# MODELO DESCRIPTIVO PARA LA COMPRENSIÓN DE FENÓMENOS POLÍTICOS EN LAS REDES SOCIALES APLICANDO MINERÍA DE DATOS

**RAMSES ALEJANDRO BRITO PLANCHART**
**VICTOR HUGO MONTILVA DUARTE**

**UNIVERSIDAD EL BOSQUE**
**PROGRAMA DE INGENIERÍA DE SISTEMAS**
**FACULTAD DE INGENIERÍA**
**Bogotá, 2020**

# MODELO DESCRIPTIVO PARA LA COMPRENSIÓN DE FENÓMENOS POLÍTICOS EN LAS REDES SOCIALES APLICANDO MINERÍA DE DATOS

**RAMSES ALEJANDRO BRITO PLANCHART**
**VICTOR HUGO MONTILVA DUARTE**

**Desarrollo Tecnológico presentado como requisito para optar al título de INGENIERO DE SISTEMAS**

**Directora**
**SANDRA MERCHAN RUBIANO**
Ingeniera de Sistemas

**UNIVERSIDAD EL BOSQUE**
**PROGRAMA DE INGENIERÍA DE SISTEMAS**
**FACULTAD DE INGENIERÍA**
**Bogotá, 2020**

Dedicatoria

**Ramses:**

*Dedico este proyecto principalmente a nuestros seres queridos que no ganaron la lucha de este duro año transcurrido. En especial a mi papá Ramses Luis Brito Renzulli, quien estuvo presente y me apoyó todo el tiempo en lo largo de mi carrera hasta el último momento que vivió con nosotros. También le dedico este trabajo a mi madre quien me sigue acompañando y está presente para darme apoyo y fortalezas para seguir adelante cada día, junto a mis hermanos y tío Alcides, quien me dio un gran apoyo moral y económico en el proceso de mi carrera profesional.*
*A mis maestros por el aprendizaje que me brindaron y a mis amigos que estuvieron presentes en el camino.*

**Victor:**

*Esta tesis va dedicada a mis padres Manuel Montilva y Eva Duarte, quienes fueron motor de ayuda y ejemplo a seguir en el transcurso de mi ciclo universitario, a mi hermano Luis Montilva quien desde la distancia me envía su incondicional apoyo en cada reto asumido. Por último, dedicada a mis profesores quienes gracias a su vocación formaron en mí conocimientos que servirán de ayuda como futuro profesional e ingeniero de sistemas.*

## AGRADECIMIENTOS

En primera instancia agradecer a todos los profesores que formaron parte de mi ciclo de aprendizaje durante la carrera, quienes, gracias a su preparación profesional, nos brindaron el conocimiento suficiente para la investigación y realización de este proyecto.

De igual manera, queremos agradecer a nuestra directora de proyecto, Sandra Merchán Rubiano, quien fue guía desde el inicio para la propuesta de este proyecto. A través de su disposición para apoyarnos con constantes retroalimentaciones durante el proceso del trabajo.

También agradecer a mi compañero de proyecto por el apoyo y trabajo realizado desde el inicio de nuestras carreras universitarias.

**CONTENIDO**


1. PROBLEMA 2
1.1 DESCRIPCIÓN DEL CONTEXTO A INTERVENIR 2
1.2 ANÁLISIS DEL CONTEXTO DESDE EL MODELO BIOPSICOSOCIAL Y CULTURAL 2
1.3 IDENTIFICACIÓN Y DESCRIPCIÓN DE LA PROBLEMÁTICA 3
2. SOLUCIÓN DE INGENIERÍA 4
2.1 OBJETIVOS: GENERAL Y ESPECÍFICOS. 4
2.1.1 Objetivo General 4
2.1.2 Objetivos Específicos
2.2 DESCRIPCIÓN DE LA SOLUCIÓN Y RESULTADOS ESPERADOS 4
2.3 ANÁLISIS DE LA SOLUCIÓN DESDE EL MODELO BPSC 5
2.4 TABLA DE ENTREGABLES 6
2.5 VARIABLES POR MEDIR 6
2.6 METODOLOGÍA 7
2.6.1 Estructura de desglose de trabajo 8
2.6.2 Aplicación de la metodología 9
2.6.3 Cronograma 10
2.7 ACUERDO CON EL CLIENTE 11
2.8 COMPONENTE ÉTICO 12
3. MARCO REFERENCIAL 13
3.1 ANTECEDENTES Y ESTADO DEL ARTE 13
3.2 MARCO TEÓRICO 17
4. DESARROLLO METODOLÓGICO 18
4.1 PROCESO INVESTIGATIVO 18
4.2 FENÓMENO POLÍTICO 22
4.3 FASES DE METODOLOGÍA CRISP - DM 23
4.3.1 Fase 1 - Comprensión del negocio 23
4.3.1.1 Determinar objetivos del negocio 23
4.3.1.1.1 ¿Qué se desea hacer? 23
4.3.1.1.2 ¿Por qué usamos minería de datos para este propósito? 24
4.3.1.1.3 ¿Cuáles son los criterios de éxito? 24
4.3.1.2 Evaluar la situación 24

4.3.1.2.1 Desde el punto de vista político/social en redes sociales 24
4.3.1.2.2 Twitter 25
4.3.1.2.3 Análisis de tendencias en noticias y redes sociales 25
4.3.1.3 Elaborar plan de proyecto 26
4.3.2 Fase 2. Comprensión de los datos 27
4.3.2.1 Recopilación de los datos 27
4.3.2.2 Descripción de los datos 30
4.3.2.3 Revisión de los datos 32
4.3.2.4 Verificación de la calidad de los datos 33
4.3.3 Fase 3. Preparación de los datos 34
4.3.3.1 Selección de los datos 34
4.3.3.2 Construcción del conjunto de datos 35
4.3.3.3 Limpieza de los datos 37
4.3.3.4 Aplicación de formato de los datos 40
4.3.4 Fase 4. Modelado 41
4.3.4.1 Selección de la técnica de modelado 41
4.3.4.2 Construcción e implementación del modelo de pruebas 42
5. RESULTADOS 44
5.1 Análisis de sentimientos general para todos los datos 44
5.2 Análisis de sentimientos para los datos relacionados a las entidades 53
5.3 Evaluación del modelo 62
5.3.1 Validación de la academia 62
5.3.2 Validación estática 64
6. ANÁLISIS DE RESULTADOS/ DISCUSIÓN 68
7. CONCLUSIONES 71
8. RECOMENDACIONES Y TRABAJOS FUTUROS 73
8.1 Recomendaciones 73
8.2 Trabajos futuros 73
9. LECCIONES APRENDIDAS 74
REFERENCIAS BIBLIOGRÁFICAS 76
ANEXOS 79

## LISTA DE TABLAS

Tabla 1. Tabla de entregables 6
Tabla 2. Variable para la validación en la academia 6
Tabla 3. Variable para la validación estática 7

## LISTA DE FIGURAS


Figura 1. Tabla comparativa entre metodología 8
Figura 2. Fases metodología CRISP-DM 8
Figura 3. WBS 9
Figura 4. Cronograma 11
Figura 5. Matriz estado del arte 18
Figura 6. Evaluación de calidad -Tipo documento 18
Figura 7. Evaluación de calidad -Autoría 18
Figura 8. Análisis Web Of Science 19
Figura 9. Categorización de componentes 20
Figura 10. Fase 1 Comprensión del negocio 23
Figura 11. Fuentes de noticias 26
Figura 12. Fase 2 Comprensión de los Datos 27
Figura 13. Solicitud para el API de Twitter 28
Figura 14. Credenciales aprobadas 29
Figura 15. Extracción mediante Full archive search collection 30
Figura 16. Perfilamiento general de los datos 31
Figura 17. Perfilamiento de las variables text, created_at 31
Figura 18. Perfilamiento de la variable author_id 32
Figura 19. Parámetros y valores definidos para la extracción 33
Figura 20. Entidades mencionadas en tweets 33
Figura 21. Fase 3 Preparación de los Datos 34
Figura 22. Campos resultantes de la extracción 35
Figura 23. Archivo JSON con datos extraídos 35
Figura 24. Archivo convertido con extensión .xlsx de los datos extraídos 36
Figura 25. Creación de dataframe en Python utilizando Pandas 36
Figura 26. Eliminación de datos duplicados 37
Figura 27. Eliminación de etiquetas “RT” y @ en los tweets 37
Figura 28. Uso de regex para eliminar expresiones regulares en los textos 38
Figura 29. Eliminación de emojis 39
Figura 30. Eliminación de stopwords utilizando librería NLTK 40
Figura 31. Columna de datos procesados 40
Figura 32. Conversión de lista array en una sola cadena de texto 41
Figura 33. Fase 4 Modelado 41
Figura 34. Algoritmo para clasificar tweets por polaridad 43
Figura 35. Condiciones para asignar una clasificación al texto 43
Figura 36.Porcentajes de tweets por etiqueta en gráfica pieChart 44
Figura 37. Cantidad de tweets por etiqueta 45
Figura 38. Dataframe con columna de etiquetado 46
Figura 39. Datos estadísticos de la métrica compound 46
Figura 40. Creación del dataframe distinguido por etiquetas 47
Figura 41. Nube de palabras total de tweets 47
Figura 42. Nube de palabras de los tweets con etiquetado positivo 48

Figura 43. Nube de palabras de los tweets con etiquetado negativo 49
Figura 44. Nube de palabras de los tweets con etiquetado neutral 50
Figura 45. Longitud promedio de los tweets por cada etiquetado 51
Figura 46. Cantidad promedio de palabras en los tweets según etiqueta 51
Figura 47. N-gramas 2 palabras generado de los tweets 52
Figura 48. N-gramas 3 palabras generado de los tweets 53
Figura 49. Dataset de las identidades identificadas 54
Figura 50. Histograma cantidad de menciones de entidad por tweet 55
Figura 51. Creación dataframes por cada etiqueta de las entidades identificadas 56
Figura 52. Dataframe entidades mencionadas en tweets negativos 57
Figura 53. Dataframe entidades mencionadas en tweets positivos 58
Figura 54. Histograma cantidad de menciones de cada entidad en tweets neutrales 59
Figura 55. Histograma cantidad de menciones de cada entidad en tweets positivos 60
Figura 56. Histograma cantidad de menciones de cada entidad en tweets negativos 61
Figura 57. Modelo sintetizado correspondiente a los dos análisis de sentimientos 62
Figura 58. División del dataset para datos de entrenamientos y de pruebas 63
Figura 59. Total de clasificaciones correctas e incorrectas para tweets con entidades 63
Figura 60. Resultados evaluación del modelo 64

# GLOSARIO

**API:** (Application programming interface) es un conjunto de herramientas y/o procedimientos que pueden ser utilizados a su vez por otro software.

**Tweet:** Un mensaje publicado en Twitter que contiene texto, fotos, GIF o video.

**Hashtags:** Conjunto de caracteres precedidos por el signo numeral que sirve para identificar o etiquetar un mensaje en las webs

**Dataset:** Conjunto o colección de datos.

**Entidad:** Objeto de la vida real que tiene propiedades y atributos que la caracterizan.

**Polaridad:** Rasgo semántico de un elemento gramatical que exige un contexto afirmativo o negativo.

# RESUMEN

Modelo descriptivo para la comprensión de fenómenos políticos en las redes sociales aplicando minería de datos

Brito Planchart Ramses Alejandro, Montilva Duarte Victor Hugo

**Resumen**

El presente estudio se enfoca en la investigación, desarrollo y análisis del uso de la minería de datos en el campo político social de Colombia, dado que las redes sociales han sido una gran fuente de información abierta para el público durante estos últimos años.

El objetivo principal de esta tesis consiste en desarrollar un modelo descriptivo sobre un fenómeno político aplicando la minería de datos a una red social, que permita comprender los actores, flujos de información e interacciones que rigen el comportamiento del fenómeno desde la perspectiva biopsicosocial y cultural. Para esto se aplicó un modelamiento de análisis sentimental sobre las opiniones expresadas en Twitter por parte de los ciudadanos colombianos con respecto al fenómeno de inmigración venezolana que ocurre actualmente en el país.

La metodología de este trabajo se basa principalmente en las etapas del proceso conocido como CRISP-DM la cual es utilizada para el desarrollo de proyectos de minería de datos. Para comenzar, se identificó un problema de investigación, el cual fue el fenómeno migratorio de los venezolanos en el país. Luego, mediante la red social seleccionada que fue Twitter se extrajeron, seleccionaron y procesaron los datos para el estudio. Para esto se hizo uso del API Tweepy de Twitter, que permitió la recolección de tweets, los cuales serán usados como los datos de este proyecto.

Posteriormente al procesamiento de los datos, se realizó un análisis de opiniones para medir la variable del problema de investigación del proyecto, que consta de la aceptación y/o rechazo de la población colombiana a partir de las expresiones obtenidas mediante la red social Twitter relacionadas con el fenómeno de la inmigración venezolana.

Palabras clave: minería, datos, extracción, fenómeno, redes sociales, política, modelo, opiniones, tweet.

**Abstract**

Keywords: *Mining, Process, Models, Data, Relation, Social networks, Politic*

---

This research focuses on the study, development, and analysis of the use of data mining in the social-political field of Colombia since social networks have been a great source of open information to the public in recent years.

The main objective of this thesis is to develop a descriptive model of a political phenomenon applying data mining to a social network, which allows us to understand the actors, information flows, and interactions that govern the behavior of the phenomenon from the biopsychosocial and cultural perspective. A sentimental analysis modeling was applied to the opinions expressed on Twitter by Colombian citizens regarding the Venezuelan immigration phenomenon that is currently happening in the country.

The methodology selected for this work was CRISP-DM, given the stages of this process, which is used for the development of data mining projects. Firstly, a research problem was identified, which was the migratory phenomenon of Venezuelans in the country. Then, through the selected social network that was Twitter, the data for the study were extracted, chosen, and processed. For this, the Twitter API Tweepy was used, which reached the collection of tweets, which will be used as the data for this project.

After processing the data, an opinion analysis was carried out to measure the variable of the project's research problem, which consists of the acceptance or rejection of the Colombian population from the expressions obtained through the social network twitter related to the phenomenon of Venezuelan immigration.

# INTRODUCCIÓN

Este proyecto de desarrollo tecnológico se centra en la investigación del uso de minería de datos en redes sociales y cómo se hace uso de estos datos abiertos para fines políticos y sociales, a partir del análisis y las ciencias de datos que involucra los procesos para recopilar información y obtener un mejor entendimiento de los datos.

Esta investigación surge a partir del impacto actual de las redes sociales y los datos dentro de la vida humana, debido al valor que esta información ha adquirido durante los últimos años. De igual manera, se ha centrado tanta atención en este tema, debido a los grandes beneficios y riesgos que los datos han marcado en esta era digital. La tecnología ha permitido la creación y almacenamiento de grandes cantidades de información, esto ha sido aprovechada por grandes empresas para convertir los datos en una ventaja competitiva al perfeccionar los productos y servicios. Esta también ha sido aprovechada en el campo político, a través del análisis de datos para estudiar el comportamiento de usuarios en redes sociales, y analizar estos datos para distintos fines políticos.

La gran cantidad de datos recopilados y almacenados, pueden generar grandes beneficios para las organizaciones y las sociedades de todo el mundo. Sin embargo, este valioso recurso puede ser obtenido sencillamente a través de medios como las redes sociales, y pueden ser utilizados para beneficios propios, sin muchas veces, el consentimiento del propietario de la información. Además, esto puede ser una ventaja para grupos políticos, quienes a través de grandes empresas facilitan el acceso y uso de estos datos, para la manipulación e intervención en campañas políticas para influir en la decisión de un proceso en un grupo de personas,

Durante los últimos años, han ocurrido diferentes casos polémicos en cuanto a este tema, así como el escándalo de Cambridge Analytica en 2018 por acceder ilegalmente a los datos de millones de usuarios de Facebook para tratar de influenciar en los resultados de las elecciones presidenciales de 2016 en Estados Unidos.

El propósito del siguiente trabajo tiene como fin el análisis y desarrollo de un modelo descriptivo aplicando minería de datos a una red social para comprender un fenómeno político actual desde la perspectiva biopsicosocial y cultural. Para el desarrollo de este proyecto de minería de datos, se emplea la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para establecer una eficaz planeación del proyecto a través de distintas fases para llevar a cabo un mejor entendimiento del problema y un organizado análisis de los datos tratados.

# 1. PROBLEMA

En el presente capítulo se realiza una descripción general del contexto en el que se desarrolla el trabajo con el fin de orientar al lector sobre el tema de estudio. En este apartado se identifican los componentes presentes en el contexto desde la perspectiva biopsicosocial y con base en esto se identifica y se describe la problemática encontrada.

## 1.1 DESCRIPCIÓN DEL CONTEXTO A INTERVENIR

Las redes sociales han sido una gran fuente de información abierta para el público durante estos últimos años. Esto ha sido aprovechado por el sector político para recopilar datos de los usuarios y analizarlos para identificar y establecer patrones que facilitan el desarrollo de campañas políticas.

Es a partir de este contexto en donde los partidos políticos encuentran la oportunidad de beneficiarse de esta información para la manipulación de datos personales a través de las redes sociales. El objetivo es agrupar los datos obtenidos para predecir comportamientos y posiciones políticas de los usuarios con el fin de ejecutar estrategias políticas y publicitarias.

## 1.2 ANÁLISIS DEL CONTEXTO DESDE EL MODELO BIOPSICOSOCIAL Y CULTURAL

Desde la perspectiva del modelo BPSC, la problemática identificada nace del uso de las redes sociales y como los datos alojados en ella son utilizados con diversos fines políticos; Campañas, Publicidad, Comunicación y Estrategia. Los anteriores ejercicios sin medir el nivel ético de los resultados que puedan causar en una sociedad.

Para el modelo biopsicosocial plasmado en este proyecto, se modelaron dos diagramas identificando el problema, uno desde la perspectiva de los usuarios de las redes sociales y el segundo desde el punto de vista de los partidos políticos quienes utilizan estos medios sociales como herramientas para elaborar estrategias políticas.

Desde este contexto expuesto se identifican las redes sociales como el principal artefacto de uso a través de los usuarios, quienes relacionan el uso de estas con creencias como: El uso de datos personales, manipulación de datos personales, falta de privacidad, tráfico de los datos personales para ser utilizados por las campañas políticas identificado como el medio de este modelo BPSC. Lo que conlleva a generar hábitos como: reducir la información compartida en redes sociales, reducir o eliminar los datos compartidos, y hasta cerrar los perfiles de las redes sociales.

Por otro lado, el segundo actor identificado en el contexto de la problemática son los partidos o cuerpos políticos quienes utilizan el artefacto como medio de

persuasión generando creencias tales como: mejorar su imagen política, predecir posición política de los usuarios, generar publicidad efectiva, entre otras. Lo que conlleva a generar hábitos como: Extraer y analizar datos desde las redes sociales, manipulando y agrupándolos para predecir y ejecutar estrategias políticas y de publicidad. Esto se puede observar detalladamente en el anexo A.

## 1.3 IDENTIFICACIÓN Y DESCRIPCIÓN DE LA PROBLEMÁTICA

La minería de datos recorre una gran variedad de campos y la política es una de las áreas donde en mayor medida afecta a la población general. Hoy en día se conoce que gran parte de los políticos utilizan este avance tecnológico para diseñar sus estrategias en campañas políticas. Esta tecnología se ha convertido en una herramienta de gran uso durante las elecciones más recientes en la actualidad, cambiando el proceso tradicional de campañas electorales, siendo un recurso principal que influyen en estas a través de la recolección y análisis de datos. Uno de los usos de la minería de datos en campañas políticas es la versatilidad de los datos que son utilizados, para medir el comportamiento de los votantes y obtener un resultado actual de los candidatos.

Estos datos son también utilizados para realizar estrategias de publicidad y promociones en diferentes medios como la televisión y redes sociales, facilitando la identificación de amplios grupos de personas para la inyección de información y material personalizado para conseguir votos esperados. Gracias a la enorme cantidad de datos disponibles, el uso de la minería de datos se convierte imprescindible para la recolección de información útil para la formulación de estrategias políticas.

Existen grandes empresas que se encargan de recopilar información personal que es autorizada por personas al aceptar términos y condiciones en distintos sitios web, donde muchas veces sin darnos cuenta autorizamos el uso de nuestros datos; para luego ser vendidos a corporaciones o partido políticos que se aprovechan de esta información para diseñar sus estrategias en campañas electorales en función del público. Estas empresas expertas en información de datos ofrecen el análisis y la clasificación de los usuarios procediendo de los datos personales como ubicación de residencia, edad, entre otros datos que son utilizados para la elaboración del material de campañas.

# 2. SOLUCIÓN DE INGENIERÍA

Para el planteamiento de los objetivos se identificó un objetivo general que pretende entregar como artefacto un modelo descriptivo para el grupo de investigaciones OSIRIS de la Universidad El Bosque, que será posible generar a partir del desarrollo de los objetivos específicos planteados en esta sección.

## 2.1 OBJETIVOS: GENERAL Y ESPECÍFICOS.

### 2.1.1 Objetivo General

Desarrollar un modelo descriptivo sobre un fenómeno político aplicando la minería de datos a una red social, que permita comprender los actores, flujos de información e interacciones que rigen el comportamiento del fenómeno desde la perspectiva biopsicosocial y cultural.

### 2.1.2 Objetivos Específicos

- Elaborar un estado del arte sobre la aplicación de la minería de datos en la comprensión de fenómenos políticos y sociales, usando los datos almacenados en redes sociales.
- Identificar un fenómeno político que sea susceptible de analizar a través de la minería de datos para proponer un experimento de ingeniería que conlleve a su modelamiento y comprensión.
- Desarrollar un proceso de minería de datos aplicando la metodología CRISP-DM a partir de datos extraídos de una red social, generando un modelo de comportamiento del fenómeno elegido.
- Evaluar si el modelo de comportamiento generado a través de CRISP-DM permite comprender el fenómeno político desde la perspectiva biopsicosocial y cultural.

## 2.2 DESCRIPCIÓN DE LA SOLUCIÓN Y RESULTADOS ESPERADOS

Los modelos de minería de datos más que algoritmos aplicados a datos, son conjuntos de datos, patrones y componentes estadísticos que pueden ser aplicados nuevos datos con el fin de generar predicciones o determinar relaciones que a simple vista no podríamos identificar. Estos modelos son aplicados en diferentes áreas sociales, una de estas la política, la cual comprende una de las áreas de mayor importancia. Por ello los políticos recurren al uso de nuevas tecnologías y técnicas que muchas veces implica el uso de redes sociales y los datos alojados en ella, por lo tanto, la solución propuesta consiste en:

Un modelo de tipo descriptivo aplicando minería de datos que permita comprender los actores, flujos de información e interacciones que rigen el comportamiento de un fenómeno político desde la perspectiva biopsicosocial y cultural.

Con el modelo descriptivo entregado al grupo de investigación OSIRIS de la Universidad el Bosque, se espera la apertura de nuevas investigaciones apoyadas en modelos biopsicosociales y culturales para el planteamiento de problemas y soluciones para comprender fenómenos políticos.
4
Se espera a partir de este proyecto, el adecuado desarrollo de un modelo descriptivo construido a partir de datos abiertos recopilados a través de las redes sociales. Este desarrollo se fundamenta en el uso de la metodología CRISP - DM y de herramientas de minería de datos como técnicas de clustering y árboles de decisión que permitan la identificación de patrones y comportamiento que ocurren dentro del fenómeno estudiado. Para evaluar este modelo se necesitará corroborar la veracidad de los datos, y se utilizarán indicadores de precisión y eficiencia para medir el rendimiento del algoritmo usado.

A partir del análisis planteado, se encuentra la oportunidad de solucionar esta problemática para comprender los actores, flujos de información e interacciones que rigen el comportamiento del fenómeno desde la perspectiva biopsicosocial y cultural. Esto permitirá la apertura de nuevas investigaciones apoyadas en un modelo biopsicosocial y cultural para el planteamiento de problemas y soluciones por parte del grupo de investigaciones OSIRIS & BIOAXIS de la Universidad El Bosque, con el fin de generar interés por el estudio de estos fenómenos políticos y aportar este conocimiento a la sociedad.

## 2.3 ANÁLISIS DE LA SOLUCIÓN DESDE EL MODELO BPSC

En la solución analizada desde la perspectiva del modelo biopsicosocial y cultural ofrecida al beneficiario (Grupo de investigación OSIRIS&BIOAXIS de la Universidad el Bosque), se identifica como artefacto el desarrollo de un modelo descriptivo aplicando la minería de datos que permita comprender los actores, flujos de información e interacciones que rigen el comportamiento del fenómeno desde las perspectiva biopsicosocial y cultural. Este artefacto pretende medir el impacto de las redes sociales que genera creencias en el grupo de investigación tales como: el uso de modelos descriptivos para entender fenómenos políticos, como recurso educativo para futuros proyectos y como base de propuesta para futuros estudios.
Esto puede generar hábitos en el grupo de investigación como la creación de modelos biopsicosociales para el planteamiento de problemas, se abren nuevas investigaciones apoyadas en un modelo biopsicosocial y cultural que permite comprender un fenómeno político, entre otros; cómo se puede ver en el anexo B.

