July), 20–24.

Al-Ghadir, A., & Azmi, A. (2016). A Study of Arabic Social Media Users—Posting Behavior and Author's Gender Prediction. Cognitive Computation, 11(1), 71–86.

Aloqaily, A., Al-Hassan, M., Salah, K., Elshqeirat, B., & Almashagbah, M. (2020). Sentiment analysis for Arabic tuits datasets: Lexicon-based and machine learning approaches. Journal of Theoretical and Applied Information Technology, 98(4), 612–623.

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016, 61–66.

Betancourt, G. (2014). LAS MÁQUINAS DE SOPORTE VECTORIAL(SVMs) Scientia Et Technica, vol. XI, núm. 27, abril, 2005, pp. 67-72

Blei, D., Ng, A. & Jordan, M. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3 2003, 991-1022.

Breiman, L. (2001). Random forests. Kluwer Academic Publishers. Manufactured in The Netherlands. Machine Learning, 45, 5–32.

Ceron, A., & Negri, F. (n.d.). The " social side " of public policies. Using sentiment analysis to support the action of policy-makers across the policy cycle. 1–16.

Chamorro, V. (2018). Clasificación de tuits mediante modelos de aprendizaje supervisado. Trabajo Fin Máster en Ingeniería Informática. Universidad complutense de Madrid.

Chen, H., Franks, P., & Evans, L. (2016). Exploring Government Uses of Social Media through Twitter Sentiment Analysis. 14(5).

Facebook Developers. (2020). Developer account settings. Facebook for Developers. Recuperado de: https://developers.facebook.com/docs/apps/developer-settings

Gallardo, J. A. (n.d.). CRISP-DM Metodología para el Desarrollo de Proyectos de Minería de Datos.

Gupta, M. (2017). Implementation of Event Extraction from Twitter using LDA. 8(5), 381–386.

Hammoe, L. (2018). Detección de tópicos utilizando el modelo LDA (Tesis de especialización). Instituto Tecnológico de Buenos Aires, Buenos Aires

Heidenreich, H. (2018). Introduction to Word Embeddings. Towards data science. Recuperado de: https://towardsdatascience.com/introduction-to-word-embeddings-4cf857b12edc

Honing Data Science. (2020). Topic Modeling using Latent Dirichlet Allocation (LDA). Honingds. Recuperado de: https://honingds.com/blog/topic-modeling-latent-dirichlet-allocation-

IBM, I. B. M. (2012). Manual CRISP-DM de IBM SPSS Modeler. IBM Corporation, 56.

Kapadia, S. (2019). Evaluate Topic Models: Latent Dirichlet Allocation (LDA). Towards data science. Recuperado de: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Kuhn M. (2019). Tha caret Package. Recuperado de: https://topepo.github.io/caret/index.html

López, S. (2007). Algoritmos de Agrupamiento Global para Datos Mezclados. Master.

Ministerio de Tecnologías de la Información y las Comunicaciones. (2018). Colombia es uno de los países con más usuarios en redes sociales en la región. MINTIC. Recuperado de https://www.mintic.gov.co/portal/604/w3-article-2713.html

Montesinos, L. (2014). Análisis de sentimientos y predicción de eventos en twitter. Memoria para optar al título de ingeniero civil eléctrico. Universidad de Chile.

Nuñez, H., Angulo, C., & Gonzalez-Abril, L. (2011). Modificación del sesgo de una SVM entrenada sobre clases no balanceadas. (1), 32–38.

Ostrowski, D. A. (2015). Using latent dirichlet allocation for topic modelling in twitter. Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015, 493–497.

Pack, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

Pandya, A., Oussalah, M., Monachesi, P., & Kostakos, P. (2020). On the use of distributed semantics of tweet metadata for user age prediction. Future Generation Computer Systems, 102, 437–452.

Pech, F. (2019). Minería de datos en Twitter con Python: Colección de datos y frecuencia de términos. Recuperado de: http://rios.tecnm.mx/cdistribuido/recursos/MinDatScr/MineriaScribble.html

Preot'iuc-Pietro, D., Lampos, V., & Aletras, N. (2015). An analysis of the user occupational class through Twitter content. 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference, 1, 1754–1764.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2014). Classifying latent user attributes in Twitter. International Conference on Information and Knowledge Management, Proceedings, 37–44.

Rivadeneira, C. (2018). Estudio del análisis de sentimientos en redes sociales para la prescripción de situaciones financieras.

Robles-Velasco, A., Cortés, P., Muñuzuri, J., & Onieva, L. (2019). Prediction of pipe failures in water supply networks using logistic regression and support vector classification. Reliability Engineering & System Safety, 106754.

Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106, 92–104.

Siddharth, S., Darsini, R., & Sujithra, M. (2018). Sentiment Analysis on Twitter Data Using Machine Learning Algorithms in Python. ISSN (Online) 2394-2320 International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 5(2), 285–291.

Singh, J., Kaur, I., & Singh, A. K. (2020). Event detection from Twitter data. 2019 4th International Conference on Information Systems and Computer Networks (ISCON), 793–798.

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who tuits? deriving the demographic characteristics of age, occupation, and social class from twitter user meta-data. PLoS ONE, 10(3).

Sobrino, J. (2018). Análisis de sentimientos en Twitter (Tesis de maestría). Universidad Oberta de Catalunya.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). Exploring topic coherence over many models and many topics. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Proceedings of the Conference, (July), 952–961.

Suri, P., & Roy, N. R. (2017). Comparison between LDA & NMF for event-detection from large text stream data. 3rd IEEE International Conference On, 1–5.

Syahputra, H., Basyar, L. K., & Tamba, A. A. S. (2020). Setiment Analysis of Public Opinion on the Go-Jek Indonesia Through Twitter Using Algorithm Support Vector Machine. Journal of Physics: Conference Series, 1462(1).

Tabares, F., & Hernández, J. (2014). Big Data Analytics: Oportunidades, Retos y Tendencias.

Tan P., Steinbach M., Karpatne A., & Kumar V. (2019). Introduction to Data Mining. 2nd edition. Pearson

Thompson, A. (2018). Sentiment Analysis for social Good. Medium. Recuperado de: https://medium.com/@BrotherAustin/sentiment-analysis-for-social-good-b13fef52b3d3

Twitter Developers. (2020). Información sobre la API de Twitter. Twitter. Recuperado de: https://help.twitter.com/es/rules-and-policies/twitter-api

Van de Loo, J., Pauw, G., & Daelemans, W. (2016). Text-Based Age and Gender Prediction for Online Safety Monitoring. International Journal of Cyber-Security and Digital Forensics, 5(1), 46–60.

Vargas, J., Pomares, A., Alvarado, J., Quintero, J., & Palacio, J. (2017). Desarrollo de un Sistema de Segmentación y Perfilamiento Digital. Procesamiento de Lenguaje Natural, 59, 163–166.

Yao, F., Chang, K., & Campbell, R. (2015). Ushio: Analyzing News Media and Public Trends in Twitter.

Zadeh, L., Abbasov, A., & Shahbazoba, S. (2016). Analysis of Twitter Hashtags: Fuzzy Clustering Approach.

Zubiaga, A., Spina, D., Martínez, R., & Fresno, V. (2015). Real-time classification of Twitter trends. Journal of the Association for Information Science and Technology, 66(3), 462–473.