## 2.4 TABLA DE ENTREGABLES

**Tabla 1.** Entregables.

| Entregable | Descripción |
|---|---|
| Tabla de datos | Se entregará al beneficiario un documento que contendrá adjunto la tabla de los datos recopilados para el estudio |
| Modelo descriptivo | Modelo descriptivo que permita la comprensión del fenómeno estudiado desde la perspectiva biopsicosocial |
| Informe final | Se entregará un informe final que documentará los resultados y aprendizajes obtenidos en el proceso de estudio y desarrollo del proyecto. |

Fuente: Elaboración propia.

## 2.5 VARIABLES POR MEDIR

En el estudio se determinaron diferentes variables teniendo en cuenta la validación de la academia y la validación estática. Las variables para presentar en esta sección se clasifican de acuerdo con las etapas en las que se van a medir. La tabla número 2, muestra las variables que se van a medir para realizar la validación de la academia, en donde se evaluará la precisión del modelo implementado. En la tabla número 3, se presenta la variable que se medirá para la validación estática con respecto al cliente, evaluando la utilidad del modelo mediante una entrevista.

**Tabla 2.** Variable para la validación de la academia.

| Variable | Tipo de Prueba | Indicadores | ¿Cómo se va a medir? | Instrumentos |
|---|---|---|---|---|
| Precisión del modelo | Prueba de precisión | Cantidad de falsos positivos | Se realiza la aplicación del algoritmo Naive Bayes sobre el modelo realizado con el fin de medir | Collab, Python |

| | | | | |
|---|---|---|---|---|
| | | | la exactitud de un clasificador | |

Fuente: Elaboración propia.

**Tabla 3.** Variable para la validación estática.

| Variable | Indicadores | ¿Cómo se va a medir? | Instrumentos | Población objetivo |
|---|---|---|---|---|
| Utilidad del modelo | Opiniones positivas-Opiniones negativas | A Partir del nivel de satisfacción expresado en opiniones por el grupo de investigación con respecto al producto entregado | Entrevista | Grupo de Investigación Osiris&Bioaxis |

Fuente: Elaboración propia.

## 2.6 METODOLOGÍA

Para el desarrollo ingenieril se emplea la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) para la organización del proyecto enfocada al análisis de los datos. Se escogió CRISP DM por su fácil aplicación a proyectos de minería de datos y su correcta estructura dividida por fases que permite una mejor compresión de la metodología. CRISP-DM a comparación de otras nos da la posibilidad de enfocarnos en entender la problemática o fenómeno a estudiar de manera efectiva y previa al estudio y comprensión de los datos.

CRISP-DM resalta sobre otras metodologías por la posibilidad crear un modelo acorde a las necesidades específicas del cliente, tiene una correcta manera de manejar errores en el desarrollo pues se nos permite regresar entre fases para realizar ajustes de ser necesario. En general es una de las metodologías más utilizadas en comparación a las demás existentes. Esta metodología incluye las tareas necesarias en cada fase y una explicación entre las relaciones entre las tareas. En la Figura 1 se muestra una comparativa entre metodologías.

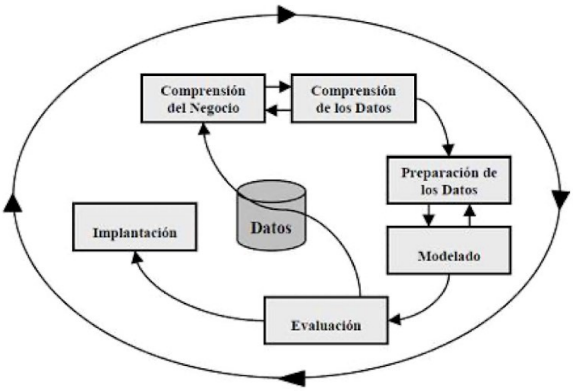
**Figura 1.** Tabla comparativa entre metodología.

| CRISP-DM | KDD | SEMMA | TDSP | Guerrilla |
|---|---|---|---|---|
| Compresión del negocio | | | Compresión del negocio | |
| Compresión de los datos | Selección | Sample | Data acquisition and understanding | Extract |
| | | Explore | | Receive |
| Preparación de los datos | Pre-procesado | Manipulate | | Load |
| | Transformación | | | |
| Modelado | Data mining | Model | Modeling | Analytics |
| Evaluación | Interpretación/ evaluación | Asses | | Consolidate |
| Explotación | | | Deployment and customer acceptance | Work products |
| | | | | Reporting |

Fuente: C. I. García Osorio, “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM,” Univ. Burgos. Mayo, 2019, doi: 10.13140/RG.2.2.34208.02566.

Ver en la figura 2 las fases de la metodología CRISP-D

**Figura 2**: Fases metodología CRISP-DM

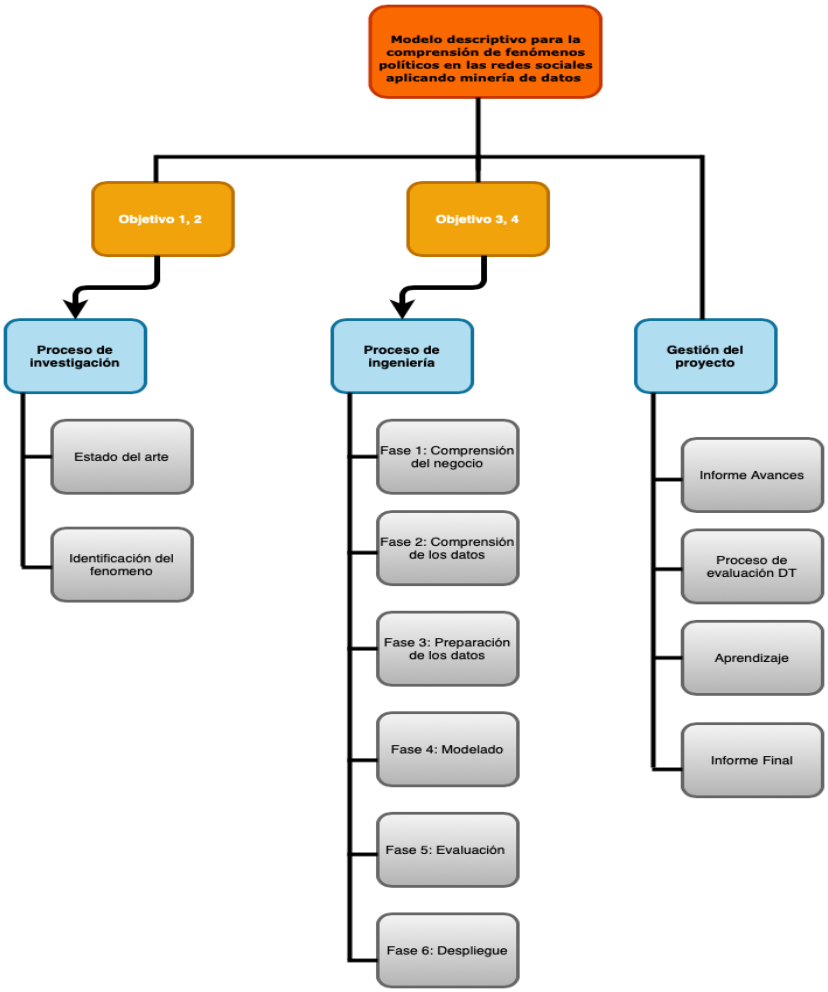



Nombre de la fuente: C. Pete et al., “Crisp-Dm 1.0,” Cris. Consort., p. 76, 2000.

### 2.6.1 Estructura de desglose de trabajo

En la figura 3 se muestra la estructura del trabajo ordenada por paquetes.

**Figura 3.** WBS




Fuente: Elaboración propia.

### 2.6.2 Aplicación de la metodología

En la elaboración de este proyecto, se utilizó el software de gestión de proyectos JIRA para la mayor eficiencia gerencial en el proyecto a la hora de organizar el tablero de control y hoja de ruta con el cronograma para las tareas pendientes, las que están en curso y las finalizadas. Se utilizó un tablero Kanban que proporciona JIRA para la administración de las tareas que fueron organizadas en 8 épicas de trabajos que contienen las respectivas actividades de cada fase de trabajo, esto permite un desarrollo simultáneo de las tareas.

Se eligió la herramienta Jira como gestor del proyecto debido principalmente a su funcionamiento sencillo y fácil de entender. También facilita la colaboración entre los miembros del equipo para la planificación y seguimiento del trabajo, permite crear tickets correspondientes a las historias de usuarios e incidencias identificadas, permite asignar un responsable a cada ticket o tarea creada, aporta información actualizada en tiempo real para mejorar el rendimiento del equipo y por último el uso

es gratuito. En el anexo A se puede detallar el tablero de control que fue utilizado gracias a esta herramienta.

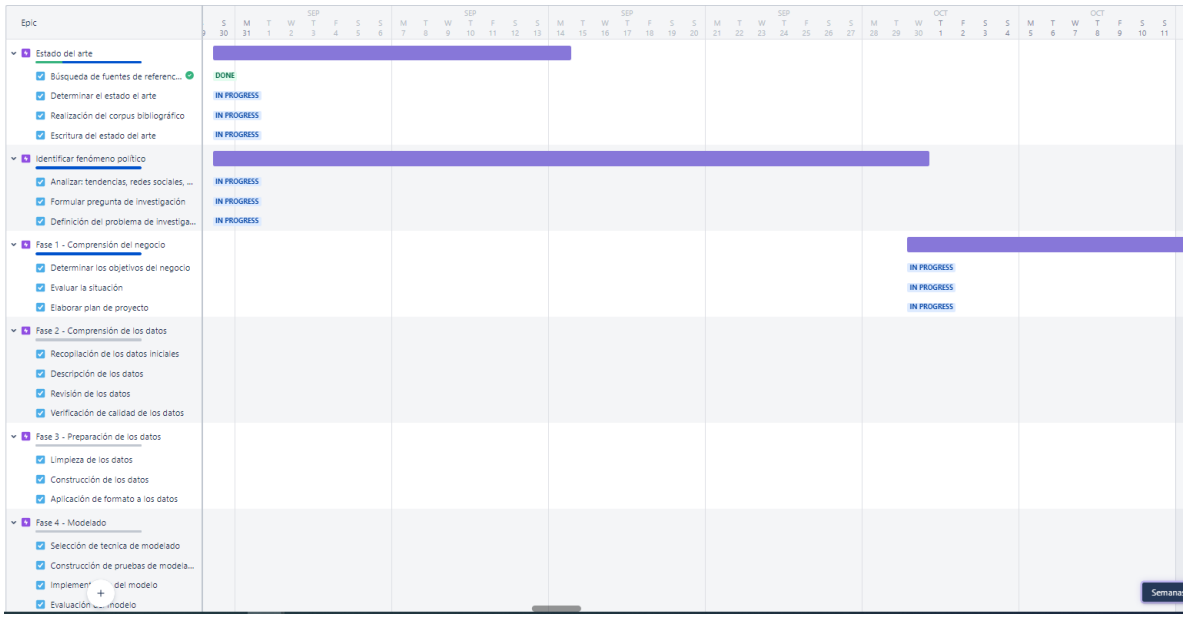
Durante el desarrollo de la gestión del proyecto se sigue un esquema de trabajo dividido en tres paquetes (Proceso de investigación, proceso de ingeniería y gestión de proyecto) los cuales contienen tareas correspondientes a cada etapa. En primer lugar, se ejecutaron las tareas correspondientes al paquete de proceso de investigación donde se formuló una ecuación de búsqueda como se observa en el anexo C que permitió una correcta indagación sobre trabajos relevantes y relacionados al tema de estudio, en este primer paquete de investigación también se busca con base al estado del arte identificar el fenómeno que se quiere comprender.

Correspondiente al segundo paquete del proceso de ingeniería se siguen seis fases propuestas por la metodología CRISP-DM; iniciando desde la comprensión del negocio, donde se establece: ¿que se desea realizar?, ¿Por qué utilizar minería de datos? e identificar los criterios de éxito; luego se ejecutan fases como comprensión y preparación de datos, modelado, evaluación y terminando en la fase de despliegue. Para un seguimiento en el desarrollo del proyecto y en cada uno de los paquetes anteriores se tomarán en cuenta los resultados de las tareas correspondientes al paquete de gestión de proyecto tales como: informes de avance, fase evaluación por parte del jurado, fase de aprendizaje y una fase informe final con el fin de documentar el proceso y el conocimiento adquirido en el desarrollo.

### 2.6.3 Cronograma

En la figura 2-4 se visualiza una muestra de la primera parte del cronograma, el cual fue construido gracias al gestor de proyectos Jira. En el Anexo C se encuentra el cronograma completo dividido por meses, desde la fecha de inicio 30 de agosto del 2020 hasta la fecha final el 12 de mayo de 2021.

**Figura 4.** Cronograma.



Fuente: Elaboración propia.

## 2.7 ACUERDO CON EL CLIENTE

En esta sección se evidencia como los integrantes del proyecto, llegaron a un acuerdo con el cliente del grupo de investigación OSIRIS&Bioaxis perteneciente a la Universidad El Bosque, en donde se establece que la titularidad de los derechos morales pertenece a nosotros en su totalidad y serán derechos irrenunciables. Los derechos patrimoniales serán cedidos al grupo de investigación OSIRIS&Bioaxis perteneciente a la Universidad El Bosque bajo los parámetros y términos de la ley y normatividad vigente.

En el Anexo F se puede observar detalladamente la carta de acuerdo entre el cliente y beneficiario. Y en el anexo G, se muestra la carta de autorización de uso y disposición del desarrollo tecnológico.

## 2.8 COMPONENTE ÉTICO

Ramses Alejandro Brito Planchart estudiante de Ingeniería de sistemas en su calidad de autor del presente trabajo y Víctor Hugo Montilva Duarte Estudiante de Ingeniería de sistemas en calidad de autor nos comprometemos en cumplir con los principios a continuación mencionados y establecidos en el código de ética de Ingeniería de Software:

**Principio 1. Sociedad:**

Nos comprometemos a aceptar la responsabilidad por el trabajo realizado, así como también ser justo y veraz en todas las afirmaciones referentes al software producido y los resultados mostrados por el mismo

**Principio 2 Cliente y empresario:**

Nos comprometemos a aplicar conocimientos solo en el área de nuestra competencia haciendo uso solo de herramientas obtenidas de manera legal y estudiadas previamente asegurando la calidad de estas.

**Principio 3. Producto:**

Garantizamos los planteamientos de objetivos adecuados y alcanzables dentro de este proyecto, así como también la identificación de la metodología de desarrollo más adecuada. Nos comprometemos a esforzarnos por entender el producto que se está desarrollando y que las especificaciones y características satisfagan las necesidades planteadas por el cliente, manteniendo siempre la integridad y veracidad de los datos utilizados.

# 3. MARCO REFERENCIAL

En esta sección se exponen las investigaciones realizadas para sustentar el desarrollo metodológico de este proyecto, a través de la indagación de antecedentes necesarios para la elaboración del estado del arte. Del mismo modo, se elabora un marco teórico con el desarrollo de la teoría con la que se fundamenta el planteamiento del problema y la metodología del desarrollo.

## 3.1 ANTECEDENTES Y ESTADO DEL ARTE

Para la elaboración del estado del arte se formuló la siguiente ecuación de búsqueda para la investigación de otros estudios relevantes similares que apoyen el proyecto.

("political phenomena" OR "political campaign" OR "presidential campaigns" OR "campaigns" OR "political strategies" OR "political elections" OR "Political Marketing") AND ("personal data" OR "personal information" OR "social networks data" OR "mass communication psychology") AND ("data mining" OR "machine learning" OR statistics).

En el Anexo C, se puede observar detalladamente los núcleos temáticos que se tuvieron en cuenta para la construcción de esta ecuación de búsqueda.

Esta fórmula se ejecutó en distintos exploradores de bases de datos académicos y profesionales sobre ingeniería y tecnología, tales como IEEE, ACM, Researchgate, Scielo, entre otros. De igual manera se realizaron búsquedas en repositorios de universidades tanto nacionales como internacionales, de los cuales se obtuvieron resultados de papers, investigaciones y tesis realizadas previamente.

En la primera sección del estado del arte, se explican las diferentes fuentes utilizadas para la recopilación de los datos a partir de las redes sociales que se han usado en distintos trabajos previos. Luego se identifican las metodologías más usadas en los proyectos de minería de datos con el fin de establecer la metodología más conveniente para este trabajo. En tercer lugar, se determinan las técnicas más eficientes y adaptables para el desarrollo del modelo descriptivo. Por último, se analizan los diferentes campos políticos que fueron estudiados para la aplicación de minería de datos.

### 3.1.1 Fuente de los datos

Luego de la investigación bibliográfica realizada se comprobó que la red social más utilizada como fuente de recopilación de datos es Twitter. Sin embargo, grandes casos como el de Cambridge analítica, utilizaron los datos de 50 millones de perfiles de usuarios de obtenidos de Facebook, analizando la información personal e infiriendo los perfiles psicológicos de cada usuario para saber cuál debía ser el

contenido inyectado en sus redes sociales para cambiar la forma de pensar de los votantes en las elecciones de Estados Unidos de 2016 **[6]**.

Otro documento como *Social media and microtargeting: Political data processing and the consequences for Germany* **[9]** que tiene como objetivo explicar cómo los políticos en Alemania crean condiciones de Microtargeting para influir sobre los votantes a través de estímulos que son formados en función de las preferencias y características del individuo a partir de los datos de Facebook. Esta red social fue elegida ya que la población alemana en Facebook es más grande y menos selectiva que la de Twitter.

Sin embargo, la gran mayoría de investigaciones y proyectos realizados previamente, prefieren recopilar datos de la red social Twitter debido a la mayor cantidad de opiniones compartidas por tweets y al proceso de recolección que brindan diferentes APIS de Twitter que permiten búsquedas y recolecciones semi automáticas de tweets, tal como se utilizó en *Modelling political disaffection from Twitter data* **[20]**. De igual manera Twitter fue utilizada en la Tesis *Twitter, la nueva era de la comunicación política: un análisis del manejo de la red social en las elecciones a la alcaldía de Bogotá 2015* **[15]** para analizar la importancia del uso de esta red social a la hora de tomar la decisión de un alcalde político, específicamente la alcaldía de Bogotá, Colombia para el año 2015.

En la mayoría de los trabajos consultados dentro del corpus bibliográfico de este proyecto, se pueden observar la preferencia de Twitter como la principal red social usada para la recolección de los datos que son utilizados para el estudio de diferentes fenómenos políticos. Esto es evidenciado en la matriz de estado de arte que fue realizada previamente a la escritura de este estado del arte. Es a partir de estos resultados que se decidió utilizar Twitter como fuente de datos en este proyecto.

### 3.1.2 Metodologías

En los trabajos analizados se determinaron las dos metodologías con mayor recurrencia de uso en proyectos de minería de datos; Entre estas metodologías se encuentra primero CRISP - DM, una metodología de fácil aplicación compuesta de una serie de fases, algunas de ellas bidireccionales. El proyecto *Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales* **[14]** dirigido a la alcaldía de Bogotá estuvo estructurado metodológicamente siguiendo las fases iterativas propuestas por CRISP-DM, el proyecto tuvo como objetivo principal determinar la relación entre el sentimiento político mostrado en las publicaciones de redes sociales como Twitter y en las encuestas de opinión pública.

En los proyectos de minería de datos CRISP - DM ha sido la metodología con más apoyo por empresas privadas, públicas e investigadores, convirtiéndose así en la

metodología más empleada gracias a su entendible guía y sucesión de fases como lo dice Galán Víctor en su tesis: *Aplicación de la metodología crisp-dm a un proyecto de minería de datos en el entorno universitario* **[16]** que tenía por objetivo la implementación y desarrollo de las seis fases propuestas por CRISP- DM en un proyecto de minería de datos relacionado con el rendimiento académico en la universidad Carlos III de Madrid.

Otra metodología que tuvo lugar en el análisis bibliográfico fue KDD, metodología comprendida por cinco frases y utilizada en los proyectos de minería de datos con el fin de identificar patrones válidos, nuevos y útiles, sin embargo de una extensa lista de trabajos revisados fue solo Corzo Daniel en su tesis de pregrado: *Modelo conceptual para determinar el próximo presidente de estado a través de datamining y Twitter* **[30]** *que estableció KDD como metodología principal de desarrollo,* pero entre las desventajas identificadas por el uso de esta metodología se identifica la falta de bidireccionalidad es decir se debe cumplir una secuencia específica que no permite retroceder entre tareas lo que puede llegar a complicar el proceso de existir errores.

Esta adecuada investigación permitió seleccionar la metodología CRISP - DM como base fundamental para el desarrollo del modelo entregado en este proyecto, gracias a su mejor adaptación de acuerdo con los requerimientos planteados en el planteamiento del problema y de la solución.

### 3.1.3 Herramientas y técnicas para minería de datos

Dentro del campo de la minería de datos se identifican dos grandes grupos: La minería de datos dirigida o supervisada y la minería de datos no dirigida. Las técnicas dirigidas también conocidas como supervisadas se caracterizan por predecir el valor de un atributo específico dentro de un conjunto de datos **[17]** donde el procedimiento consiste en realizar predicciones sobre los datos partiendo de una relación entre el valor desconocido y valores conocidos. Por su parte, las técnicas no dirigidas se caracterizan por el uso de algoritmos para identificar patrones o tendencia de los datos estudiados.

El grupo de interés para este proyecto es el de minería de datos no dirigidas donde se encuentra que las técnicas más utilizadas son: En primer lugar, detección automática de Cluster cuyo objetivo es formar grupos donde cada dato dentro del grupo sean lo más parecidos entre sí y lo más diferentes a otros grupos **[17]**. En segundo lugar, reglas de asociación que funcionan para la detección de patrones de comportamientos y asociaciones entre los datos y por último detección de desviaciones que consiste en identificar cambios significativos dentro del conjunto de datos estudiados con respecto a estados o valores pasados **[31]**.

Las herramientas utilizadas para realizar estudios de minería de datos se clasifican en herramientas libres y licenciados, la diferencia entre estos grupos radica en que

el acceso a las técnicas y recursos en softwares licenciados son pagos, por su parte los softwares o herramientas libres son totalmente gratuitos por ello radica el interés en este grupo, entre los software libres destacan: Weka, un software con entorno gráfico que conforma un grupo completo de técnicas, algoritmos y esquemas que permiten el procesamiento y analizado de datos**[17]**. RapidMiner es otro entorno libre que permite el análisis de información por medio del encadenamiento de operadores y el uso de técnicas dentro del entorno, RapidMiner es considerada líder dentro de los softwares libres de minería de datos, sin embargo suele ser más compleja de manejar que Weka; por último encontramos el lenguaje de programación Python y sus diferentes librerías y modelos para la manipulación de datos que permiten desde la extracción hasta el análisis de información alojada en diferentes bases de datos, páginas web, redes sociales, entre otras .

Como resultado de la investigación, la experiencia previa y la libre licencia se determinó el uso de Python y sus librerías como la herramienta principal para la realización del proceso de minería de datos en el desarrollo del proyecto.

### 3.1.4 Campos políticos

Durante la investigación de fuentes de referencias de este proyecto, se identificaron distintos fenómenos políticos estudiados para aplicar minería de datos con diferentes fines. Para facilitar la identificación de estos campos políticos, se clasificaron los documentos hallados por regiones geográficas, agrupados en categorías tales como: Antecedentes internacionales, Antecedentes Nacionales (Colombia) y antecedentes locales (Bogotá).

En los antecedentes locales se puede identificar la Tesis anteriormente mencionada en el apartado de las fuentes de los datos: *Twitter, la nueva era de la comunicación política: un análisis del manejo de la red social en las elecciones a la alcaldía de Bogotá 2015* **[15]** del cual se determinar la influencia de las redes sociales como herramienta de estrategia política dentro de las campañas de los candidatos a la alcaldía de ese entonces, Clara López Obregón, Enrique Peñalosa, Rafael Pardo y Francisco Santos.

En cuanto a los antecedentes locales, se identifica otro campo político similar en una tesis de pregrado en Cartagena titulada como *Uso de la red social "Facebook" en la política cartagenera, basado en el Modelo de la recepción selectiva. Caso del alcalde Dionisio Vélez, periodo 2013-2015* **[27]** en donde se busca diagnosticar el impacto de la red social Facebook en la política cartagenera durante las elecciones gubernamentales del alcalde de Cartagena de Indias. Dionisio Vélez. Ambos campos políticos están basados en el estudio de elecciones de alcaldías regionales en Colombia. Estas investigaciones aportan a este proyecto en cuanto a la toma de decisión sobre el fenómeno político que se identificará de manera local o nacional para el desarrollo del modelo descriptivo.

## 3.2 MARCO TEÓRICO

El uso de las redes sociales en el ámbito político y social durante los últimos años ha incrementado rápido. Un estudio realizado por la universidad de Oxford en el año 2019 con una muestra de 70 países determinó que al menos un partido o entidad política de cada país utiliza estos medios para fines de manipulación **[5]**. Los políticos aprovechan el uso de datos para realizar modelos que les permitan determinar qué contenido mostrar o de qué temas hablar, para de esta manera lograr un cambio objetivo en la mentalidad de los usuarios y votantes **[6]** así es como las redes sociales han mostrado gran efectividad como medio de estrategia política, comunicación política y construcción de una mejor imagen **[22]**.

La minería de datos es una tecnología que permite la recolección y agrupación de datos; es un campo multidisciplinario de la informática utilizado en diversas áreas incluyendo la política en donde se lleva a cabo un proceso con el objetivo de descubrir, analizar y extraer patrones en grandes conjuntos de datos **[7]**.El término es aplicado a cualquier tipo de información a gran escala y alojada en cualquier fuente accesible, por ende los políticos muchas veces recurren sin consentimiento informado a los datos alojados en redes sociales como Facebook o Twitter puesto que son los datos con mayor facilidad de acceso y vulnerabilidad de manipulación**[13]**, con base en estos datos logran establecer estrategias y modelos que pueden llegar persuadir o cambiar el pensamiento de los votantes o población objetiva **[6].**

Los modelos descriptivos permiten describir y/o comprender un problema o tendencia social. El principal objetivo del modelo es recolectar y ordenar información a través de medios visuales para describir las relaciones existentes entre los datos, identificando patrones y las características más representativas de la colección de datos para clasificar la información analizada con base en unas variables preestablecidas **[17]**. De esta manera se obtienen datos precisos para evaluar las características de un fenómeno en particular **[20]**.

Los datos extraídos a partir de las redes sociales se han visto manipulados para su uso en actividades políticas. Un claro ejemplo de este fenómeno es el caso de Facebook quienes vendieron los datos de 87 millones de usuarios a la empresa Cambridge Analytica para ser utilizados con el objetivo de influenciar las elecciones presidenciales de Estados Unidos en 2016 y el referéndum del Brexit **[6]**.

# 4. DESARROLLO METODOLÓGICO

## 4.1 PROCESO INVESTIGATIVO

Para lograr dar cumplimiento con los objetivos específicos número uno y dos fue necesario implementar una serie de herramientas que ayudarían en el proceso de desarrollo. En primer lugar, se realizó una matriz del estado del arte, documento cuyo objetivo era detallar la información más relevante relacionada a cada trabajo consultado para el desarrollo de la presente tesis; entre la información que esta matriz contiene se encuentra: Título, tipo de documento (artículo, tesis, revista), resumen general. A su vez esta guía funcionó para la elección correcta de los antecedentes relacionados con el proyecto y la categorización de componentes temáticos.

**Figura 5. Matriz estado del arte.**

| Autor | Titulo | Tipo de trabajo | Resumen |
|---|---|---|---|
| Boldyreva, Elena | Cambridge Analytica: Ethics And Online Manipulation With Decision-Making Process | Paper/Conferencia | Las tecnicas o procesos de recolección de datos son actualmente utilizadas en diferentes ambitos con el fin de analizar, estudiar, predecir fenomenos o factores en diferentes areas sociales incluyendo en estas la politica. El presente estudio busca explicar como fue llevado a cabo este proceso por la compañia Cambridge Analytica y cuales fueron las tecnicas y plataformas utilizadas en el camino. AMT (Amazon Mechanical Turk) fue la plataforma clave en el proceso pues logro atraer una cantidad considerable de usuarios cuyas cuentas estarian vinculadas directamente a la red social Facebook y los datos alojados en ella, sin embargo aplicando tecnicas de analiticas fue posible extraer los datos personales de 340 personas mas por cada usuario de AMT. Los datos recolectados siguieron un proceso de analisis y clasificación basados en un modelo de perspectiva disposicional que le permitiria a Cambridge Analityca determinar caracteristicas especificas por grupo de usuarios. El proceso logro recolectar los datos personales de al |

Fuente: Elaboración propia.

Con el fin de determinar qué bibliografía era de calidad para su uso, fue necesario realizar un documento de evaluación el cual tuvo como objetivo en una escala de 1 a 10 medir la calidad de información que presentaban estos documentos, dicha calidad era medida con base en la búsqueda y análisis previo en WebOfScience de las áreas núcleo y sus exponentes que conforman este proyecto, durante la evaluación se tuvieron en cuenta componentes como: autoría, fecha de publicación, tipo de documento y tipo de contenido; cada componente con información específica para evaluar.

**Figura 6. Evaluación de calidad -Tipo documento.**

| ID | ¿Es un articulo cientifico? | ¿Cuál es la revista de publicación? | ¿Factor de impacto? | Puntaje |
|---|---|---|---|---|
| 1 | 5 | 3,5 | 3,5 | **4,0** |
| 2 | 5 | 2,5 | 2,5 | **3,3** |

Fuente: Elaboración propia.

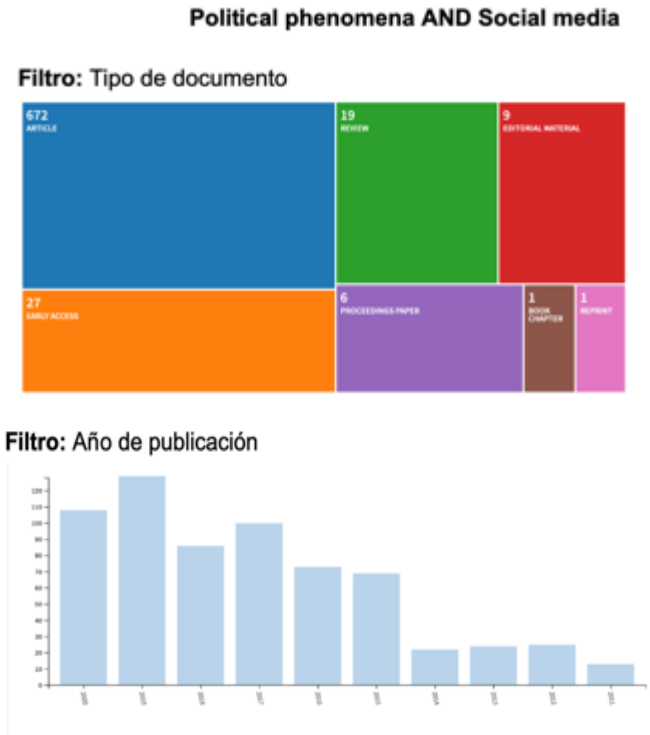
**Figura 7. Evaluación de calidad -Autoría.**

| ID | ¿Quien es el autor? | ¿Es citado por otros autores? | ¿Es relevante en la materia? | Puntaje |
|---|---|---|---|---|
| 1 | 3 | 2,5 | 2,5 | **2,7** |
| 2 | 2,5 | 2,5 | 3 | **2,7** |

Fuente: Elaboración propia.

Uno de los factores que permitió evaluar correctamente las bibliografías encontradas fue primeramente determinar cuáles eran los exponentes en las áreas de interés (Minería de datos, Big Data, Machine Learning, entre otras). En segundo lugar, fue necesario determinar quiénes eran los autores más relevantes dentro de las áreas de interés, siguiente a esto se identificaron los tipos de documentos más frecuentes (artículo, tesis, revista, etc.) y, por último, se realizó una identificación de cuáles eran los años con más publicaciones de trabajos relacionados al campo de investigación.

**Figura 8. Analisis Web of Science.**




Fuente: Elaboración propia.

Para concluir se realizó un documento que cumplía la función de categorización de componentes, el cual buscaba clasificar parte de la información alojada en los documentos en cuatro áreas principales; referentes al campo político, redes sociales, softwares utilizados y técnicas o herramientas de minería de datos. Estas categorías se clasificaron por colores, en donde el color azul era referente a la categoría campo político, que abarca todos los términos encontrados en los resúmenes, asociados con política. En Verde se categorizaron las redes sociales usadas en documentos. En color amarillo los softwares y plataformas utilizadas en los proyectos, y en color rojo, las técnicas y herramientas de minería de datos mencionadas.

**Figura 9. Categorización de componentes.**

| Autor | Título | Resumen |
|---|---|---|
| Boldyreva, Elena | Cambridge Analytica: Ethics And Online Manipulation With Decision-Making Process | Las técnicas o procesos de recolección de datos son actualmente utilizadas en diferentes ámbitos con el fin de analizar, estudiar, predecir fenómenos o factores en diferentes áreas sociales incluyendo en éstas la política. El presente estudio busca explicar cómo fue llevado a cabo este proceso por la compañía Cambridge Analytica y cuáles fueron las técnicas y plataformas utilizadas en el camino. AMT (Amazon Mechanical Turk) fue la plataforma clave en el proceso pues logró atraer una cantidad considerable de usuarios cuyas cuentas están vinculadas directamente a la red social Facebook y los datos alojados en ella, sin embargo aplicando técnicas de analiticas fue posible extraer los datos personales de 340 personas más por cada usuario de AMT. Los datos recolectados siguieron un proceso de análisis y clasificación basados en un modelo de perspectiva disposicional que le permitiría a Cambridge Analityca determinar características específicas por grupo de usuarios. El proceso logró recolectar los datos personales de alrededor de 87 millones de personas lo que permitió establecer qué contenido mostrar con el fin de lograr un cambio en la mentalidad de los votantes. |
| Solano D, 2014 | Aplicación de Minería de datos para la identificación de patrones de comportamiento en las organizaciones enfocado en prácticas de impresión | La presente tesis tiene como objetivo un estudio de minería de datos con el fin de identificar patrones relacionados a las impresiones realizadas en una empresa. El estudio siguió una metodología CRISP - DM para cumplir de manera correcta cada etapa en la minería de datos, en el desarrollo del trabajo se hicieron uso de técnicas dirigidas y no dirigidas como lo son árboles de decisión para mejores tomas de decisión sobre los datos analizados y técnicas de detección automática de cluster para visualizar patrones de comportamiento en el conjunto de datos. La fase de modelado fue posible gracias al uso de algoritmos como el SimpleKMeans, Apriori y BestFirst que fueron puestos en ejecución en el software Weka. El informe final mostró los patrones encontrados, los |

| | | |
|---|---|---|
| | | días de mayor demanda de impresión y los picos más altos por cada departamento identificado. |
| Monti, Corrado & Rozza, Alessandro & Zappella, Giovanni & Zignani, Matteo & Arvidsson, Adam & Colleoni, Elanor. | Modelling political disaffection from Twitter data | El presente trabajo tuvo como objetivo determinar si existía una relación entre el desafecto político expresado en redes sociales y el desafecto en las encuestas de opinión pública. El proyecto siguió fases como la recolección de datos de manera semiautomática gracias al uso de una API de la red social Twitter que permitió la recolección de alrededor de 35 millones tweets; La segunda fase consistió en el etiquetado donde se determinaron tres etiquetas con valores respectivos para cada una. Para determinar el desafecto en los tweets recolectados se utilizó un algoritmo cuya tarea principal era distinguir entre los tweets siguiendo el criterio de etiquetado y clasificándolos en una de las etiquetas, en la fase final se determinaba si algún tweet en particular iba dirigido a un partido o dirigente político. Analizando todos estos factores en relación a los datos recolectados encontraron una relación directamente proporcional entre el desafecto mostrado en redes sociales y el visto en las encuestas públicas. Se basan en conocer el número y ritmo de las publicaciones de los candidatos, identifican los temas centrales en los mensajes emitidos por los candidatos en sus páginas de Facebook y Twitter. Se evalúa el nivel de interactividad y el flujo de comunicación entre ciudadanos y políticos a través de las redes sociales en mención, además se compara los resultados de análisis de la campaña en redes sociales con los resultados de las elecciones generales del 2017. Por último determinar si existe un lenguaje bidireccional entre los internautas y los candidatos |

Fuente: Elaboración propia.

## 4.2 IDENTIFICACIÓN DEL FENÓMENO POLÍTICO

En el transcurso de los últimos años la migración de venezolanos se ha visto reflejada en la ciudad de Bogotá. Este fenómeno migratorio ha coincidido con un aumento delincuencial y xenofobia dentro de la capital del país. Esta migración ha generado distintas opiniones dentro de la población en Colombia, incluyendo malestar social debido al rechazo a la población venezolana, y por otra parte la aceptación y apoyo hacia los migrantes que se vieron afectados por el régimen del actual gobierno venezolano.

A partir del análisis realizado previamente de tendencias y noticias, se pudieron identificar diferentes grupos de opiniones de usuarios en los medios y redes sociales del país. Con esto se determinó la variable de interés la cual es la aceptación y/o rechazo hacía los venezolanos en Colombia. A través del estudio realizado, se encontraron diversas posiciones políticas y sociales que podrían influir en factores como: el aumento delincuencial relacionado a los inmigrantes venezolanos, disminución de oportunidades laborales (falta de empleo), aumento de xenofobia hacia los venezolanos, la aprobación y/o intolerancia hacia los ciudadanos venezolanos en el país.

De estos factores, se identifican como principales actores a los ciudadanos Bogotanos, que expresan su opinión con respecto a los factores del fenómeno migratorio. Esto corresponde al estudio visto desde la perspectiva social del fenómeno.

Por otro lado, desde el punto de vista político, se busca analizar las acciones que ha optado el gobierno para afrontar con este éxodo venezolano al país vecino. Para esto se identifican las opiniones y planes de ayudas realizados por parte del gobierno colombiano, que pueden afectar los niveles de aceptación o rechazo por parte de los ciudadanos, con respecto a factores mencionados anteriormente provocados por la migración.

A partir de la indagación realizada y el proceso investigativo, se estableció la siguiente pregunta del problema con el fin de medir el nivel aprobación de los ciudadanos colombianos, con respecto al fenómeno identificado.

¿Cómo se expresa la aceptación o rechazo por parte de la población colombiana, respecto a la gestión política y social en relación con el fenómeno de inmigración venezolana que ocurre en el país?

## 4.3 FASES DE METODOLOGÍA CRISP - DM

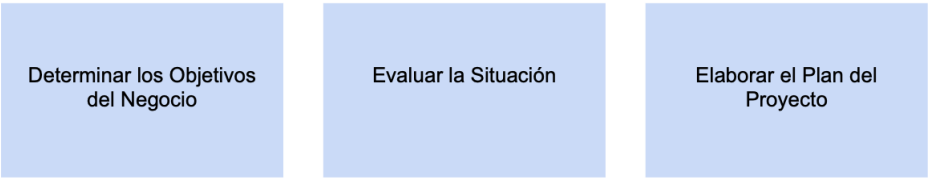
El presente capítulo está dedicado al desarrollo metodológico. En este se presenta como se llevó a cabo el proceso de seis fases que comprende la metodología CRISP - DM, de igual manera se muestran los pasos y diagramas que soporta el artefacto planteado como solución.

### 4.3.1 Fase 1 - Comprensión del negocio

Consiste en la fase inicial y se enfoca en la compresión de los objetivos del proyecto para después convertir el conocimiento de los datos en la definición del problema de minería de datos y por último se elabora un plan preliminar para lograr los objetivos propuestos.

En la figura 10, refleja las actividades que componen la fase de comprensión del negocio.

**Figura 10.** Fase 1 Comprensión del negocio.




Fuente: Elaboración propia.

#### 4.3.1.1 Determinar objetivos del negocio

##### 4.3.1.1.1 ¿Qué se desea hacer?

Se desea hacer un modelo descriptivo sobre el fenómeno político identificado (movimiento migratorio de venezolanos a Colombia) y el manejo que las entidades políticas han tenido

Para esto se tendrán en cuenta la siguiente variable en el fenómeno a analizar:
la aceptación o rechazo de la población colombiana a partir de las expresiones de las redes sociales en relación con la gestión política con respecto al fenómeno de la inmigración venezolana
Los datos se extraerán con base en estas variables de selección. Posteriormente se hará un análisis de sentimientos en redes sociales que conlleve a identificar el

MBPSC de la variable a caracterizar: Aceptación o rechazo de la población colombiana, en relación con la gestión política con respecto al fenómeno de la inmigración venezolana

#### 4.3.1.1.2 ¿Por qué usamos minería de datos para este propósito?

La minería de datos a través de sus diferentes prácticas y técnicas nos permite identificar conductas, comportamientos y patrones que no podemos ver a simple vista. Por ello para entender los factores que determinan el comportamiento de un fenómeno tan grande como lo es la política y su papel dentro de redes sociales es necesario identificar las interacciones y patrones que suceden dentro del sistema estudiado. La minería funciona como una herramienta bastante útil para darle uso sobre los datos recolectados y definir los factores, características y semejanzas más relevantes que rigen el comportamiento de un fenómeno político-social lo que posteriormente permitirá modelar la información y obtener el modelo descriptivo deseado.

#### 4.3.1.1.3 ¿Cuáles son los criterios de éxito?

Los criterios de éxito fueron establecidos a partir del problema/necesidad identificada(a) y del entregable planteado. El principal criterio de éxito para el presente trabajo será poder identificar patrones de comportamientos o conductas no visibles previamente en los datos de estudio. Un segundo criterio de éxito es poder establecer si el modelo descriptivo efectivamente me permite entender el fenómeno estudiado. Por último, el tercer criterio de éxito sería el poder modelar desde la perspectiva biopsicosocial y cultural tomando las interacciones y patrones identificados gracias a la minería de datos realizada, para este último es indispensable haber cumplido con los dos primeros criterios mencionados.

### 4.3.1.2 Evaluar la situación

#### 4.3.1.2.1 Desde el punto de vista político/social en redes sociales

Para la identificación del fenómeno político que será tratado a través de minería de datos para este proyecto de investigación, en donde se realizará como desarrollo tecnológico un modelo descriptivo para la comprensión de los flujos de información, actores e interacciones que rigen el comportamiento del fenómeno, se observaron y analizaron diferentes noticias y tendencias, durante diferentes periodos en donde ocurrieron diferentes hechos delictivos en donde se vieron involucrados personas de nacionalidad venezolana, además se tomó como parámetro de búsqueda, otras fechas en donde no ocurrió ningún evento, con el fin de comparar y analizar las opiniones de los ciudadanos a partir de Twitter, con respecto a la aprobación de los inmigrantes.

#### 4.3.1.2.2 Twitter

Se utiliza la red social Twitter como principal fuente de información y noticias, debido a que será el medio o artefacto por el cual se recolectarán, analizarán y verificarán los datos obtenidos a través de un proceso de minería de datos. Para realizar dicho proceso, se contempla el uso de herramientas como Tweepy que es una librería para Python la cual se conecta con el API de Twitter y permite obtener la captura de información, datos y estadísticas de tweets de manera semiautomática lo que permite agilizar el proceso de conseguir los datos. Esta API de Twitter tiene diferentes funcionalidades tales como la publicidad: ya que permite la creación de campañas publicitarias, seguimiento y alcance de estas; analítica: ya que permite la recolección y el análisis de los atributos de los tweets recolectados en tiempo real; y por último el acoplamiento: ya que permite la creación de bots y diferentes formas interactuar con usuarios de Twitter y generar contenido.

#### 4.3.1.2.3 Análisis de tendencias en noticias y redes sociales.

Durante la indagación de los temas más relevantes tratados y discutidos en las noticias, revistas y redes sociales, se tuvo una constante exploración de las tendencias de Colombia y Bogotá durante los últimos meses mencionados previamente. A lo largo de estas revisiones, se encontraron debates e intercambios de opiniones sobre diferentes temas polémicos de la actualidad tales como la aprobación por parte de los colombianos, con respecto a la inmigración de ciudadanos venezolanos en el país entre otras controvertidas opiniones.

A partir de la exploración y el análisis de estas tendencias, se seleccionará como problema de investigación, para muestra del modelo a desarrollar, las opiniones mostradas por parte de ciudadanos colombianos con respecto a la inmigración de venezolanos en el país y los problemas que se han generado, conllevando a opiniones xenófobas expresadas en la red social Twitter.

**Figura 11. Fuentes de noticias**

Fuentes_Noticias

Archivo Editar Ver Insertar Formato Datos Herramientas Complementos Ayuda Última modificación de VICTOR MONTILVA el 24 de noviembre de 2020

| | A | B | C |
|---|---|---|---|
| 1 | Hipervínculo | Título de la referencia | Año Publicación |
| 2 | https://www.bancomundial.org/es/news/feature/2019/11/26/migracion-venezolana-4500-kilometros-entre-el-abandono-y-la-oportunidad | Migración venezolana: 4.500 kilómetros entre el abandono y la oportunidad | 2019 |
| 3 | https://data.humdata.org/organization/immap | iMMAP | N/A |
| 4 | https://migrationdataportal.org/es?i=stock_abs_&t=2019 | PORTAL DE DATOS MUNDIALES SOBRE LA MIGRACIÓN | 2019 |
| 5 | https://datosmacro.expansion.com/demografia/migracion/inmigracion/venezuela | El número de inmigrantes desciende en Venezuela | |
| 6 | https://www.acnur.org/noticias/press/2019/6/5cfa5eb64/refugiados-y-migrantes-de-venezuela-superan-los-cuatro-millones-acnur-y.html | Refugiados y migrantes de Venezuela superan los cuatro millones: ACNUR y OIM | 2019 |
| 7 | https://ukdiss.com/examples/social-media-in-political-communication-9877.php#_ENREF_18 | Role of Social Media in Political Communication | 2019 |
| 8 | https://www.infobae.com/politica/2019/03/11/internet-y-elecciones-como-las-redes-sociales-irrumpen-en-la-comunicacion-politica/ | Internet y elecciones: cómo las redes sociales irrumpen en la comunicación política | 2019 |
| 9 | https://www.redalyc.org/jatsRepo/5177/5177527177019/html/index.html | Aspectos legales al utilizar las principales redes sociales en Colombia | 2020 |
| 10 | https://www.mincit.gov.co/CMSPages/GetFile.aspx?guid=c52859bb-0056-46f8-a97d-c9d23e41ee98 | Redes Sociales y Participación | N/A |
| 11 | https://www.semana.com/confidenciales-semanacom/articulo/en-twitter-piden-la-renuncia-de-claudia-lopez-y-daniel-quintero/648910/ | En Twitter piden la renuncia de Claudia López y Daniel Quintero | 2020 |
| 12 | https://pluralidadz.com/politica/claudia-lopez-yo-soy-la-alcaldesa-de-bogota-no-la-de-twitter/ | Claudia López: "yo soy la alcaldesa de Bogotá, no la de Twitter" | 2020 |
| 13 | https://www.eltiempo.com/bogota/claudia-lopez-la-alcaldesa-de-bogota-se-pronuncia-tras-ser-victima-de-acto-de-homofobia-536229 | Las palabras de Claudia López tras ser víctima de ataque homofóbico | 2020 |
| 14 | https://www.eltiempo.com/bogota/claudia-lopez-alcaldesa-de-bogota-dijo-que-algunos-venezolanos-estan-metidos-en-la-criminalidad-546180 | Claudia López propone deportar a venezolanos metidos en el crimen | 2020 |
| 15 | https://www.semana.com/nacion/articulo/xenofobia-en-colombia-contra-los-venezolanos/569808/ | Alerta por xenofobia en contra de los venezolanos en Colombia | 2018 |
| 16 | https://www.elespectador.com/noticias/el-mundo/asi-se-mueve-la-xenofobia-contra-los-migrantes-venezolanos-en-redes/ | Así se mueve la xenofobia contra los migrantes venezolanos en las redes | 2020 |
| 17 | https://elpais.com/sociedad/2020-08-18/los-migrantes-venezolanos-en-colombia-afrontan-la-pandemia-entre-la-vulnerabilidad-y-la-xenofobia.html | Los migrantes venezolanos en Colombia afrontan la pandemia entre la vulnerabilidad y la xenofobia | 2020 |
| 18 | https://www.voanoticias.com/america-latina/colombia-protestas-provocan-xenofobia-contra-migrantes-venezolanos | Protestas en Colombia provocan xenofobia contra migrantes venezolanos | 2019 |
| 19 | https://www.rcnradio.com/politica/onu-alerta-brotes-de-xenofobia-contra-migrantes-venezolanos-en-colombia | ONU alerta brotes de xenofobia contra migrantes venezolanos en Colombia | 2019 |
| 20 | https://www.elnacional.com/mundo/virus-xenofobia-contra-venezolanos-ha-aumentado-en-colombia/ | Con el virus la xenofobia contra venezolanos ha aumentado en Colombia | 2020 |
| 21 | https://cnnespanol.cnn.com/2020/04/05/alerta-colombia-ivan-duque-pide-solidaridad-con-los-venezolanos-desplazados-en-medio-de-la-pandemia-de-coronavirus/ | Iván Duque pide solidaridad con los venezolanos desplazados en medio de la pandemia de coronavirus | 2020 |
| 22 | https://www.eltiempo.com/bogota/tutela-contra-claudia-lopez-por-comentarios-xenofobos-fue-rechazada-550244 | Rechazan tutela contra López por supuestos comentarios xenófobos | 2020 |
| 23 | https://www.larepublica.co/economia/la-oim-lanzo-campana-en-contra-de-la-xenofobia-hacia-los-migrantes-venezolanos-3020285 | OIM lanzó campaña en contra de la xenofobia hacia refugiados y migrantes venezolanos | 2020 |
| 24 | http://www.scielo.org.co/pdf/pacla/v21n4/0122-8285-pacla-21-04-00960.pdf | Xenofobia y periodismo: Colombia y la migración venezolana | 2018 |
| 25 | https://es.wikipedia.org/wiki/Sentimiento_antivenezolano | Sentimiento antivenezolano | 2020 |
| 26 | https://www.europapress.es/internacional/noticia-oea-lanza-campana-contra-discriminacion-xenofobia-venezolanos-desplazados-20200625022157.html | La OEA lanza una campaña contra la discriminación y la xenofobia hacia los venezolanos desplazados | 2020 |
| 27 | https://oxfamilibrary.openrepository.com/bitstream/handle/10546/620890/bp-si-pero-no-aqui-251019-es.pdf?sequence=8&isAllowed=y | SÍ, PERO NO AQUÍ Percepciones de xenofobia y discriminación hacia migrantes de Venezuela en Colombia, Ecuador y Perú | 2019 |
| 28 | https://globalizateradio.com/2020/06/12/procurador-de-colombiadebemos-evitar-actos-de-xenofobia-contra-los-migrantes-venezolanos-que-retornan-a-la-frontera/ | Procurador De Colombia:"Debemos Evitar Actos De Xenofobia Contra Los Migrantes Venezolanos | 2020 |

Fuente: Elaboración propia.

#### 4.3.1.3 Elaborar plan del proyecto

En la última tarea propuesta en esta fase, se busca realizar un plan de proyecto que facilite la organización durante la ejecución del trabajo y donde se establezca el tiempo estimado de cada etapa a partir de la aplicación de la metodología escogida.

**Etapa 1:** Se realiza una investigación exhaustiva en noticias y redes sociales con el fin de recopilar información y datos relevantes que puedan servir en el presente estudio. Tiempo estimado: 3 - 4 semanas.

**Etapa 2:** En esta etapa se describe y se realiza una revisión general de los datos recopilados anteriormente, con el fin de determinar la calidad de los datos. Tiempo estimado: 2 - 3 semanas.

**Etapa 3:** Se realiza la preparación de los datos (limpieza, construcción y aplicación de formatos) con el fin de facilitar la minería de datos. Tiempo estimado. Tiempo estimado: 4 - 5 semanas.

**Etapa 4:** Corresponde a la elección de técnicas de modelado y la aplicación de estas sobre los datos. Tiempo estimado: 5 - 6 semanas.

**Etapa 6:** Análisis de los resultados que se obtuvieron en la fase previa Tiempo estimado: 2 - 3 semanas.

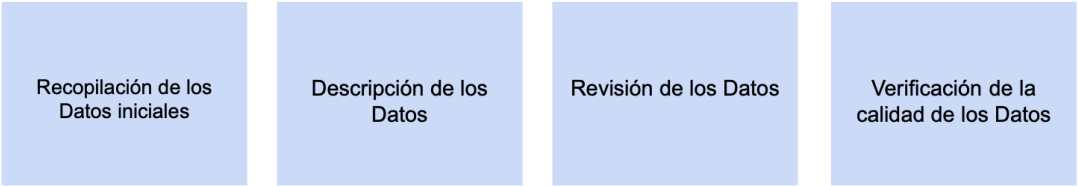
**Etapa 7:** Producción de informes con los resultados obtenidos. Tiempo estimado: 1 - 2 semanas.

**Etapa 8:** Presentación al beneficiario de los resultados finales Tiempo estimado: 1 semana.

### 4.3.2 Fase 2. Comprensión de los datos

En la segunda fase se comienza la recolección inicial de datos, y se realizan las tareas que permiten familiarizarse con estos (identificar problemas, descubrir conocimiento sobre los datos, descubrir subconjuntos). Este es el primer acercamiento que tendremos sobre los datos los cuales deben ser revisados detenidamente para lograr contar con una información de calidad para su posterior estudio. En la figura 12, se refleja las actividades que componen la fase de comprensión de los datos.

**Figura 12.** Fase 2 Comprensión de los Datos

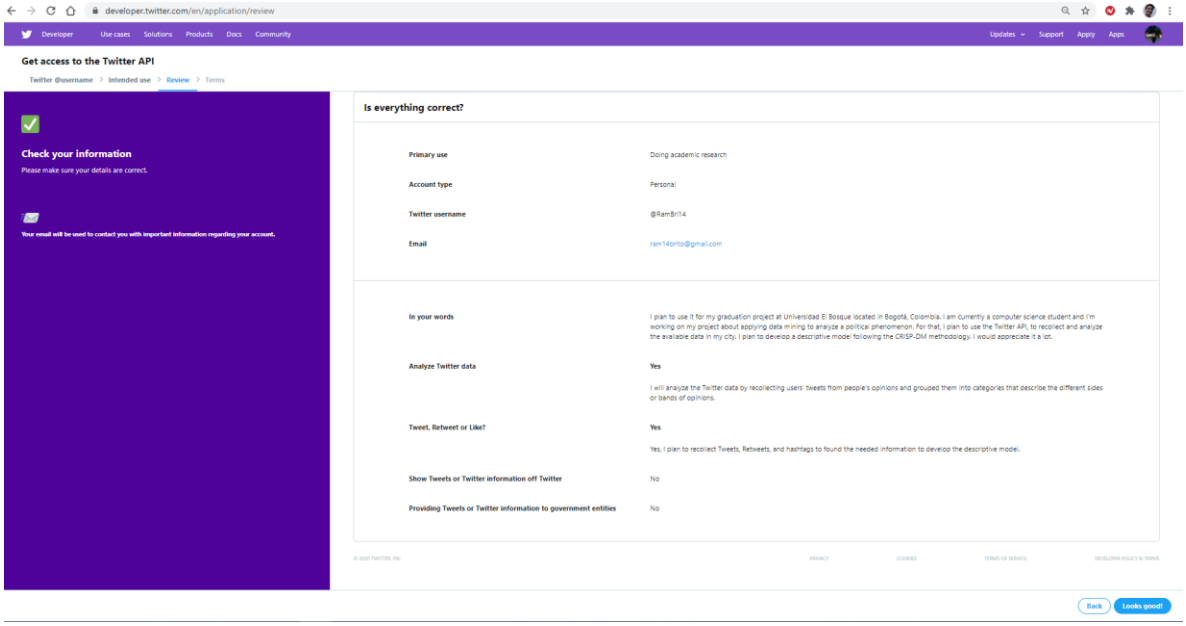



Fuente: Elaboración propia.

#### 4.3.2.1 Recopilación de los datos iniciales

Para la recopilación de los datos, se hizo uso de la API de Twitter, que permite la recolección de tweets de la red social. Para lograr esto, se requiere solicitar acceso al API de Twitter, mediante el portal web de Twitter developer https://developer.twitter.com/en/application/intent.

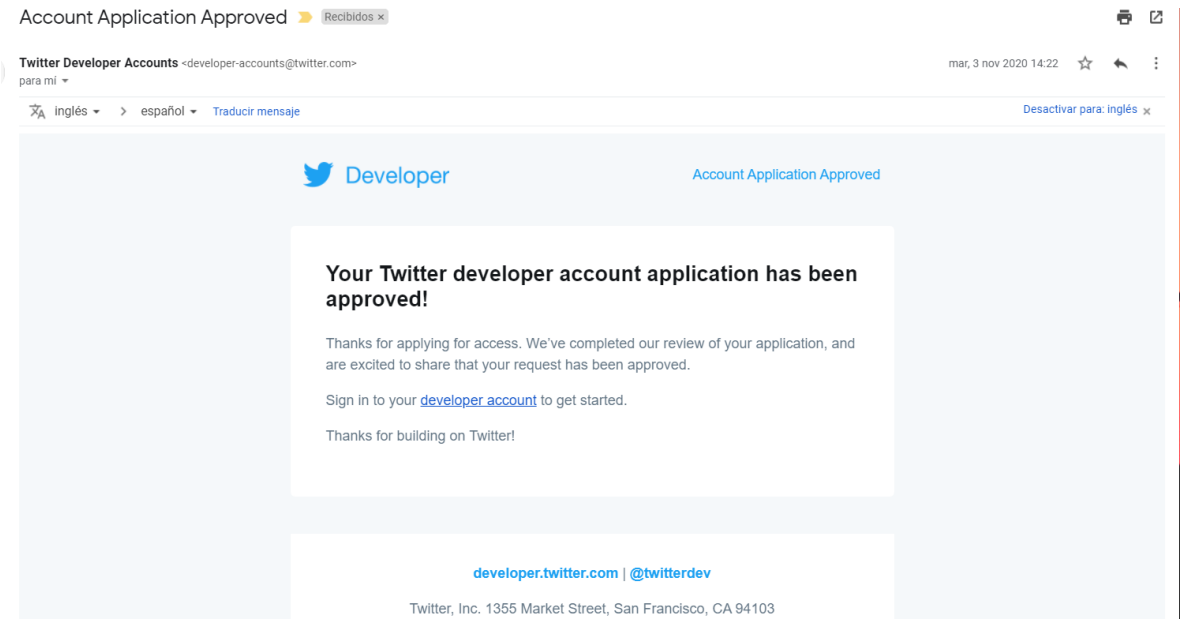
**Figura 13.** Solicitud para el API de Twitter.




Fuente: Elaboración propia.

Una vez enviada esta solicitud, fueron concedidos los Tokens o claves requeridos para hacer uso de esta API.

**Figura 14.** Credenciales aprobadas.

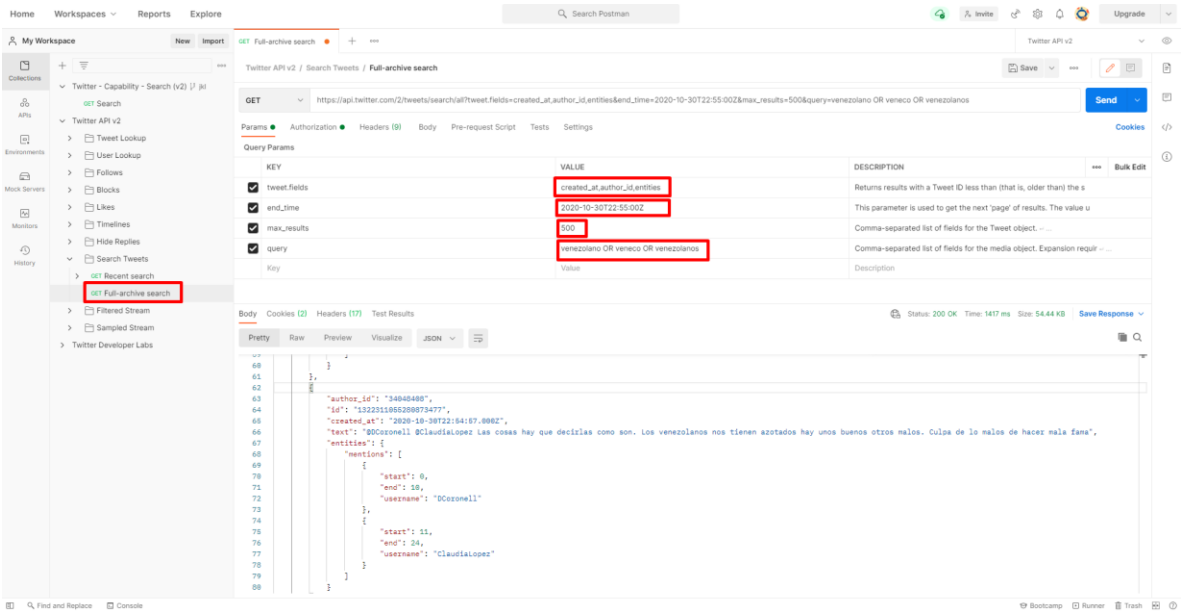



Fuente: Elaboración propia.

Para esta primera tarea de esta fase, se obtuvo un conjunto de datos a través de la extracción de tweets, utilizando la API de Twitter V2 a través de la herramienta *Postman.* Con esto se recopilaron inicialmente 32.454 tweets de Twitter, utilizando un conjunto de palabras claves como “venezolano”, “veneco”, “venezolanos”, en donde usuarios de Twitter realizaban publicaciones con estos términos para referirse a los actores principales del fenómeno migratorio.

Para la extracción de estos datos, se seleccionaron diferentes rangos de fechas dentro del año 2020 y 2021. Se recopilaron tweets publicados durante los meses septiembre, octubre y noviembre en donde a partir de la recolección de noticias realizada anteriormente, se pudo aprovechar ocurrencias y discusiones en Twitter con respecto a los migrantes venezolanos. Una de ellas debido al asesinato ocurrido el 30 de octubre en una estación de Transmilenio en la ciudad de Bogotá. Por lo que resultaron muchas interacciones en la red social, involucrando en menciones a la alcaldesa de Bogotá Claudia López, así como también otras entidades distritales de la ciudad.
Durante el rango de fechas entre el 10 de marzo del 2021 al 16 de marzo del mismo año, en donde ocurrió un hecho delictivo en donde un ciudadano de nacionalidad venezolana acabó con la vida de un policía nacional. Durante, se expresaron muchas personas mediante Twitter para rechazar este hecho.
Para la extracción de estos datos se hizo uso de la colección de *Full archive search* de la API de Twitter V.2 a través de la herramienta *Postman* para obtener la recopilación de estos datos.

**Figura 15.** Extracción mediante Full archive search collection.




Fuente: Elaboración propia.

A partir de esa búsqueda se obtuvieron diferentes archivos tipo JSON para cada una de las consultas realizadas, ya que se realizaron diferentes consultas para diferentes rangos de horas en distintos días.

#### 4.3.2.2 Descripción de los datos

Luego de conformado el *dataset*, haciendo uso de la librería *pandas profiling* en *python* se realizó un reporte de análisis sobre los datos que fue exportado en formato html. A partir del reporte generado se pudieron identificar cuatro variables dentro del conjunto de datos de las cuales dos eran de tipo numéricas y dos de tipo categóricas; las observaciones totales corresponden al número de datos alojados dando como resultado un total de 10.764 observaciones. El *dataset* tuvo un total de 514 celdas vacías lo que representó una carencia de información del 1.2% en relación al total de celdas.

**Figura 16. Perfilamiento general de los datos.**

Overview | Warnings 7 | Reproduction

Dataset statistics

| | |
|---|---|
| Number of variables | 4 |
| Number of observations | 10764 |
| Missing cells | 514 |
| Missing cells (%) | 1.2% |
| Duplicate rows | 0 |
| Duplicate rows (%) | 0.0% |
| Total size in memory | 336.5 KiB |
| Average record size in memory | 32.0 B |

Variable types

| | |
|---|---|
| Numeric | 2 |
| Categorical | 2 |

Fuente: Elaboración propia.

En relación con las variables obtenidas, encontramos que dos de tipo categóricas tienen el atributo de *High_Cardinality* lo que en estadística descriptiva representa aquellos datos que pueden tomar una gran cantidad de valores distintos [32]. En relación con la variable *text* observamos que contiene 10764 valores distintos y sin tener valores perdidos, similar el caso de la variable *created_at* la cual tiene un total de 9596 valores distintos, y sin presentar pérdidas de celdas.

**Figura 17. Perfilamiento de las variables text, created_at**

text
Categorical
HIGH CARDINALITY
UNIFORM
UNIQUE

| | |
|---|---|
| Distinct | 10764 |
| Distinct (%) | 100.0% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 84.2 KiB |

created_at
Categorical
HIGH CARDINALITY
UNIFORM

| | |
|---|---|
| Distinct | 9596 |
| Distinct (%) | 89.1% |
| Missing | 0 |
| Missing (%) | 0.0% |
| Memory size | 84.2 KiB |

Fuente: Elaboración propia.

Con respecto a la variable numérica author_id observamos un total de 8150 valores distintos y un total de 514 celdas vacías lo que representa el 4.8% con respecto al total de celdas, siendo esta entre todas las variables con mayor pérdida de información.

**Figura 18. Perfilamiento de la variable author_id**

author_id
Real number ($\mathbb{R}_{\geq 0}$)
MISSING

| | | | |
|---|---|---|---|
| Distinct | 8150 | Minimum | 12864 |
| Distinct (%) | 79.5% | Maximum | $1.371208475 \times 10^{18}$ |
| Missing | 514 | Zeros | 0 |
| Missing (%) | 4.8% | Zeros (%) | 0.0% |
| Infinite | 0 | Negative | 0 |
| Infinite (%) | 0.0% | Negative (%) | 0.0% |
| Mean | $4.295877478 \times 10^{17}$ | Memory size | 84.2 KiB |

Fuente: Elaboración propia.

#### 4.3.2.3 Revisión de los datos

En esta tercera tarea de la segunda fase, se validó visualmente que los datos extraídos, fueran los esperados de la extracción y necesarios para la construcción del *dataset* utilizado para el desarrollo del modelo. En esta etapa se verifica que los datos sean de utilidad y no existan valores nulos que actúen como ruido en la aplicación de los algoritmos, así como también identificar inconsistencias del texto, tales como emoticones, símbolos, caracteres especiales, entre otros caracteres que pudieran complicar el proceso de análisis de sentimientos, para luego en la siguiente Fase de preparación de los datos (Fase 3) poder realizar una limpieza de estas inconsistencias.

Para esta etapa, se especificó en la herramienta postman los parámetros de búsqueda más útiles para el análisis. En primer lugar, el parámetro de tweet.fields que arroja valores como la fecha de creación del tweet, el número de identificación del usuario que realizó la publicación, y las entidades que se mencionan en el tweet. Otro parámetro es el start_time y el end_time, especificando el año, mes día, hora, minutos y segundos en cada uno de estos dos campos, con el formato de fecha YYYY-MM-DDTHH:mm:ss (ISO 8601/RFC 3339). En este caso no se tuvo en cuenta el parámetro de start_time ya que, al realizar la búsqueda, la API arroja como resultados los tweets más recientes publicados desde la fecha final establecida (end_time) hasta el último de los 500 primeros tweets.

En estos campos también se definieron las palabras clave del query, que en este caso fueron las palabras claves que se deseaban para hacer la búsqueda de los tweets. También se aclaró el número máximo de resultados, que es 500 (dado que es el límite máximo por consulta que permite la API).

**Figura 19. Parámetros y valores definidos para la extracción.**

Query Params

| KEY | VALUE |
|---|---|
| tweet.fields | created_at,author_id,entities |
| end_time | 2020-10-30T22:55:00Z |
| max_results | 500 |
| query | venezolano OR veneco OR venezolanos |
| Key | Value |

Fuente: Elaboración propia.

#### 4.3.2.4 Verificación de la calidad de datos

Para esta última tarea, se comprobó que los datos (tweets) fueran reales, escogiendo al azar varios tweets extraídos y realizando posteriormente una búsqueda manual del usuario de Twitter, dentro de la red social, e identificar el tweet que fue extraído. Además, se verificó que el contenido de los tweets, fueran útiles para lograr un adecuado análisis de opiniones, con respecto al modelo descriptivo que se desea elaborar. Esta verificación del contenido de los tweets se realizó con respecto a la variable a medir de la validación de la academia, en donde se mide la aceptación y/o rechazo de los usuarios de Twitter, acerca del fenómeno político escogido.

Para este proceso también se tuvo en cuenta la relación de las entidades con los términos de búsqueda establecidos, con el fin de verificar que se mencionasen estas entidades en los tweets relacionados al fenómeno político seleccionado, como se puede observar en la figura, el archivo resultante en formato JSON, identifica las entidades mencionadas en el texto.

**Figura 20. Entidades mencionadas en tweets.**

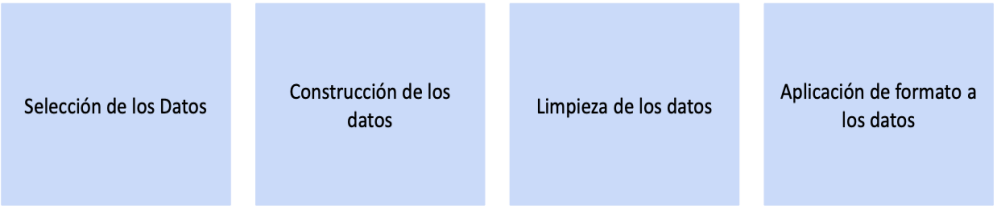

```
{
    "text": "RT @JoJakubowicz: A propósito de los comentarios de @ClaudiaLopez la alcalde xenófoba de Bogotá, los venezolanos cometen menos del 1% de lo…",
    "author_id": "95501672",
    "entities": {
        "annotations": [
            {
                "start": 89,
                "end": 94,
                "probability": 0.9704,
                "type": "Place",
                "normalized_text": "Bogotá"
            }
        ],
        "mentions": [
            {
                "start": 3,
                "end": 16,
                "username": "JoJakubowicz"
            },
            {
                "start": 52,
                "end": 65,
                "username": "ClaudiaLopez"
            }
        ]
    },
    "created_at": "2020-10-30T23:58:53.000Z",
    "id": "1322327144563986433"
```

Fuente: Elaboración propia.

### 4.3.3 Fase 3. Preparación de los datos

La fase de preparación abarca todas las tareas necesarias para la construcción del conjunto de datos finales que serán estudiados y utilizados en las herramientas de modelado. Entre las tareas encontramos selección de tablas, registros y atributos, transformación y limpieza de datos.
Ver la figura 21, refleja las actividades que componen la fase de preparación de los datos.

**Figura 21:** Fase 3 Preparación de los Datos




Fuente: Elaboración propia.

#### 4.3.3.1 Selección de los datos

Durante esta primera tarea de la preparación de datos, se escogieron los datos que fueran útiles para la construcción del dataset, seleccionando por región los tweets que fueron recopilados, esto gracias a la ubicación geográfica que fue extraída en la recolección de los datos. Se seleccionaron todos los tweets que fueron publicados dentro de toda Colombia, ya que, al momento de recopilar los datos, se obtuvieron tweets publicados desde cualquier lugar del mundo. Es necesario seleccionar los tweets dentro de Colombia, ya que es la región en donde se estudia el fenómeno político-social escogido.

Luego de este proceso, se tomaron los archivos JSON resultantes de la búsqueda configurada con la fecha de inicio y fin deseada, y el valor de la palabra deseada. Se obtuvo como resultado los tweets especificados con los siguientes campos resultados:

- text: cadena de caracteres del texto obtenido de un tweet.
- author_id: Código de identificación del usuario de Twitter.
- entities: identificador de identidades mencionadas en el texto y otras anotaciones como lugares.
- created_ at: fecha de publicación del tweet.
- id: número de identificación único de un tweet

**Figura 22. Campos resultantes de la extracción.**


```
"text": "@ClaudiaLopez Su Libertador fue venezolano. No lo olvide a la hora de encasillar. Y recuerde además que muchos hermanos colombianos escogieron vivir en Venezuela y viven
    aún aquí.",
"author_id": "533246155",
"entities": {
    "annotations": [
        {
            "start": 152,
            "end": 160,
            "probability": 0.9957,
            "type": "Place",
            "normalized_text": "Venezuela"
        }
    ],
    "mentions": [
        {
            "start": 0,
            "end": 13,
            "username": "ClaudiaLopez"
        }
    ]
},
"created_at": "2020-10-30T23:58:14.000Z",
"id": "1322326982093230081"
```


Fuente: Elaboración propia.

#### 4.3.3.2 Construcción del conjunto de datos

Debido a que el formato inicial de los datos extraídos a través de postman se encontraban en formato JSON fue necesario convertir los dataset extraídos en archivos de tipo .xlsx. Además de esto, fue necesario unificar los diferentes archivos obtenidos por consultas, ya que como se mencionó anteriormente, fue necesario realizar varias peticiones en Postman con 500 tweets cada una. A partir de esto se construyó un único dataset de 32.454 datos para un evento socio político.

**Figura 23. Archivo JSON con datos extraídos.**


```
GET https://api.twitter.com/2/tweets/search/all?tweet.fields=created_at,author_id,entities&end_time=2020-10-30T23:59:00Z&max_results=500&query=@claudialopez
Send
Params Authorization Headers (9) Body Pre-request Script Tests Settings Cookies
Body Cookies (2) Headers (17) Test Results Status: 200 OK Time: 2.01 s Size: 52.34 KB Save Response
Pretty Raw Preview Visualize JSON
902    {
903        "text": "@ClaudiaLopez Su Libertador fue venezolano. No lo olvide a la hora de encasillar. Y recuerde además que muchos hermanos colombianos escogieron vivir en Venezuela y viven
               aún aquí.",
904        "author_id": "533246155",
905        "entities": {
906            "annotations": [
907                {
908                    "start": 152,
909                    "end": 160,
910                    "probability": 0.9957,
911                    "type": "Place",
912                    "normalized_text": "Venezuela"
913                }
914            ],
915            "mentions": [
916                {
917                    "start": 0,
918                    "end": 13,
919                    "username": "ClaudiaLopez"
920                }
921            ]
922        },
923        "created_at": "2020-10-30T23:58:14.000Z",
924        "id": "1322326982093230081"
925    },
926    {
927        "text": "@ClaudiaLopez Hola señora @ClaudiaLopez  soy Gustavo Ramírez un joven  padre de familia y un ciudadano trabajador que se vino Huyendo por Razones Políticas, y  me encuentro
               aquí en Antioquia,  así que  #NoSoyDelincuente le agradezco no generalice.",
928        "entities": {
929            "hashtags": [
930                {
931                    "start": 202,
932                    "end": 219,
933                    "tag": "NoSoyDelincuente"
934                }
935            ],
```


Fuente: Elaboración propia.

**Figura 24. Archivo convertido con extensión .xlsx de los datos extraídos.**

| | A | B | C |
|---|---|---|---|
| 1 | author_id | text | created_at |
| 2 | 1217237490530123777 | RT @SeijasLesbia: Son sean tan falta de respeto.... Tengan un poco de dignidad, como se les ocurre este Twitter.  La gente se esta muriendo... | 2020-11-02T21:44:55.000Z |
| 3 | 1274736262910443520 | RT @PetroDivisa: El Capitalismo sin inversión Social se vuelve salvaje, depredador e inhumano. Muchos de los miles de venezolanos que retor... | 2020-11-02T21:44:55.000Z |
| 4 | 4046968707 | Luego de varios meses cerrada debido a la pandemia de Covid-19, reabrimos nuestras puertas este lunes 02 de noviembre. Con distancia física, uso obligatorio de masc | 2020-11-02T21:44:50.000Z |
| 5 | 221104315 | RT @CATERINAV: #2Nov Dice Gustavo Petro: "Si le vendemos alimentos a Venezuela, los venezolanos no vendrían en masa a Colombia.En una prop... | 2020-11-02T21:44:48.000Z |
| 6 | 222875387 | RT @partidoUNT: #2Nov \| Nuestro equipo de @untaragua se juramenta en el Comando Regional de cara a la Consulta Popular, comprometidos con t... | 2020-11-02T21:44:45.000Z |
| 7 | 4649138123 | RT @VTVcanal8: #AUDIO \| Protector de Táchira, @FreddyBernal: Le hemos propuesto a Colombia que elaboremos un protocolo de ingreso para los... | 2020-11-02T21:44:41.000Z |
| 8 | 622263524 | RT @GNB_DeltaD619: #02Nov POLÍTICO-MILITAR VENEZOLANO #FANB #GNBYPuebloEsperanzaAbsoluta #CienciaEsSoberanía @GNBGDjuvenal @GNBoficialVE... | 2020-11-02T21:44:40.000Z |
| 9 | 274171163 | RT @saludhernandezm: Ahora @ClaudiaLopez culpa a Migración Colombia por el asesinato del señor en Transmilenio. Y estigmatiza a los venezol... | 2020-11-02T21:44:39.000Z |
| 10 | 3346119844 | RT @libertad003: VE#2Nov\|\| ¡Fuego sagrado!, es la llamarada que nuestra #FANB arraiga en el más profundo ser del Oficial y Soldado venezola... | 2020-11-02T21:44:39.000Z |
| 11 | 874318940054458368 | RT @MercalTrujillo_: EFEMÉRIDES Este #2Nov se cumplen 40 años del paso a la inmortalidad del maestro César Rengifo. Un gran artista venezol... | 2020-11-02T21:44:39.000Z |
| 12 | 974476269315338243 | RT @RedRadioVe: VE Este domingo, el jefe de Estado venezolano, @NicolasMaduro, informó la activación de nuevas instalaciones médicas en 3... | 2020-11-02T21:44:36.000Z |
| 13 | 708478690984726528 | @Itzeldt2388 @KattyValeroLea1 @elpais_america @realDonaldTrump No entiendo, los venezolanos apoyando a locotrump que cojones tienen,  por eso están jodidos | 2020-11-02T21:44:35.000Z |
| 14 | 77774057 | RT @CarlosNovoT: Colaboracionistas, al lado de eso que llaman "izquierda mundial", ligan por Biden. Otros, queremos que Trump repita.Es p... | 2020-11-02T21:44:35.000Z |
| 15 | 222875387 | RT @partidoUNT: Exhortamos a todos venezolanos a participar unidos en la consulta popular, de esta manera haremos visible para la Comunidad... | 2020-11-02T21:44:34.000Z |
| 16 | 4649138123 | RT @PetroDivisa: El Capitalismo sin inversión Social se vuelve salvaje, depredador e inhumano. Muchos de los miles de venezolanos que retor... | 2020-11-02T21:44:32.000Z |
| 17 | 4649138123 | RT @ConElMazoDando: Beisbol profesional venezolano arranca el 15 de noviembre https://t.co/ZY5q5fWuS2  #6DVictoriaPerfecta https://t.co/SoM... | 2020-11-02T21:44:31.000Z |
| 18 | 949856740014739457 | RT @VTVcanal8: Presidente Maduro anuncia el inicio del béisbol venezolano para el 15 de noviembre#FlexibilizaciónSeguraYVigiladahttps:/... | 2020-11-02T21:44:31.000Z |
| 19 | 2332415768 | RT @EAristeguieta: ¡AUCH! Tribunal español niega extradición de un venezolano por considerar ilegítimos a Tarek William Saab y Maikel Moren... | 2020-11-02T21:44:26.000Z |
| 20 | 4649138123 | RT @VillegasPoljak: Presidente @NicolasMaduro acogió propuesta de cultores venezolanos, que consignaron sus firmas, y decretó el #31Oct com... | 2020-11-02T21:44:24.000Z |
| 21 | 505073107 | RT @GNB_antidrogas1: Desde la .@GNB_Lara iniciamos desplegados esta semana de flexibilización de la Cuarentena, bajo el método 7+7, instand... | 2020-11-02T21:44:23.000Z |
| 22 | 1294030786392465408 | RT @NicolasMaduro: Mis felicitaciones y mi reconocimiento a los 3.432 nuevos profesionales que se suman a la batalla por la salud, por la v... | 2020-11-02T21:44:23.000Z |
| 23 | 974476269315338243 | RT @RedRadioVe: #REDviral Ante la imposibilidad de salir a comprar en algún establecimiento, un grupo de jóvenes venezolanos decidieron l... | 2020-11-02T21:44:18.000Z |
| 24 | 806088468715671553 | RT @VamosJuntosNE: "Este 6D ganaremos la Asamblea Nacional para acompañar a nuestro Presidente @NicolasMaduro a garantizar la Paz y la pros... | 2020-11-02T21:44:15.000Z |
| 25 | 1150552243 | RT @teleiguana: Cada día el filósofo y comunicador venezolano, Pérez Pirela @maperezpirela los acompaña #DesdeDondeSea con reflexiones, aná... | 2020-11-02T21:44:14.000Z |
| 26 | 124172948 | Seguridad, el primer pensamiento de los venezolanos ante la inauguración de los cajeros dolarizados https://t.co/l75o2H4jMr | 2020-11-02T21:44:13.000Z |
| 27 | 1344707382 | RT @EugeniaTunja: @CarlosOAlbornoz Afortunadamente @MigracionCol respondió, claro y sencillo a la sra @Bogota @ClaudiaLopez. https://t.co/W... | 2020-11-02T21:44:12.000Z |
| 28 | 302026759 | RT @cmdtemiliciab1: Los Milicianos continúan efectuando el abordaje, concientizando al Pueblo venezolano sobre el Plan de Bioseguridad o... | 2020-11-02T21:44:12.000Z |
| 29 | 1073311668157980674 | RT @sergioulloa1: Hay un descontrol total. Es impresionante la cantidad de cubanos ilegales en Chile. Y hablamos sólo de cubanos. Imaginen... | 2020-11-02T21:44:12.000Z |
| 30 | 864410774 | RT @dariomanuel: #LibereraRodneyAlvarez Obrero político preso más Antiguo de #Venezuela su Delito defender a los trabajadores venezolano 9... | 2020-11-02T21:44:10.000Z |
| 31 | 1036206139 | RT @DonaMed_VE: ¿Quieres un "mejor país"? Conviértete en un venezolano ejemplar. #LideraConElBuenEjemplo :) | 2020-11-02T21:44:10.000Z |
| 32 | 781451864 | RT @El_Cooperante: #2Nov \| Puerto Rico PR pide extradición de un colombiano y un venezolano por tráfico de cocaína  https://t.co/1zBFvBkm... | 2020-11-02T21:44:09.000Z |
| 33 | 178350656 | RT @ReinaldoProfeta: Venezolanos que piensan votar por #Biden miren eso. https://t.co/eY4JAZHY66 | 2020-11-02T21:44:08.000Z |
| 34 | 103707413 | RT @SoyArepita: Soda Stereo tocando al amanecer en El Cafetal #Caracas, Miguel Ríos, Paralamas, Fito Páez...Hace 29 años (#2Nov 1991) com... | 2020-11-02T21:44:07.000Z |
| 35 | 316093543 | RT @JHdzA3: @AdamSmithFree Personas, seres humanos huyendo de regímenes: cubanos en balsas atraviesan el mar, venezolanos las trochas, por... | 2020-11-02T21:44:05.000Z |
| 36 | 2767590820 | Beisbol profesional venezolano arranca el 15 de noviembre https://t.co/ZY5q5fWuS2  #6DVictoriaPerfecta https://t.co/SoMqatEyYB | 2020-11-02T21:44:03.000Z |
| 37 | 3245584270 | RT @VilmaPREZEchev2: Ley de Inmigración urgente, 150 venezolanos diarios, por frontera, quien mierda Gobierna este País,  que tan difícil p... | 2020-11-02T21:44:00.000Z |
| 38 | 733531024248832001 | RT @ramonacosta33: Trump: "Hay una razón por la cual Biden atrae el respaldo del socialista colombiano Gustavo Petro, del venezolano Diosda... | 2020-11-02T21:43:58.000Z |
| 39 | 50472273 | RT @VenezuelaQuilla: Smart Clinic, Unidad Móvil de Salud: Una oportunidad para la integraciónLa asociación Venezolanos en Barranquilla se... | 2020-11-02T21:43:57.000Z |
| 40 | 917079245175828485 | RT @melfossi: Los venezolanos les debemos HONOR a quienes con su vida dejaron en evidencia que el régimen revolucionario de HRCH y del ileg... | 2020-11-02T21:43:52.000Z |
| 41 | 254333960 | Seguimos luchando por recuperar la democracia en Venezuela. Los Comandos para la Consulta Popular, han sido juramentados en todo el país. #JuntosAlzamosLaVoz cc | 2020-11-02T21:43:44.000Z |
| 42 | 1320832516736110593 | RT @VamosJuntosNE: "Este 6D ganaremos la Asamblea Nacional para acompañar a nuestro Presidente @NicolasMaduro a garantizar la Paz y la pros... | 2020-11-02T21:43:44.000Z |
| 43 | 305479775 | RT @cmdtemiliciab1: Los Milicianos continúan efectuando el abordaje, concientizando al Pueblo venezolano sobre el Plan de Bioseguridad o... | 2020-11-02T21:43:43.000Z |
| 44 | 1286620867254984706 | RT @greider_son: @jcsigo Aquí los medios Venezolanos todos están bajo las Ordenes de Régimen y siempre han dado ganador a Biden, aun sabien... | 2020-11-02T21:43:42.000Z |
| 45 | 737530153 | RT @jlopezbogota: @AntonioSanguino @juanadarco1 El plan del Uribismo Duquismo, llenar el país de Venezolanos para darles cédulas chimbas p... | 2020-11-02T21:43:41.000Z |

Fuente: Elaboración propia.

Teniendo el dataset en el formato deseado, fue cargado en el programa y a través de la librería pandas se convirtió en un dataframe el cual nos permitió un mejor trabajo sobre los datos.

**Figura 25. Creación de dataframe en Python utilizando Pandas.**

```
[47] df = pd.read_excel('/content/data_1.xlsx')
     df
```

| | text | created_at | author_id |
|---|---|---|---|
| 0 | @AT_P8 @CaracolRadio @ClaudiaLopez Cual xenofo... | 2021-03-11T12:59:54.000Z | 1.354621e+18 |
| 1 | @BluRadioCo @ricarospina @hectorriveross @Zule... | 2021-03-11T12:59:51.000Z | 1.222109e+18 |
| 2 | @pilysuspendida @mauro_rivera14 Veneco hpppppp... | 2021-03-11T12:59:47.000Z | 1.443747e+08 |
| 3 | @anibal86co @CaracolRadio @Millonariosco @Clau... | 2021-03-11T12:58:40.000Z | 1.354621e+18 |
| 4 | @Citytv Ahora que no lo deporte que pague por ... | 2021-03-11T12:57:50.000Z | 1.277993e+18 |
| ... | ... | ... | ... |
| 4530 | RT @JULYNEIL1: @PTV_Noticias De Seguro Esos Ex... | 2021-03-15T12:45:36.000Z | 1.249912e+18 |
| 4531 | @PTV_Noticias De Seguro Esos Extranjeros "Vene... | 2021-03-15T12:45:17.000Z | 1.249912e+18 |
| 4532 | EN YOUTUBE IGUAL UN ODIO HACIA VENEZUELA Y MAD... | 2021-03-15T12:42:31.000Z | 4.954423e+08 |
| 4533 | @Alejakls @luisferheroch Está vieja así será d... | 2021-03-15T12:33:30.000Z | 1.370221e+18 |
| 4534 | RT @Gatocabron13: Recibiendo buena verga de ve... | 2021-03-15T12:31:22.000Z | 1.307056e+18 |

4535 rows × 5 columns

Fuente: Elaboración propia.

#### 4.3.3.3 Limpieza de los datos

En primer lugar, se eliminaron todos los datos duplicados, ya que, en Twitter, al momento de retweetear un tweet, se publica como un tweet diferente, pero con el mismo texto.

**Figura 26. Eliminación de datos duplicados.**

```
[ ] df.drop_duplicates(subset ="text",
                       keep = False, inplace = True)
```

Fuente: Elaboración propia.

Para la limpieza de los datos, primero se eliminaron las etiquetas de Retweet “RT” de los tweets, que indican que los tweets fueron compartidos y publicados nuevamente por el usuario. Posteriormente se eliminó el signo @ seguido del usuario mencionado en el texto de un tweet.

**Figura 27. Eliminación de etiquetas “RT” y @ en los tweets.**

```
#Limpieza
remove_rt = lambda x: re.sub('RT @\w+: ',"" ",x)
rr = lambda x: re.sub("(@[A-Za-z0-9]+)", " ",x)
rrt = lambda x: re.sub("\s([@#][\w_-]+)", " " , x)

#df['texto_procesado'] = df['text'].apply(lambda x: limpiar_tokenizar(x))
df['texto_preproc'] = df['text'].map(remove_rt).map(rr).map(rrt)
df
```

| | author_id | text | created_at | texto_preproc |
|---|---|---|---|---|
| 2 | 4.046969e+09 | Luego de varios meses cerrada debido a la pand... | 2020-11-02T21:44:50.000Z | Luego de varios meses cerrada debido a la pand... |
| 6 | 6.222635e+08 | RT @GNB_DeltaD619: #02Nov POLÍTICO-MILITAR VEN... | 2020-11-02T21:44:40.000Z | POLÍTICO-MILITAR VENEZOLANO \n ... |
| 11 | 7.084787e+17 | @Itzeldt2388 @KattyValeroLea1 @elpais_america ... | 2020-11-02T21:44:35.000Z | _america No entiendo, los venezolanos a... |
| 12 | 7.777406e+07 | RT @CarlosNovoT: Colaboracionistas, al lado de... | 2020-11-02T21:44:35.000Z | Colaboracionistas, al lado de eso que llaman ... |

Fuente: Elaboración propia.

Además, para este proceso de limpieza, se hizo uso de las librerías NLTK (Natural Language Toolkit) y Regex, para limpiar las expresiones regulares que vienen incluidas muchas veces en los textos de los tweets. Con el paquete Regex se eliminan expresiones regulares tales como signos de puntuación, números, caracteres especiales, espacios en blanco, urls de páginas webs (palabras que empiezan por "http"), y, además, se convirtió todo el texto en minúsculas para una mejor comprensión de los datos.

**Figura 28. Uso de Regex para eliminar expresiones regulares en los textos.**

```
import re
def limpiar_tokenizar(texto):
    '''
    Esta función limpia y tokeniza el texto en palabras individuales.
    El orden en el que se va limpiando el texto no es arbitrario.
    El listado de signos de puntuación se ha obtenido de: print(string.punctuation)
    y re.escape(string.punctuation)
    '''

    # Se convierte todo el texto a minúsculas
    nuevo_texto = texto.lower()

    # Eliminación de páginas web (palabras que empiezan por "http")
    nuevo_texto = re.sub('http\S+', ' ', nuevo_texto)

    # Eliminación de signos de puntuación
    regex = '[\\¡\\!\\"\\#\\$\\%\\&\\'\\(\\)\\*\\+\\,\\-\\.\\/\\:\\;\\<\\=\\>\\¿\\?\\\\[\\\\\\]\\^_\\`\\{\\|\\}\\~]'
    nuevo_texto = re.sub(regex , ' ', nuevo_texto)

    # Eliminación de números
    nuevo_texto = re.sub("\d+", ' ', nuevo_texto)

    #Eliminación de caracteres especiales
    nuevo_texto = re.sub('\s([@#][\w_-]+)', ' ', nuevo_texto)
    re.sub(r'!+(?=.*\!)','',nuevo_texto)

    # Eliminación de espacios en blanco múltiples
    nuevo_texto = re.sub("\\s+", ' ', nuevo_texto)

    # Tokenización por palabras individuales
    nuevo_texto = nuevo_texto.split(sep = ' ')

    # Eliminación de tokens con una longitud < 2
    nuevo_texto = [token for token in nuevo_texto if len(token) > 1]

    return(nuevo_texto)
```

Fuente: Elaboración propia.

También se utilizó Regex para eliminar todos los emojis que fueron usados por los usuarios en cada uno de los tweets.

**Figura 29. Eliminación de emojis.**

```
import re
def deEmojify(text):
    regrex_pattern = re.compile(pattern = "["
        u"\U0001F600-\U0001F64F"  # emoticons
        u"\U0001F300-\U0001F5FF"  # symbols & pictographs
        u"\U0001F680-\U0001F6FF"  # transport & map symbols
        u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
        u"\U00002500-\U00002BEF"  # chinese char
        u"\U00002702-\U000027B0"
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        u"\U0001f926-\U0001f937"
        u"\U00010000-\U0010ffff"
        u"\u2640-\u2642"
        u"\u2600-\u2B55"
        u"\u200d"
        u"\u23cf"
        u"\u23e9"
        u"\u231a"
        u"\ufe0f"  # dingbats
        u"\u3030"
                          "]+", flags = re.UNICODE)
    return regrex_pattern.sub(r'',text)
```

Fuente: Elaboración propia.

En cuanto a la librería NLTK, esta permitió eliminar stopwords (palabras vacías) o palabras sin significado cuya función es modificar o acompañar otras palabras. Algunas stopwords son los artículos, pronombres, preposiciones, adverbios, entre otras palabras. Esto se realiza con el fin de tener un mejor procesamiento de los datos durante la etapa del análisis de sentimiento.

**Figura 30. Eliminación de stopwords utilizando librería NLTK.**

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import SnowballStemmer
from nltk.corpus import stopwords

nltk.download('stopwords')
nltk.download('punkt')

stop_words = set(stopwords.words('spanish'))
stop_words = stop_words.union(set(['me','si', 'le', 'da', 'mi', 'su', 'qué','esto','con','tu','si', 'te', 'tus', 'a','se',
                                   'por','lo','mis','ha','el','tú','tu','fue','he', 'ya', 'un', 'una', 'es','del', 'las',
                                   'los', 'en', 'que', 'y', 'la','de','no','lo','si','era','pero','se','esos','has','es']))

def fast_preproc(text):
  words = word_tokenize(text)
  words = [word for word in words if not word in stop_words]
  try:
    text = " ".join(str(word) for word in words)
  except Exception as e:
    print(e)
    pass
  return text
```

Fuente: Elaboración propia.

#### 4.3.3.4 Aplicación de formato de los datos

Posterior a la construcción del conjunto de datos fue necesarios aplicar y cambiar algunas columnas consecuencia de que al limpiar los datos y aplicar la librería re nos devuelve datos en formato de contenedor de listas llegando a dificultar la aplicación del análisis.

Así se almacenaron los datos procesados.

**Figura 31. Columna de datos procesados.**

| |
|---|
| [no, es, xenofobia, es, realidad, ahora, mante... |
| [ahora, que, no, lo, deporte, que, pague, por,... |

Fuente: Elaboración propia.

Para convertir el formato fue necesario aplicar una función corta que gracias al método toString() transformó una lista de array en una sola cadena de texto por cada array de la lista.

**Figura 32. Conversión de lista array en una sola cadena de texto.**

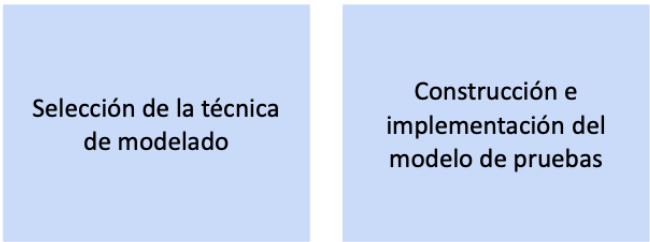
```
[ ] procT = []
    for i in df['texto_preproc']:
      procT.append(listToString(i))
    procT

     'ver imbécil si el veneco disparo contra el policia es porque hay agentes en las calles es que el veneco no le teme matar porque va rec
     'triple hp veneco',
     'maldito veneco ojalá se pudra en la cárcel no llegue un enérgico juez lo deje libre',
     'veneco hijo de puta',
     'que tal bogotá si organizamos un plantón en la plaza mayor le éximos al gobierno sacar sin contemplaciones cuánto veneco aparezca en l
     'vieja inepta la solución es sacar todo veneco que no aporte nada en el país ayudar al colombiano los de acá si necesitamos esos progra
     'yo digo "qué veneco" unas veces al día',
     'entonces el asesino del policía es veneco duquenomasvenecos',
     'veneco el problema realmente nunca fueron los peninsulares durante la guerra muerte los españoles que peleasen por la republica eran b
     'me avisa cuando esa sarta de cabrones decida conversar como gente civilizada lo que no hay en venezuela es pueblo ni sociedad el venec
     'el sr los venezolanos les da trabajo garantías al colombiano la espalda no hay trabajo el veneco trabaja por menos plata los quiere na
     'tu perfil dice economista no periodista tratas de escribir como tal pero tu ignorancia te delata más millones de colombianos hicieron
     'sapegato tierrúo nveneco profundo entrecerró los ojos prensó los labios ncésar carenero chocolate oscuro',
     'yo si veo que un veneco le pega un policía conmigo también defiendo los policías punto',
     'venezolano los venezolanos son aquellos ciudadanos provenientes de venezuela país ubicado en la parte septentrional de américa del sur
     'veneco muchos de ellos solo vienen ganarsela fácil hacer daño gente trabajadora miserables de mier',
     'si lees esto veneco malandrín no mates para robar no mates no mates',
     'veneco claro pero está rebelión fue más antipeninsular que anticriolla por lo menos se pretendía',
     'veneco tenía que ser el hijo de puta',
     'arrugó la nariz cuando se acercó tanto de esa forma quién se cree para intentar intimidarlo así vos qué noops se le salió lo veneco pe
     'veneco muchos de ellos solo vienen ganarsela fácil hacer daño gente trabajadora miserables de mier aclaro no son todos',
     'la mayoría de mis problemas es porque soy veneco',
     'duque es mas veneco que colombiano',
     'muy bien por el valiente policía que dió de baja al bandido veneco capturó al acecino de de su compañero un saludo fuera venecos gonor
     'tanto que joden requisando gente buscando malandros hasta que se encuentran uno veneco',
     'veneco tenía que ser el hijo de puta',
     'el veneco magallanes',
     'señora claudia lópez con todo respeto pero se imagina si partir de ahora nosotros los comunicadores nos referimos usted en de esta mar
     'veneco shit',
     'el gol de la tambien sale de pelota parada pero no veo nadie webeando al veneco por eso',
     'el hpta mató al policía era veneco sorpresa en redes no les gusta digan la nacionalidad porque alegan eso es xenofobia tal claro como
     'si fuera un colombiano el que la estuviese cortando estría preso pero como un veneco solo es "llamado de atención" por qué todo el mur
     'veneco comparto aunque las consecuencias no me parecieron del todo malas sino ambivalentes ya que muchos negros indígenas se unieron l
     'betancourtm no vale pero si era un veneco forma parte de la cultura venezolana admirar la belleza de la mujer si me gusta se lo digo r
```

Fuente: Elaboración propia.

### 4.3.4 Fase 4. Modelado

En esta fase se escogen y se aplican las técnicas de modelado y algoritmos que más se adecuen a el problema (sin restricción de cantidad) se establecen los parámetros y valores óptimos a utilizar. Ver la figura 33, refleja las actividades que componen la fase de modelado.

**Figura 33:** Fase 4 Modelado




Fuente: Elaboración propia.

#### 4.3.4.1 Selección de la técnica de modelado

En esta primera etapa de la fase cuatro, se optó por utilizar la técnica de análisis sentimental sobre los datos extraídos con la finalidad de obtener el tipo de polaridad en cada uno de los tweets de estudio. Este proceso comenzó con el uso del API v2 de Twitter y *Postman,* que funcionó como herramienta para la extracción de los datos que conforman el modelo, para la limpieza de los tweets fue necesario el uso del módulo re que funciona para identificar caracteres especiales, palabras y signos

de puntuación a partir del uso de expresiones regulares, adicionalmente se utilizó la biblioteca *NLTK* para lematizar el texto alojado en los tweets y la eliminación de stopwords; por último se seleccionó la técnica de análisis sentimental la cual busca determinar la polaridad de los tweets extraídos para ello, fue necesario el uso de la librería VADER la cual nos proporcionó un modelo de análisis de sentimientos que fue aplicado sobre el dataset, este modelo trabaja sobre unos rangos entre -1,1; donde aquellos tweets que den unos resultados más cercano a -1 corresponden a una polaridad negativa mientras que aquellos más cercanos a uno representan una polaridad positiva, siendo 0 neutral.

#### 4.3.4.2 Construcción e implementación del modelo de pruebas

Para realizar el análisis sentimental sobre los tweets extraídos se hizo uso del VADER (Valence Aware Dictionary and Sentiment Reasoner), una librería de python específicamente para realizar análisis de sentimiento en redes sociales [33]. VADER es una librería Open Source que implementa una combinación de reglas y léxicos de palabras relacionadas con un sentimiento para de esta manera detectar la métrica de polaridad (positiva, negativa, neutral, compound) en un texto y determinar así el sentimiento general relacionado [33]. El modelo VADER tiene un buen comportamiento gracias a una gran cantidad de léxicos de palabra que maneja los cuales fueron extraídas de reseñas y clasificadas manualmente por personas.

Entre las características que más destacan a VADER para el análisis de sentimientos son las tres reglas que implementa:

- Incrementa la intensidad de la polaridad si reconoce un signo de exclamación. Ej.: *¡Cuanto más la escucho más me gusta!*

- Incrementa la intensidad de la polaridad si reconoce palabras en mayúscula dentro de textos en minúscula. Ej.: *nunca escucha NADA*

- Incrementar la intensidad de la polaridad si reconoce palabras que indican intensidad. Ej.: *Ella es extremadamente alto*

Para hacer el análisis de sentimiento se programó un script en Python utilizando la librería de VADER modificada. La siguiente imagen contiene un extracto del código implementado para evaluar los tweets.

El modelo implementado trae como resultado de análisis cuatro etiquetas o métricas con sus respectivos valores: polaridad positiva, polaridad negativa, polaridad neutra y *compound* que representa la suma de todas las anteriores. A través de condicionales y según los valores obtenidos por cada etiqueta son asignados a una serie de listas para posteriormente ser representado de manera gráfica y porcentual.

**Figura 34. Algoritmo para clasificar tweets por polaridad.**

```
for tweet in tweets:


    tweet_list.append(tweet)
    analysis = TextBlob(tweet.text)
    entidades.append(tweet.entities)
    ubicacion.append(tweet.user.location)
    score = SentimentIntensityAnalyzer().polarity_scores(tweet.text)
    #diccionarios
    neg = score['neg']
    neu = score['neu']
    pos = score['pos']
    comp = score['compound']
    polaridad += analysis.sentiment.polarity
```

Fuente: Elaboración propia.

**Figura 35. Condiciones para asignar una clasificación al texto.**

```
    if neg > pos:
        negativo_list.append(tweet)
        negativo += 1

    elif pos > neg:
        positivo_list.append(tweet)
        positivo += 1

    elif pos == neg:
        neutral_list.append(tweet)
        neutral += 1

positivo = porcentage(positivo, noOfTweet)
negativo = porcentage(negativo, noOfTweet)
neutral = porcentage(neutral, noOfTweet)
polaridad = porcentage(polaridad, noOfTweet)
```

Fuente: Elaboración propia.

El porcentaje total de tweets obtenidos para cada etiqueta se determina dividiendo el total de tweets encontrados por etiqueta entre el total de tweets dentro del dataset.

*#total tweets por etiqueta / #total de tweets en el dataset*

# 5. RESULTADOS

Los resultados arrojados de este estudio se dividieron en 3 secciones, en donde la primera sección corresponde al análisis de sentimientos general aplicado a todos los datos obtenidos inicialmente y en la segunda se muestran los resultados obtenidos para los tweets recopilados relacionados con las entidades identificadas. La siguiente sección corresponde a las respectivas evaluaciones del modelo, la primera con respecto a la validación de la academia en donde se mide la variable relacionada a la precisión del modelo. Mientras que la siguiente evaluación abarca la validación estática que se realizó a través de una entrevista al representante del grupo de investigación Osiris&Bioaxis de la universidad el bosque, para medir la utilidad del modelo para trabajos futuros.

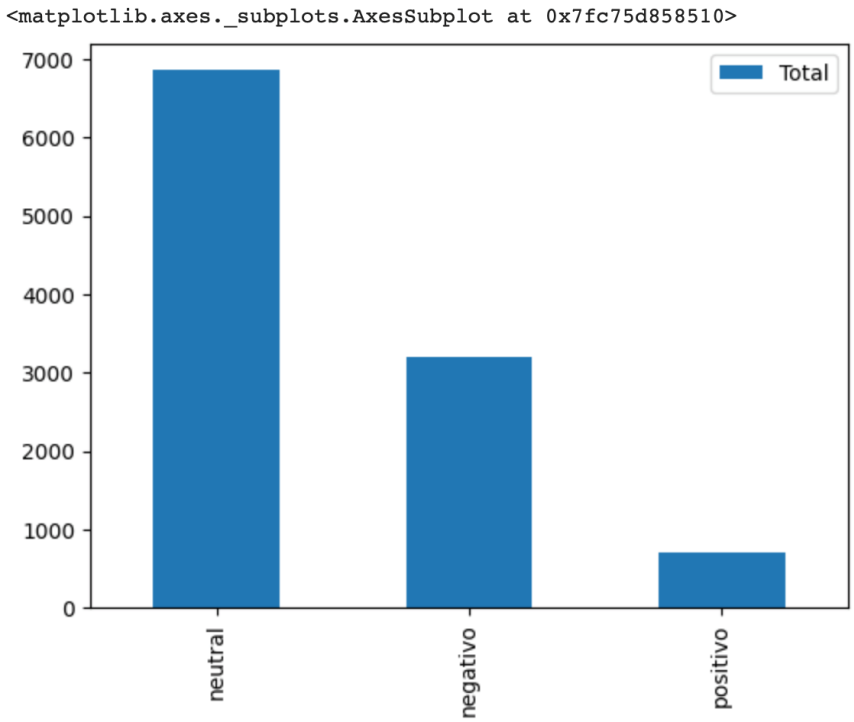
## 5.1 Análisis de sentimientos general para todos los datos

Inicialmente, se realizó un análisis de sentimientos para todos los datos recopilados en la extracción de datos, con un total de 32.454, sin embargo, luego del preprocesamiento en donde se limpiaron los textos y eliminaron los tweets duplicados por retweet, quedó finalmente un conjunto de datos de 10.764.
Al total de 10.764 tweets que quedaron después del preprocesamiento se les aplicó el modelo VADER, clasificándolos según la polaridad presentada en sus textos. Como resultado se obtuvieron 6854 neutrales, 3203 negativos, 707 positivos como se observa en la siguiente gráfica.

**Figura 36. Porcentajes de tweets por cada etiqueta en gráfica PieChart**

Fuente: Elaboración propia.

**Figura 37. Cantidad de tweets por etiqueta.**




```
count_values_in_column(df,"sentiment")
```

| | Total |
|---|---|
| **neutral** | 6854 |
| **negativo** | 3203 |
| **positivo** | 707 |

Fuente: Elaboración propia.

Se agregaron al dataframe las columnas correspondientes a la polaridad de cada dato, junto a la subjetividad, el resultado del análisis de sentimiento (si es neutral, negativo o positivo) y los valores arrojados para cada una de estas categorías.

**Figura 38. Dataframe con columnas de etiquetado.**

| | text | created_at | author_id | id | lang | texto_preproc | texto_procesado | polaridad | subjectividad | sentiment | neg | neu | pos | compound |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | @AT_P8 @CaracolRadio @ClaudiaLopez Cual xenofo... | 2021-03-11T12:59:54.000Z | 1.354621e+18 | 1369996504774115328 | es | cual xenofobia ahora el veneco mantenido en la... | xenofobia ahora veneco mantenido cárcel impuestos | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 1 | @EluRadioCo @ricarospina @hectorriveros @Zule... | 2021-03-11T12:59:51.000Z | 1.222109e+18 | 1369996490593210368 | es | nla gran diferencia es que ni joe biden ni el ... | nla gran diferencia joe biden santísimo papa e... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 2 | @pilysuspendida @mauro_rivera14 Veneco hpppppp... | 2021-03-11T12:59:47.000Z | 1.443747e+18 | 1369996475208511488 | es | rivera veneco hpppppppppp | rivera veneco hpppppppppp | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 3 | @aniball6co @CaracolRadio @Millonariosco @Clau... | 2021-03-11T12:58:40.000Z | 1.354621e+18 | 1369961938881412 | es | no es xenofobia es realidad ahora mantener el ... | xenofobia realidad ahora mantener veneco cárce... | 0.0 | 0.00 | negativo | 0.087 | 0.866 | 0.047 | -0.2500 |
| 4 | @Cityfv Ahora que no lo deporte que pague por ... | 2021-03-11T12:57:50.000Z | 1.277993e+18 | 1369995985494171648 | es | ahora que no lo deporte que pague por las cosa... | ahora deporte pague cosas hacen quieren quedar... | 0.0 | 0.00 | negativo | 0.071 | 0.929 | 0.000 | -0.2960 |
| 6 | @NoticiasCaracol Ay, pero esa no fue la entrad... | 2021-03-11T12:57:29.000Z | 2.851331e+08 | 1369995896805539840 | es | ay pero esa no fue la entrada tanto veneco que... | ay entrada veneco aplaudieron mundo entero | 0.0 | 0.00 | negativo | 0.128 | 0.872 | 0.000 | -0.2960 |
| 7 | @martinvillada1 @CaracolRadio @Miguel_2821 @Cl... | 2021-03-11T12:57:17.000Z | 1.354621e+18 | 1369995845723164672 | es | ver imbécil si el veneco disparo contra el pol... | ver imbécil veneco disparo policía agentes cal... | 0.0 | 0.00 | negativo | 0.056 | 0.914 | 0.030 | -0.2500 |
| 8 | @pilysuspendida Triple HP veneco | 2021-03-11T12:53:35.000Z | 1.478835e+18 | 1369994912977084416 | es | triple hp veneco | triple hp veneco | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 9 | @pilysuspendida Maldito veneco ojalá se pudra ... | 2021-03-11T12:52:58.000Z | 3.392138e+08 | 1369994758844743680 | es | maldito veneco ojalá se pudra en la cárcel no ... | maldito veneco ojalá pudra cárcel llegue enérg... | 0.0 | 0.00 | negativo | 0.121 | 0.879 | 0.000 | -0.2960 |
| 11 | @pilysuspendida Veneco hijo de puta! | 2021-03-11T12:49:50.000Z | 1.287769e+18 | 1369993970957320192 | es | veneco hijo de puta | veneco hijo puta | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 12 | @rcnradio que tal Bogotá si organizamos un plá... | 2021-03-11T12:49:48.000Z | 1.352327e+18 | 1369993961905983488 | es | que tal bogotá si organizamos un plantón en la... | tal bogotá organizamos plantón plaza mayor éxi... | 0.0 | 0.00 | negativo | 0.077 | 0.923 | 0.000 | -0.5574 |
| 14 | @mluciaramirez Vieja inepta, la solución es sa... | 2021-03-11T12:40:20.000Z | 1.280422e+08 | 1369991578694127616 | es | vieja inepta la solución es sacar todo veneco ... | vieja inepta solución sacar veneco aporte país... | 0.0 | 0.00 | negativo | 0.051 | 0.949 | 0.000 | -0.2960 |
| 15 | Yo digo: "qué veneco" unas 20 veces al día | 2021-03-11T12:39:36.000Z | 7.842839e+07 | 1369991393314213888 | es | yo digo "qué veneco" unas veces al día | digo " veneco " unas veces día | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 16 | Entonces el asesino del policía es veneco ????... | 2021-03-11T12:38:51.000Z | 1.242567e+18 | 1369991207632441344 | es | entonces el asesino del policía es veneco duqu... | entonces asesino policía veneco duquenomasvenecos | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 18 | @elwebo696 @Juan_veneco El problema realmente ... | 2021-03-11T12:33:58.000Z | 1.307757e+18 | 1369989977149427712 | es | veneco el problema realmente nunca fueron los ... | veneco problema realmente nunca peninsulares g... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 19 | @ajcaleca Me avisa cuando esa sarta de cabrone... | 2021-03-11T12:33:49.000Z | 1.327814e+18 | 1369989941044871168 | es | me avisa cuando esa sarta de cabrones decida c... | avisa sarta cabrones decida conversar gente ci... | 0.0 | 0.00 | negativo | 0.150 | 0.850 | 0.000 | -0.5267 |
| 20 | @lcvelez @ClaudiaLopez El sr @IvanDuque a los... | 2021-03-11T12:33:08.000Z | 3.996662e+08 | 1369989769216860160 | es | el sr los venezolanos les da trabajo garantías... | sr venezolanos trabajo garantías colombiano es... | 0.0 | 0.00 | negativo | 0.051 | 0.949 | 0.000 | -0.2960 |
| 21 | @SdrodriguezT Tu perfil dice economista no per... | 2021-03-11T12:31:53.000Z | 1.096597e+18 | 1369989454073638912 | es | tu perfil dice economista no periodista tratas... | perfil dice economista periodista tratas escri... | 0.0 | 0.00 | negativo | 0.061 | 0.939 | 0.000 | -0.4184 |
| 22 | "... Sapegato tierrüo."#VirVeneco profundo. En... | 2021-03-11T12:29:02.000Z | 1.369787e+18 | 1369988735186657280 | es | sapegato tierrüo nveneco profundo entrecerró l... | sapegato tierrüo nveneco profundo entrecerró o... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 23 | @ELTIEMPO Yo si veo que un veneco le pega a un... | 2021-03-11T12:28:38.000Z | 4.236026e+08 | 1369988636226228224 | es | yo si veo que un veneco le pega un policía con... | veo veneco pega policía conmigo defiendo polic... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 24 | Venezolano: Los venezolanos son aquellos ciuda... | 2021-03-11T12:26:47.000Z | 1.030531e+18 | 1369988169689600000 | es | venezolano los venezolanos son aquellos ciudad... | venezolano venezolanos aquellos ciudadanos pro... | -0.4 | 0.55 | negativo | 0.099 | 0.901 | 0.000 | -0.5267 |
| 25 | RT @FerchoAragon14: @Cityfv Veneco H P, muchos... | 2021-03-11T12:26:45.000Z | 2.851331e+08 | 1369988163440541184 | es | veneco muchos de ellos solo vienen ganarsela f... | veneco solo vienen ganarsela fácil hacer daño ... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 26 | @PoliciaBogota Si lees esto veneco malandrín: ... | 2021-03-11T12:22:25.000Z | 1.872705e+09 | 1369987069779582976 | es | si lees esto veneco malandrín no mates para ro... | lees veneco malandrín mates robar mates mates | 0.0 | 0.00 | negativo | 0.452 | 0.548 | 0.000 | -0.8439 |
| 27 | @Juan_veneco @VirgiliusNovus Claro, pero está ... | 2021-03-11T12:21:01.000Z | 1.253657e+18 | 1369986718292709376 | es | veneco claro pero está rebelión fue más antipe... | veneco claro rebelión antipeninsular anticriol... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 29 | RT @Miguel_2821: @ELTIEMPO @scrappymas Veneco ... | 2021-03-11T12:19:55.000Z | 4.088806e+08 | 1369986440768159744 | es | veneco tenía que ser el hijo de puta | veneco ser hijo puta | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 30 | Arrugó la nariz cuando se acercó tanto de esa ... | 2021-03-11T12:19:19.000Z | 1.369787e+18 | 1369986291601973248 | es | arrugó la nariz cuando se acercó tanto de esa ... | arrugó nariz acercó forma quién cree intentar ... | 0.0 | 0.00 | negativo | 0.109 | 0.891 | 0.000 | -0.6298 |
| 32 | @Cityfv Veneco H P, muchos de ellos solo viene... | 2021-03-11T12:19:02.000Z | 1.183754e+18 | 1369986218918809600 | es | veneco muchos de ellos solo vienen ganarsela f... | veneco solo vienen ganarsela fácil hacer daño ... | 0.0 | 0.00 | negativo | 0.099 | 0.901 | 0.000 | -0.2960 |
| 34 | La mayoría de mis problemas es porque soy..... | 2021-03-11T12:17:21.000Z | 6.093579e+08 | 1369985794522415104 | es | la mayoría de mis problemas es porque soy veneco | mayoría problemas veneco | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 35 | Duque es mas veneco que colombiano | 2021-03-11T12:15:43.000Z | 1.343578e+18 | 1369985384634060800 | es | duque es mas veneco que colombiano | duque mas veneco colombiano | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 37 | @NoticiasCaracol muy bien por el valiente poli... | 2021-03-11T12:13:00.000Z | 1.352327e+18 | 1369984703198023680 | es | muy bien por el valiente policía que dió de ba... | bien valiente policía dió baja bandido veneco ... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 39 | Tanto que joden requisando gente buscando mala... | 2021-03-11T12:11:54.000Z | 2.837590e+08 | 1369984423555436544 | es | tanto que joden requisando gente buscando mala... | joden requisando gente buscando malandros encu... | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 40 | @ELTIEMPO @scrappymas Veneco tenía que ser el ... | 2021-03-11T12:11:09.000Z | 1.271358e+09 | 1369984234123886592 | es | veneco tenía que ser el hijo de puta | veneco ser hijo puta | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |
| 41 | @ms30863 @IvanCepedaCast El veneco Magallanes 😂😂😂 | 2021-03-11T12:09:10.000Z | 3.527507e+08 | 1369983736813658112 | es | el veneco magallanes | veneco magallanes | 0.0 | 0.00 | neutral | 0.000 | 1.000 | 0.000 | 0.0000 |

Fuente: Elaboración propia.

Se obtuvo también la suma de todos los valores anteriores, dentro de la columna compound, y se generó un reporte estadístico en donde se observa la suma total (**count)**, el promedio (**mean)** la desviación estándar (**std**), el elemento con el valor más bajo (**min**) y los cuartiles en valores porcentuales (**25%, 50%, 75%, max**).

**Figura 39. Datos estadísticos de la métrica compound.**

```
df['compound'].describe()
```

```
count    10764.000000
mean        -0.101257
std          0.246730
min         -0.970500
25%         -0.296000
50%          0.000000
75%          0.000000
max          0.861900
Name: compound, dtype: float64
```

Fuente: Elaboración propia.

Se encontró una media en la polaridad de -0.101 lo que representa que los sentimientos expresados por los colombianos en sus tweets son en general de polaridad negativa o neutral; al mismo tiempo observamos que la polaridad con el valor mínimo encontrada dentro del conjunto de datos fue de valor -0.97 lo que representa un tweet negativo en su totalidad, por el contrario la polaridad con el

valor máximo encontrada dentro del conjunto de datos de valor 0.86 lo que representa un tweet totalmente positivo. Con referencia a los cuartiles encontramos que el 25% o menos presentan una polaridad negativa y el 75% o más presentan polaridad entre neutral y positiva.

A partir de los sentimientos resultantes, se construyó una colección de datos correspondiente a cada valor, neutral, positivo y negativo con el fin de generar nubes de palabras para cada una de las polaridades. Posteriormente se generó una nube de palabras para todas las palabras y otras tres para cada uno de estos valores, de tal manera que puedan visualizar las palabras más repetidas categorizadas por sentimiento.

**Figura 40. Creación de dataframes distinguidos por etiqueta.**

Creación de Dataframes para sentimientos positivos, negativos y neutrales

```
df_negativo = df[df["sentiment"]=="negativo"]
df_positivo = df[df["sentiment"]=="positivo"]
df_neutral = df[df["sentiment"]=="neutral"]
```

Fuente: Elaboración propia.

**Figura 41. Nube de palabras del total de tweets.**

Wordcloud para todos los tweets

```
create_wordcloud(df["texto_procesado"].values)
```

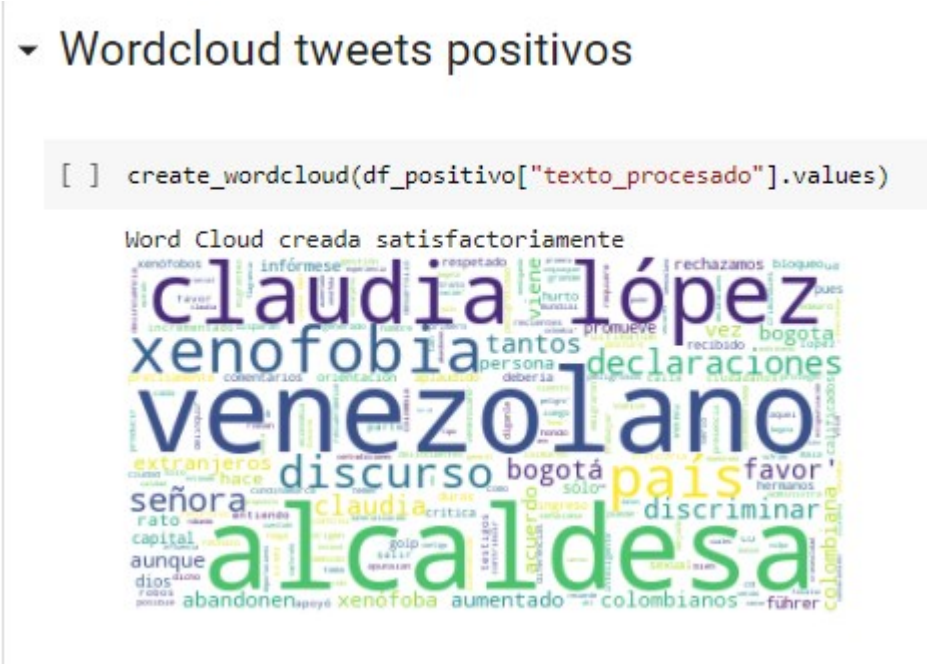
Word Cloud creada satisfactoriamente

Fuente: Elaboración propia.

En esta primera nube de palabras se pueden observar las palabras mayor frecuencia en todos los tweets procesados, en donde se destacan los términos de

búsqueda principales como “venezolano”, seguido de la entidad política más mencionada que fue la alcaldesa de Bogotá Claudia López.

**Figura 42. Nube de palabras de los tweets con etiquetado positivo.**




Fuente: Elaboración propia.

En la nube de palabras para los tweets clasificados positivamente, encontramos nuevamente la palabra clave “venezolano”, con esto se puede inferir que este adjetivo es el más utilizado para referirse a los migrantes venezolanos en el país.

**Figura 43. Nube de palabras de los tweets con etiquetado negativo.**

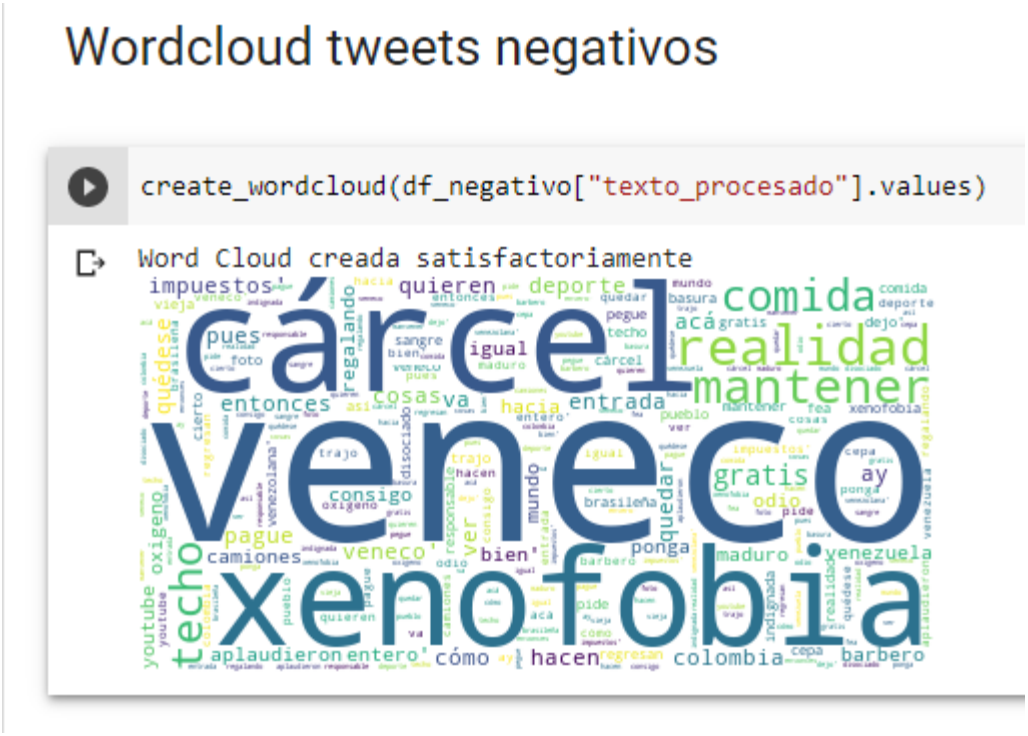
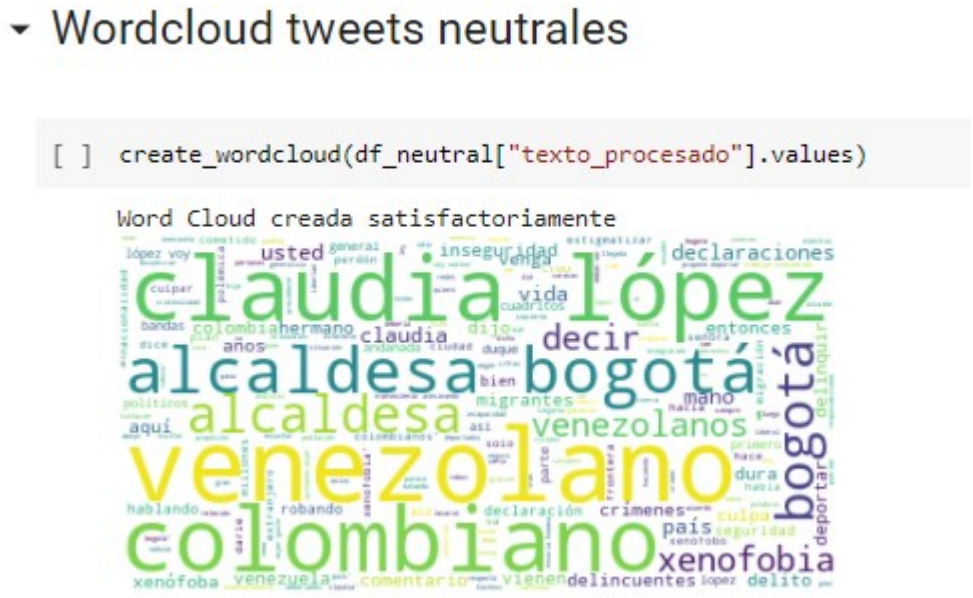

Wordcloud tweets negativos
create_wordcloud(df_negativo["texto_procesado"].values)
Word Cloud creada satisfactoriamente
cárcel
veneco
xenofobia
realidad
mantener
comida
techo
gratis
entonces
cosas
entrada
hacia
pues
igual
quieren
deporte
impuestos
consigo
pague
camiones
veneco
bien
ponga
maduro
venezuela
aplaudieron
entero
cómo
hacen
colombia
barbero
ay
odio
acá
basura
pegue
sangre
foto
cárcel
quedar
mundo
trajo
hacen
brasileña
pide
youtube
oxigeno
venezolana
disociado
responsable
ver
entrada
regalando
quédese
vieja
cierto
regresan
fea
xenofobia
pueblo
ver
cepa
indignada
realidad
quédese
venezuela
regalando
aplaudieron
responsable
entero
colombia
cepa
gratis
barbero
quieren
pueblo
techo
va
aplaudieron
responsable
regalando
aplaudieron
responsable
regalando
aplaudieron
responsable
regalando

**Figura 44. Nube de palabras de los tweets con etiquetado neutral.**

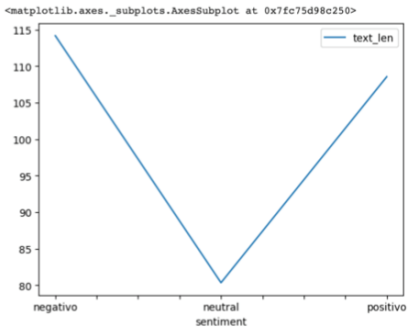



Fuente: Elaboración propia.

En la nube de palabras correspondiente a los tweets clasificados como neutrales, encontramos términos imparciales como venezolano, colombiano, Bogotá, Claudia López, y entre otras palabras que no tienen una tendencia ni positiva ni negativa.

Se realizó un análisis sobre el texto en los tweets con la finalidad de encontrar la longitud promedio en cada uno de los publicados clasificados por cada etiqueta, de la misma manera para determinar la cantidad promedio de palabras en las publicaciones. Para el primer caso se determinaron unas longitudes promedio de 115 caracteres para los tweets negativos, seguido de 108 para los tweets de polaridad positiva y por último un promedio de 80 caracteres para los tweets neutrales.

**Figura 45. Longitud promedio de los tweets por cada etiquetado.**

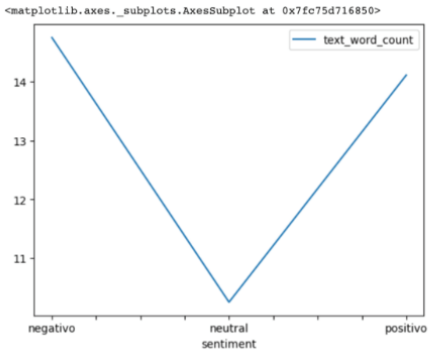



| sentiment | text_len |
|---|---|
| negativo | 114.18 |
| neutral | 80.35 |
| positivo | 108.56 |

Fuente: Elaboración propia.

Para el segundo caso se realizó un conteo del promedio de palabras utilizadas en los tweets clasificados por cada polaridad, para el caso de los tweets negativos se utilizan alrededor de 15 palabras en promedio, los positivos por su parte utilizan alrededor de 14 palabras, mientras que los neutrales solo los conforman en promedio 10 palabras.

**Figura 46. Cantidad promedio de palabras en los tweets según etiquetado.**




| sentiment | text_word_count |
|---|---|
| negativo | 14.75 |
| neutral | 10.25 |
| positivo | 14.11 |

Fuente: Elaboración propia.

Se realizó un análisis de N-gramas clasificando en series de dos y tres palabras con el objetivo de determinar las palabras vecinas con mayor frecuencia y determinar las series de palabras con más recurrencia dentro de los tweets analizados.

Para N-grama de 2 palabras, encontramos en su mayoría palabras neutrales, entre las que destacan: claudia López, ser veneco, venezolanos Colombia, venezolanos delincuentes, entre otros.

**Figura 47. N-grama de dos palabras generado a partir del total de tweets.**

```
('claudia lópez', 81),
('cantor pueblo', 77),
('hace años', 69),
('cubanos venezolanos', 69),
('nicolás maduro', 66),
('ee uu', 65),
('mandela venezolano', 64),
('venezolanos colombia', 60),
('ser veneco', 53),
('palabra veneco', 53),
('mayoría venezolanos', 52),
('ciudadanos venezolanos', 52),
('gobierno venezolano', 50),
('hermanos venezolanos', 48),
('declaraciones alcaldesa', 48),
('migración colombia', 47),
('debe ser', 44),
('consulta popular', 38),
('activista político', 38),
('colombianos venezolanos', 37),
('venezolanos colombianos', 36),
('venezolanos delincuentes', 36),
```

Fuente: Elaboración propia.

Para el N-gramas de 3 palabras se encuentra en su mayoría la presencia de palabras de carácter o polaridad neutrales, destaca en este la presencia de menciones sobre la alcaldesa de Bogotá.

**Figura 48. N-grama de tres palabras generado a partir del total de tweets.**

```
('declaraciones alcaldesa bogotá', 40),
('alcaldesa bogotá venezolanos', 27),
('alí primera voz', 24),
('voz inspiración aliento', 24),
('inspiración aliento pueblo', 24),
('aliento pueblo venezolano', 24),
('músico cantante compositor', 24),
('ely rafael primera', 24),
('personas humildes quieran', 24),
('humildes quieran alguna', 24),
('quieran alguna ayuda', 24),
('alguna ayuda económica', 24),
('ayuda económica pequeña', 24),
('económica pequeña concluir', 24),
('pequeña concluir pequeño', 24),
('concluir pequeño negocio', 24),
('leopoldo lópez mandela', 23),
('oficiales venezolanos alertaron', 23),
('venezolanos alertaron nicolás', 23),
('alertaron nicolás maduro', 23),
('venezolano alí primera', 23),
('lópez mandela venezolano', 22),
('pueblo venezolano alí', 22),
('cantante compositor poeta', 22),
('cantante eterno alí', 21),
('eterno alí primera', 21),
```

Fuente: Elaboración propia.

## 5.2 Análisis de sentimientos para los datos relacionados a las entidades

Luego de realizar el procesamiento y análisis de la información de los tweets extraídos, se construyó un nuevo conjunto de datos en donde se almacenaron únicamente todos aquellos Tweets en donde se mencionó alguna entidad política, a través de la mención de una cuenta distrital de Bogotá o del país completo. Entre estas se encuentran Claudia López, alcaldesa de Bogotá, GobiernoBTA, Duque, Iván Duque, Gustavo Petro, Migración Colombia. De esta manera se obtuvieron 962 tweets con dichas menciones.

**Figura 49. Dataset de las entidades identificadas.**

```
[3] df = pd.read_excel('entities-1.xlsx')
```

```
[6] df
```

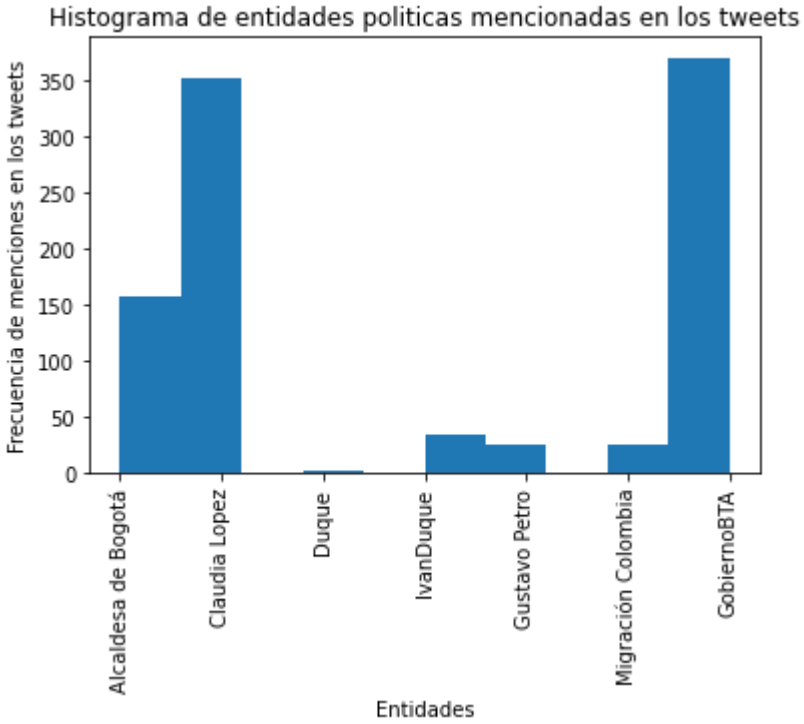
| | Entidad | Tweet |
|---|---|---|
| 0 | Alcaldesa de Bogotá | RT @Javier_Chirinos: Según la Alcaldesa de Bog... |
| 1 | Claudia Lopez | RT @NoticiasCaracol: “Deberíamos deportarlos s... |
| 2 | Claudia Lopez | RT @BluRadioCo: Claudia López propone deportar... |
| 3 | Claudia Lopez | RT @JuanCamilo: #InseguridadEnBogotaEs consecu... |
| 4 | Claudia Lopez | RT @tavo2366: Según Claudia López, éramos un r... |
| ... | ... | ... |
| 957 | GobiernoBTA | @AlcaldChapinero @GobiernoBTA Vengan, ya que e... |
| 958 | GobiernoBTA | @LuisErnestoGL @GobiernoBTA @SaninPazC Totalme... |
| 959 | GobiernoBTA | @SaninPazC @GobiernoBTA Gracias por la denunci... |
| 960 | GobiernoBTA | @AlcaldChapinero @GobiernoBTA Gracias! También... |
| 961 | GobiernoBTA | RT @camilocogollo11: Hoy terminamos nuestras v... |

962 rows × 2 columns

Fuente: Elaboración propia.

Posteriormente se utilizó un histograma que representa gráficamente la distribución de frecuencias de cada una de las variables de entidades mencionadas en los tweets.

**Figura 50. Histograma cantidad de menciones de entidad por tweet.**




Fuente: Elaboración propia.

A partir de esta gráfica, podemos observar que las entidades con más frecuencia de menciones en los tweets recolectados son Claudia López y Gobierno BTA, seguido de alcaldesa de Bogotá. Esto se debe principalmente a los rangos de fecha establecidos para la búsqueda y recolección de datos, ya que la mayoría de estas fechas están relacionadas a hechos delincuenciales ocurridos en Bogotá. Estos días fueron tomados como referencia de búsqueda puesto que, a partir de estos hechos ocurridos en la ciudad de Bogotá, hubo una alta interacción en las redes sociales e intercambio de opiniones por parte de los usuarios, en donde mayormente se mencionó a la alcaldesa de la ciudad de Bogotá.

Posterior a esto, se agregó la columna de *sentiment* al conjunto de datos de las entidades mencionadas. Esta columna corresponde a la clasificación otorgada a cada tweet a partir del análisis de sentimientos realizado anteriormente. De esta manera obtuvimos todos los datos clasificados por neutrales, positivos, o negativos. A continuación, se agruparon los tweets por estas polaridades y se construyeron tres subconjuntos de datos para cada una de las categorías.

**Figura 51. Creación dataframes por cada etiqueta de las entidades identificadas.**

```
[ ] neutral = polaridades.query("sentiment == 'neutral'")
    positivo = polaridades.query("sentiment == 'positivo'")
    negativo = polaridades.query("sentiment == 'negativo'")
```

```
[ ] neutral
```

| | Entidad | texto_procesado | sentiment |
|---|---|---|---|
| 0 | Alcaldesa de Bogotá | según alcaldesa bogotá delinquir acción inhere... | neutral |
| 2 | Claudia Lopez | claudia lópez propone deportar ' contemplación... | neutral |
| 3 | Claudia Lopez | consecuencia dejar manos claudia lópez venezol... | neutral |
| 4 | Claudia Lopez | según claudia lópez remanso paz hermanos venez... | neutral |
| 5 | Claudia Lopez | aplausos claudia lopez crea xenofobia dura par... | neutral |
| ... | ... | ... | ... |
| 955 | GobiernoBTA | vez usar dineros recompensas deben usarlo paga... | neutral |
| 957 | GobiernoBTA | vengan tan diligentes salle fritos transporte ... | neutral |
| 959 | GobiernoBTA | gracias denuncia haremos revisión respectiva c... | neutral |
| 960 | GobiernoBTA | gracias vallaron parque público carrera calle | neutral |
| 961 | GobiernoBTA | hoy terminamos visitas alcaldías locales parte... | neutral |

698 rows × 3 columns

Fuente: Elaboración propia.

**Figura 52. Dataframe de entidades mencionadas en tweets negativos.**

```
[16] negativo
```

| | Entidad | texto_procesado | sentiment |
|---|---|---|---|
| 1 | Claudia Lopez | “ deberíamos deportarlos contemplación ” claud... | negativo |
| 7 | Duque | don youtuber retos divertidos duque visto mal ... | negativo |
| 9 | Claudia Lopez | colombiano venezolano asumir consecuencias ser... | negativo |
| 10 | Claudia Lopez | xenofobia nadie menos venezolanos además deuda... | negativo |
| 12 | Claudia Lopez | ninguna excepción venezolanos buen obrar mal o... | negativo |
| ... | ... | ... | ... |
| 947 | GobiernoBTA | corona halloween ciudad control | negativo |
| 948 | GobiernoBTA | solución terminar extorsiones poner agentes po... | negativo |
| 949 | GobiernoBTA | claudia lópez inepta farcsante miserable menti... | negativo |
| 950 | GobiernoBTA | tranco calle totalmente imposible policía regu... | negativo |
| 958 | GobiernoBTA | totalmente ilegal calle pobres ricos colombian... | negativo |

219 rows × 3 columns

Fuente: Elaboración propia.

**Figura 53. Dataframe de entidades mencionadas en tweets positivos.**

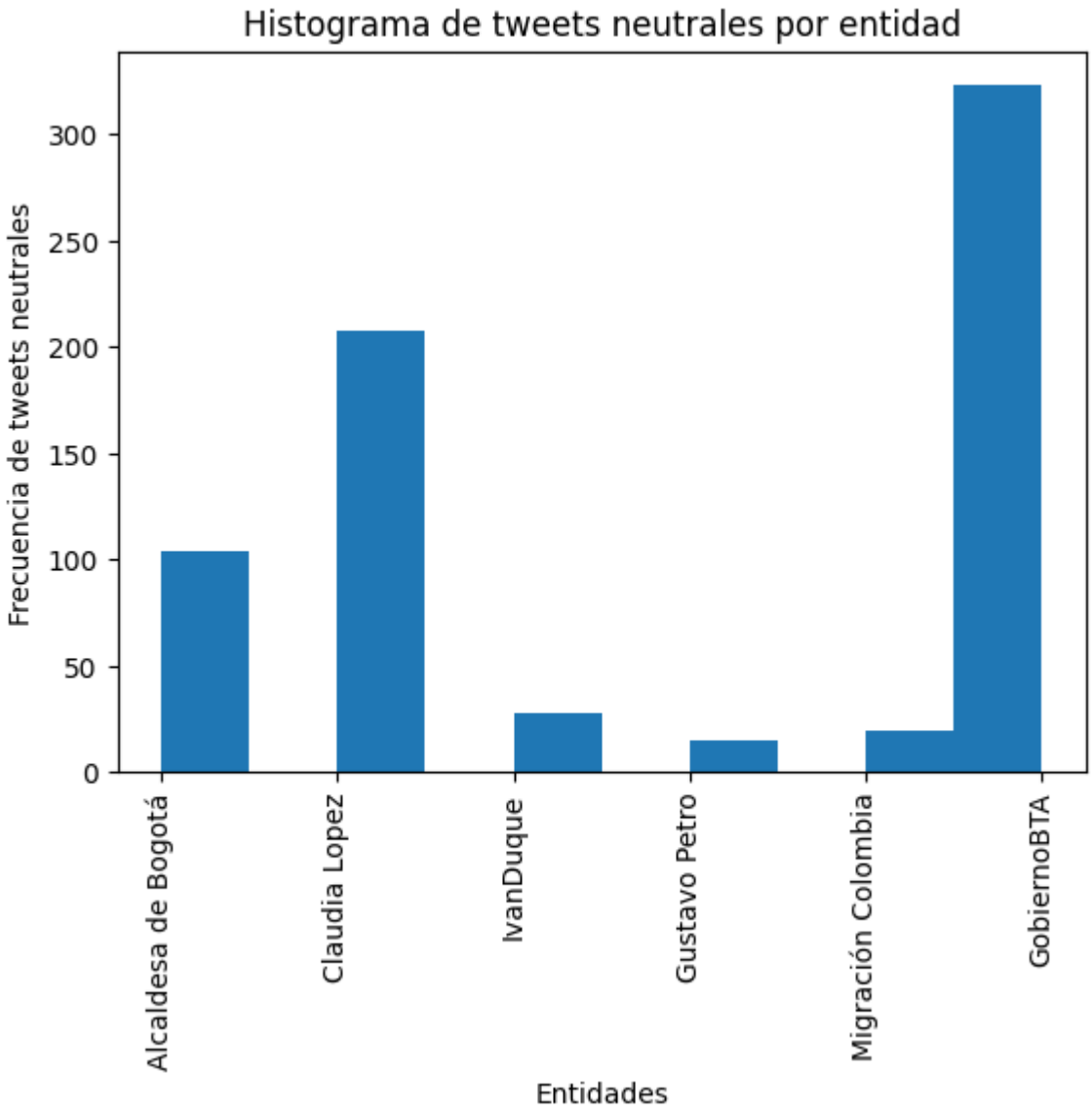
positivo

| | Entidad | texto_procesado | sentiment |
|---|---|---|---|
| 8 | Duque | favor pueden decir fascista corrija éste discu... | positivo |
| 15 | Alcaldesa de Bogotá | señora alcaldesa dios infórmese discriminar ve... | positivo |
| 40 | Alcaldesa de Bogotá | duras crítica recibido alcaldesa bogotá claudi... | positivo |
| 75 | Alcaldesa de Bogotá | sepaque hondo rechazo generado recientes decla... | positivo |
| 80 | Alcaldesa de Bogotá | aunque diferencias sra alcaldesa acuerdo postu... | positivo |
| 103 | Alcaldesa de Bogotá | claudia lópez materialización contradicciones ... | positivo |
| 116 | Alcaldesa de Bogotá | cuestión claudia venido hace rato discurso est... | positivo |
| 162 | Alcaldesa de Bogotá | rechazamos declaraciones señora alcaldesa insi... | positivo |
| 207 | Claudia Lopez | claudia lópez peló cobre comentario xenófobo v... | positivo |
| 209 | Claudia Lopez | claudia lópez peló cobre comentario xenófobo v... | positivo |
| 218 | Claudia Lopez | señora alcaldesa dios infórmese discriminar ve... | positivo |
| 226 | Claudia Lopez | claudia lópez peló cobre comentario xenófobo v... | positivo |

Fuente: Elaboración propia.

Esto se realizó con el fin de formar histogramas, para visualizar la frecuencia de las entidades mencionadas en los tweets, por cada una de las categorías relacionadas a la polaridad de sentimiento.

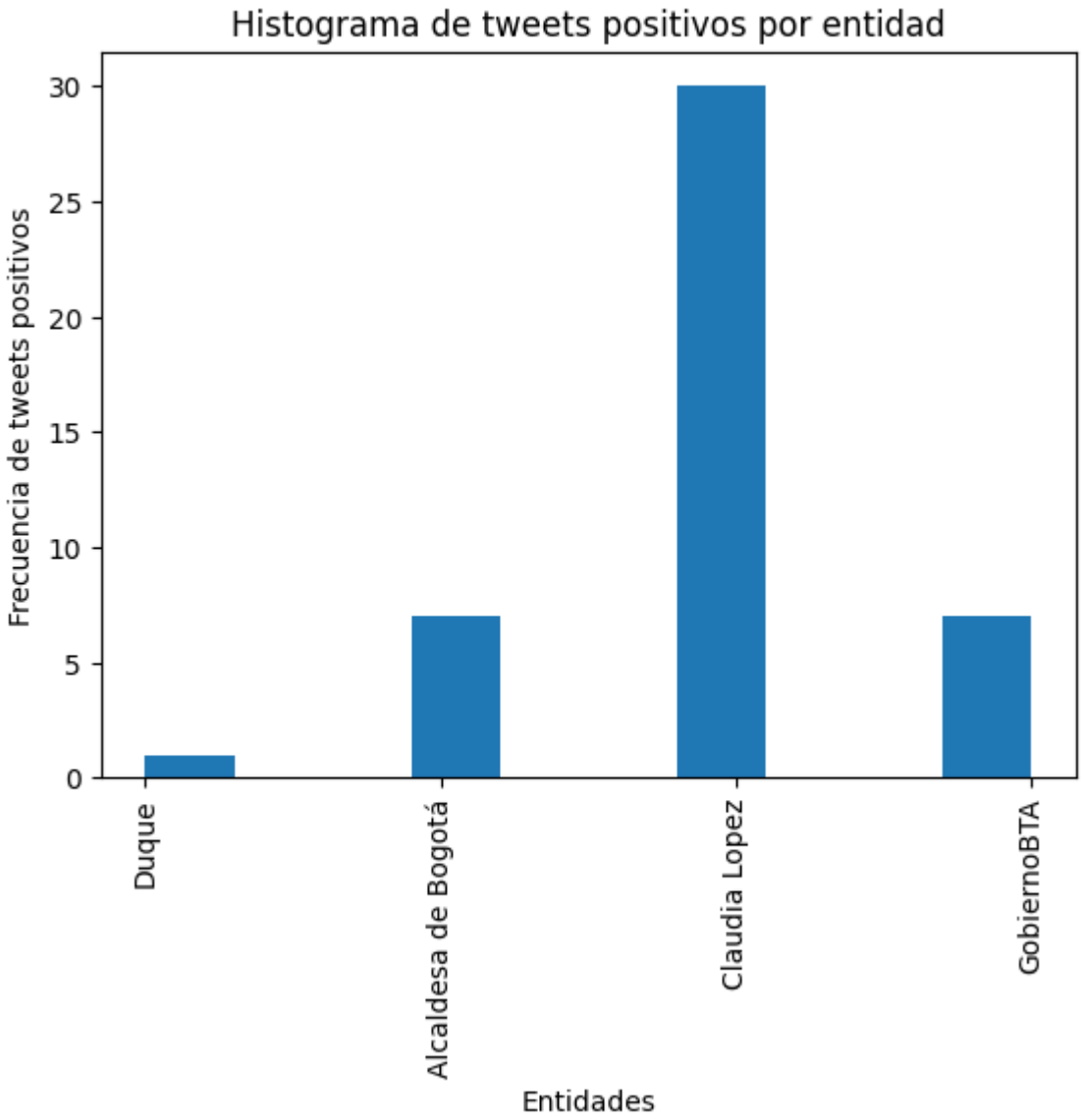
**Figura 54. Histograma cantidad de menciones de cada entidad en tweets neutrales.**




Fuente: Elaboración propia.

En este primer histograma realizado para medir la frecuencia de los tweets neutrales por entidad, se observa que la mayoría de los tweets neutrales, se relacionan con la mención de la cuenta GobiernoBTA, seguido de Claudia López. Se puede observar que, de los 698 tweets con clasificación neutral, más de 300 corresponden a comentarios en donde se menciona a GobiernoBTA.

**Figura 55. Histograma cantidad de menciones de cada entidad en tweets positivos.**

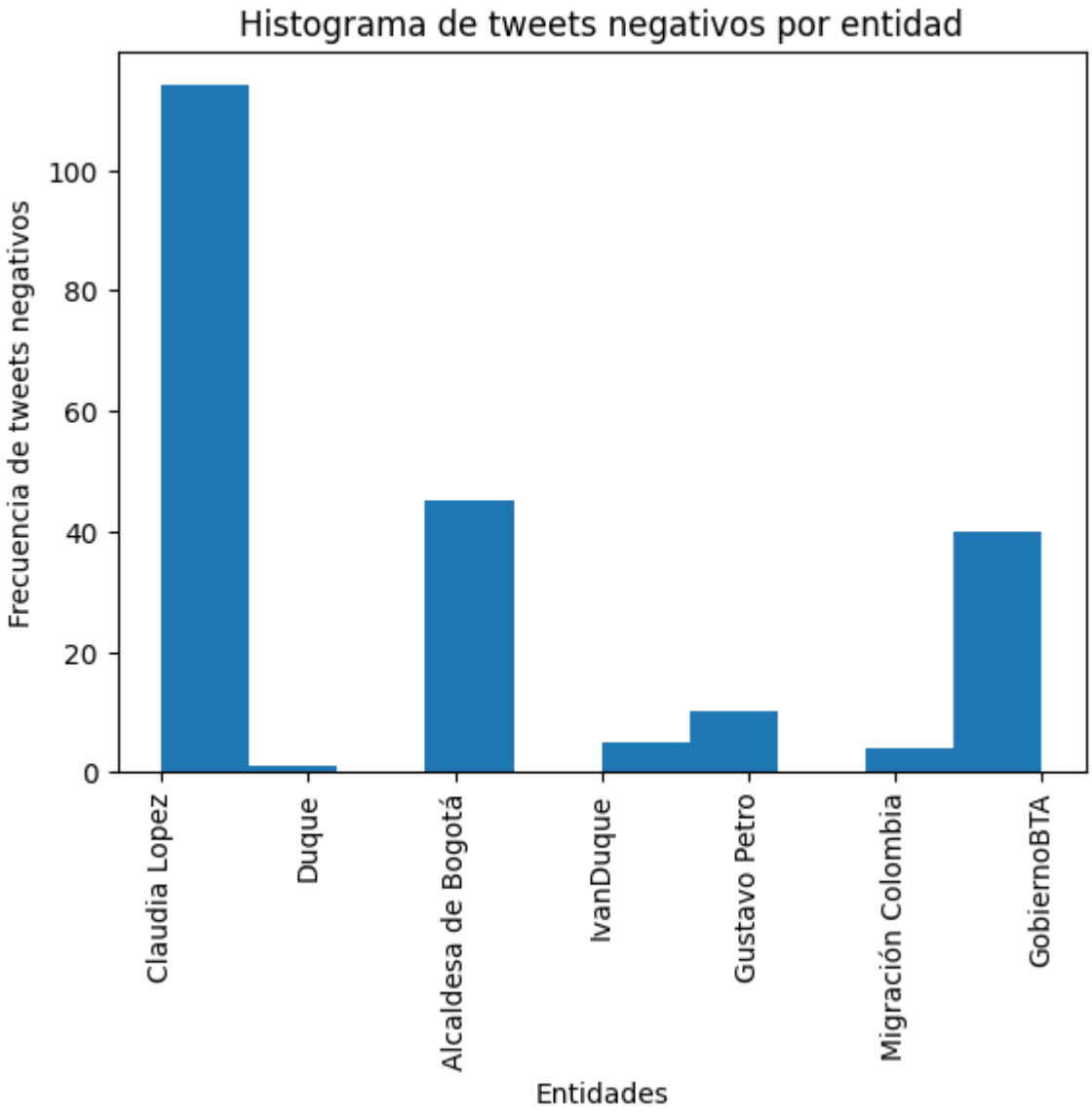



Fuente: Elaboración propia.

En este segundo histograma se observa que la mayoría de los tweets **positivos**, son relacionados con Claudia López. Sin embargo, son pocos los tweets positivos, con un total de 45 dentro de los 962 agrupados de las entidades mencionadas. De

estos 45 positivos, 30 corresponden a publicaciones en donde se mencionan a Claudia López.

**Figura 56. Histograma cantidad de menciones de cada entidad en tweets negativos.**

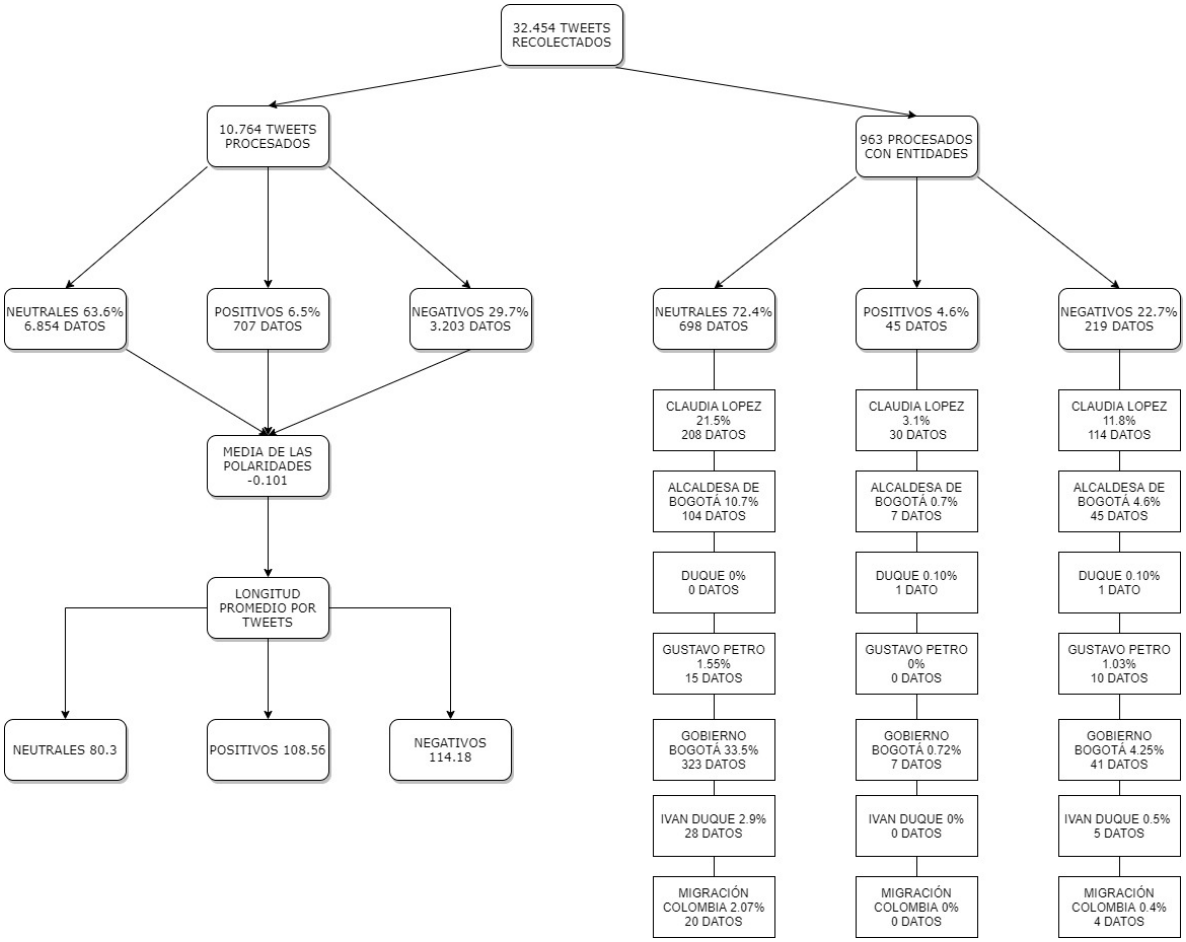



Fuente: Elaboración propia.

Este último histograma corresponde a la frecuencia de tweets negativos en donde se mencionan a las entidades. En esta se encuentra nuevamente *Claudia López* seguido de *alcaldesa de Bogotá*, con más frecuencia en los tweets negativos.

Con esta observación se puede inferir que las entidades con más menciones en los tweets negativos corresponden a la cuenta de la alcaldesa de Bogotá *claudia López*, seguido del término que se le denomina que es *alcaldesa de Bogotá*. esto refleja gran rechazo de los usuarios de las redes sociales hacía la alcaldesa con respecto al actual fenómeno migratorio de venezolanos en Colombia.

**Figura 57. Modelo sintetizado correspondiente a los dos análisis de sentimientos.**




Fuente: Elaboración propia.

## 5.3 Evaluación del modelo

### 5.3.1 Validación de la academia

Para realizar la evaluación del modelo, se validó la variable correspondiente a la validación de la academia que corresponde a la precisión del modelo. Para esto se definió inicialmente un entrenamiento del modelo con los datos resultantes obtenidos del análisis de sentimiento utilizando la herramienta VADER, en donde se obtuvo un conjunto de datos con un total de 10.764 tweets procesados y su

respectivos etiquetados en donde se clasificaron los tweets con polaridad positiva, negativa o neutral. Para esto fue necesario dividir el conjunto de datos en dos partes. En la primera parte se utilizaron 8.073 tweets (equivalente a un 75%) para el entrenamiento y el restante 25% igual a 2.961 datos, fue utilizado para realizar las pruebas de la evaluación.

**Figura 58. División del dataset para datos de entrenamientos y de pruebas.**

```
[86] from sklearn.model_selection import train_test_split

[87] x_train, x_test, y_train, y_test = train_test_split(all_features, data.sentiment,
                                                         test_size=0.25)
```

Fuente: Elaboración propia.

De estos 2691 datos, 1864 tweets fueron clasificados correctamente en relación a sus respectivas polaridades y un total de 827 clasificaciones fueron clasificadas incorrectamente.

**Figura 59. Total de clasificaciones correctas e incorrectas para tweets con entidades.**

```
[94] print(f'{correct} clasificaciones correctas')

     1864 clasificaciones correctas

[95] incorrect = y_test.size - correct

[96] print(f'{incorrect} clasificaciones incorrectas')

     827 clasificaciones incorrectas
```

Fuente: Elaboración propia.

De esta manera, con ayuda de la librería sklearn.metrics, se evaluaron métricas tales como la precisión, que es definida como el número de verdaderos positivos divididos entre el número de verdaderos positivos más falsos positivos y el accuracy definido como qué tan cercanos fueron las medidas al valor verdadero.

***VP:*** *Verdaderos positivos*
***VN:*** *Verdaderos negativos*
***FP:*** *Falsos positivos*
***VN:*** *Verdaderos negativos*

***Precision*** = VP / VP + FP

***Accuracy*** = VP + VN / VP + VN + FP + FN

También se evalúa el *recall* que se refiere al porcentaje total de resultados correctamente clasificados por el algoritmo.

***Recall*** = VP / VP + FN

A partir de los resultados obtenidos en la evaluación, se puede observar que para la polaridad positiva se obtuvo una precisión media (0.53%) y un recall parcialmente alto (0.72%) por lo que se puede inferir que el modelo clasifica correctamente los tweets positivos sin embargo puede fallar en algunos etiquetados.

Por otro lado, con respecto a la polaridad negativa, se obtuvieron una precisión y un recall relativamente altos (0.81% y 0.74%). Con esto se puede deducir que el modelo clasifica de manera correcta la mayoría de los tweets con sentimiento negativo.

Por último, para la etiqueta neutral se obtuvo una baja precisión con un 0.47% y un bajo recall (0.10%) lo que indica que los tweets neutrales son clasificados de manera incorrecta.

**Figura 60. Resultados evaluación del modelo.**

```
print(classification_report(y_test, classifier.predict(x_test), target_names=['positivo','negativo','neutral']))

              precision    recall  f1-score   support

    positivo       0.53      0.72      0.61       819
    negativo       0.81      0.74      0.78      1688
     neutral       0.47      0.10      0.17       184

    accuracy                           0.69      2691
   macro avg       0.61      0.52      0.52      2691
weighted avg       0.71      0.69      0.69      2691
```

Fuente: Elaboración propia.

### 5.3.2 Validación estática

En esta sección de evaluación del modelo, se validó la variable correspondiente a la validación estática en donde se evalúa la utilidad del modelo a partir del nivel de satisfacción expresado en la opinión por el co-líder del grupo de investigaciones OSIRIS & BIOAXIS de la Universidad El Bosque.

Para realizar esta evaluación se realizó una entrevista semiestructurada a el representante del grupo de investigaciones OSIRIS & BIOAXIS de la Universidad El Bosque, Carlos Delgado.

Para esta entrevista se formularon las siguientes preguntas:

1. ¿A partir del contexto planteado se puede identificar una problemática social
2. cultural que se adapte a la creación de un modelo BPSC?
3. ¿Cree usted que la variable seleccionada fue la apropiada para estudiar el fenómeno político definido?
4. ¿En el artefacto implementado se identifican los actores principales relacionados al fenómeno político estudiado?
5. ¿En el artefacto implementado se relacionan las interacciones de la población con las principales entidades o actores políticos identificados?
6. ¿Qué hábitos y creencias se logran identificar en el modelo descriptivo?
7. ¿Es posible adaptar el modelo implementado en estudios futuros seleccionando un fenómeno político social diferente?
8. ¿El modelo satisface sus necesidades para el estudio de fenómenos políticos a partir de la información recopilada mediante las redes sociales?
9. ¿La técnica de modelado escogida fue pertinente para medir la variable establecida?

Luego de la entrevista realizada al ingeniero representante del grupo de investigación OSIRIS & BIOAXIS de la Universidad El Bosque, Carlos Delgado, se obtuvieron las respuestas a las preguntas definidas previamente.

Dando respuesta a la primera pregunta **(1) ¿A partir del contexto planteado se puede identificar una problemática social cultural que se adapte a la creación de un modelo BPSC?**, el investigador respondió lo siguiente:

A partir del contexto planteado y de los resultados mostrados, sí es posible identificar una problemática social. Gracias a los resultados en donde se pudieron observar términos despectivos, se puede encontrar una problemática de juicio y rechazo a los migrantes provenientes de un país distinto. El ingeniero coloca como ejemplo su situación personal de cuando vivió en la ciudad de Caracas, Venezuela durante dos años, en donde pudo percibir en esa época la discriminación hacia ciudadanos colombianos que emigraron a Venezuela entre los años 70 y 80, en donde se escuchaban palabras despectivas hacías estas personas, sin embargo, resalta que muchas veces, esas palabras despectivas son utilizadas cuando los migrantes cometen actos incorrectos y delincuenciales en el país destino.
Sin embargo, resalta que esta opinión es un poco ignorante con respecto al conocimiento de los hechos ocurridos, y asegura que gracias al modelo BPSC podría reflejar de manera objetiva la situación y encontrar variables que permitan

estudiar el problema para llegar a buscar posibles soluciones que se pueda hacer al respecto.

Con respecto a la segunda pregunta definida, **(2) ¿Cree usted que la variable seleccionada fue la apropiada para estudiar el fenómeno político definido?** la respuesta del investigador fue que:
En una primera instancia si, sin embargo, dado los resultados del modelo implementado puede ampliarse y sondear de manera que se definan nuevas variables de estudio como, por ejemplo: Las expresiones en relación con venezolanos que estudian o trabajan dentro de Colombia de manera honesta, esto de manera que se genere un estudio más cualitativo con mayor cantidad de resultados. Como recomendación futura el ingeniero Carlos plantea la posibilidad de determinar que tanto daño causa las opiniones mostradas en los tweets con relación al fenómeno escogido.

Para la tercera pregunta **(3) ¿En el artefacto implementado se identifican los actores principales relacionados al fenómeno político estudiado?** la respuesta fue que:

Si, se logra identificar ciertos actores en los resultados como lo son la población venezolana en Colombia, la alcaldesa, el presidente, el ex candidato a la presidencia Gustavo Petro entre otros, sin embargo el ingeniero Carlos plantea recomendación a futuro clasificar en grupos aquellas personas que se expresan en relación al fenómeno, aquellas que apoyan o juzgan dichas expresiones y las personas que generan ruido, adicionalmente el ingeniero Carlos nos plantea realizar un estudio de atributos como el tipo de persona (natural o jurídica) o rango de edades de aquellas personas que publican los tweets. Estas recomendaciones con el fin de generar una mayor cantidad de grupos de actores que cumplan una mejor caracterización.

Siguiendo con la cuarta pregunta **(4) ¿En el artefacto implementado se relacionan las interacciones de la población con las principales entidades o actores políticos identificados?** la respuesta fue la siguiente:

Esta respuesta está relacionada con la respuesta de la anterior pregunta, ya que se debe un poco más sobre los actores, con más detalle para obtener mejores relaciones. Sin embargo, se resalta que no es posible relacionar las interacciones de los usuarios con las principales entidades identificadas, ya que el modelo realiza un análisis de sentimientos de los tweets, pero no se realiza un análisis de las relaciones entre las entidades. Al no realizar un estudio de los actores, no se puede obtener un análisis de estas relaciones en cuanto a las entidades, sin embargo, resalta que este no es el objetivo del proyecto, pero puede ser tomado en cuenta para futuros trabajos.

Con respecto a la quinta pregunta **(5) ¿Qué hábitos y creencias se logran identificar en el modelo descriptivo?,** se obtuvo la siguiente respuesta:

A partir de los resultados se logran identificar unos posibles hábitos como puede ser: la expresión de opiniones generalizadas en la red social Twitter sin tener una información verídica respecto a lo que se opina, esto como resultado de acciones de una minoría de personas en relación con malos actos. Para las creencias se puede pensar que las expresiones negativas se deben a la creencia que se tiene que los migrantes les quitaran el trabajo a los nativos, esto pudiendo ser un posible factor que genera rechazo hacia la población venezolana.

En la sexta pregunta **(6) ¿Es posible adaptar el modelo implementado en estudios futuros seleccionando un fenómeno político social diferente?,** se obtuvo como respuesta que:

Sí es posible, ya que hay fenómenos que sí se pueden trabajar bajo la variable de aceptación y/o rechazo, y expone como ejemplo la situación de polarización política en Colombia.

En la pregunta siete **(7) ¿El modelo satisface sus necesidades para el estudio de fenómenos políticos a partir de la información recopilada mediante las redes sociales?,** el ingeniero Carlos respondió que:

Sí fueron útiles los datos recopilados para estudiar este fenómeno, adicionalmente el grupo de investigación tiene interés en el método de recolección, para poder ser replicado en clases de primeros semestres de la carrera de ingeniería de sistemas. Además, el modelo puede ser utilizado para encontrar una mejor toma de decisiones a partir del árbol de resultados obtenidos.

Para la octava y última pregunta **(8) ¿La técnica de modelado escogida fue pertinente para medir la variable establecida?,** la respuesta fue la siguiente:

Según los resultados obtenidos y el uso de esta técnica en estudios previos se considera que fue pertinente utilizar un análisis de sentimientos como técnica de modelado para medir la variable de interés, por su parte el ingeniero delgado también nos indica que, aunque el modelo no puede ser matemáticamente preciso ya que se basa en un estudio de léxicos es recomendado hacer un ajuste de precisión al modelo en relación con el etiquetado neutral para una mejor clasificación.

# 6. ANÁLISIS DE RESULTADOS/ DISCUSIÓN

En primer lugar, la elección de la metodología seleccionada fue la más apropiada para el desarrollo de este proyecto, debido a su correcta estructuración dividida por fases, que facilitó una mejor comprensión de la estructura del proyecto con respecto a las necesidades específicas del cliente. Esto gracias a la adecuada manera de corregir los errores al permitir regresar entre las fases previas realizar ajustes necesarios.

Durante el desarrollo del proyecto, se presentaron varias dificultades que afectaron los tiempos de ejecución y entregas al cliente. Inicialmente se realizaron ajustes en el estado del arte debido a cambios del enfoque del proyecto, ya que, al mismo tiempo, surgieron inconvenientes al momento de seleccionar el fenómeno político de estudio. Esto fue conversado con el cliente para definir puntualmente el fenómeno, ya que en primera instancia se había establecido estudiar la aceptación y rechazo de la población colombiana con respecto a la gestión de la alcaldesa de Bogotá Claudia López. Posteriormente esto fue modificado y aprobado para estudiarla aceptación y/o rechazo de la población colombiana a partir de las expresiones obtenidas mediante la red social Twitter relacionadas con el fenómeno de la inmigración venezolana.

Otro inconveniente que surgió durante el desarrollo del proyecto fue con respecto a la elección de la herramienta a utilizar para la extracción y recolección de los datos. Esto dado que inicialmente se propuso utilizar *Tweepy* como librería para la recolección de los datos, sin embargo, luego de revisar los límites de datos extraídos, se decidió realizar un web *scraping* a páginas web de noticias y extraer las noticias más significativas con respecto al fenómeno escogido. Dado que esta última opción no se adaptó al objetivo principal del proyecto, se decidió utilizar la herramienta tag *explorer* para recopilar tweets de Twitter. Esta herramienta no permitió extraer tweets entre un rango de fechas en específico, por lo cual, luego de varias pruebas, se descartó. Finalmente, luego de indagaciones en otros proyectos similares revisados, se determinó utilizar la API de Twitter para realizar la extracción de los datos correctamente cumpliendo con todos los requisitos solicitados.

Para el cumplimiento de los objetivos específicos uno y dos fue necesaria la aplicación de ciertas herramientas y documentos que ayudaron agilizar el proceso en la construcción del corpus bibliográfico, redacción del estado del arte y marco teórico. La matriz de estado de arte permitió tener un orden sobre las bibliografías que sirvieron de apoyo para la composición del documento y el desarrollo del artefacto; la evaluación de calidad permitió escoger aquellos trabajos que cumplieron con una buena puntuación lo que a su vez significa que poseían información útil y de calidad, para dar cumplimiento con la evaluación de calidad fue necesaria realizar una investigación en *Web Of Science* la cual nos permitió conocer aquellos documentos más relevantes en las áreas de interés y los autores relevantes.

Por último, se generó un documento donde se clasificaron componentes encontrados en los trabajos consultados esto permitió tener idea de que fenómeno podríamos escoger para modelar, qué herramientas utilizar para el modelo o donde encontrar los datos para el estudio.

El objetivo tres fue el de más extenso trabajo y donde más cambios se presentaron, para dar finalizado con este objetivo fue necesario una ardua investigación inicial para determinar qué herramientas serían las ideales para la extracción de los datos, procesado y modelado teniendo en cuenta esto el siguiente paso fue extraer los datos que tuvieran mayor cantidad de información de calidad y verídica, luego de utilizar numerosas herramientas se definió la API para la extracción desde *postman* ya que era la que permitió mayor extracción de tweets en el menor tiempo posible.

El modelo implementado logró cumplir con el propósito planteado en este proyecto, en donde se deseaba analizar los comentarios de las personas a través de Twitter, ya que mostró un adecuado funcionamiento con respecto a la clasificación que se buscaba con el conjunto de datos obtenidos. Con esta información se pudieron analizar cada dato y sacar una polaridad aproximada al sentimiento presentado en cada tweet. Luego de analizar los datos sobre el fenómeno escogido, se obtuvieron como gran parte de los resultados, valores neutrales con respecto al fenómeno ocurrido durante una fecha particular. La siguiente polaridad con mayor presencia en los datos, fue el del sentimiento con valor negativo, relacionados al rechazo por parte de la población colombiana. Y por último se encontró que la polaridad positiva, fue la de menor impacto y presencia en los comentarios publicados por los usuarios del país Colombia en cuanto a la aceptación de los venezolanos frente al fenómeno migratorio que se vive actualmente.

Luego de realizar la evaluación de modelo, a partir de la validación de la academia, en donde se buscaba evaluar principalmente la precisión del modelo, se obtuvo que el modelo acertaba correctamente con la clasificación de los tweets con sentimiento negativos, dado que esta variable arrojó un porcentaje de precisión y de *recall* altos. Sin embargo, con respecto a la polaridad neutra, se encontró que el modelo no clasifica correctamente los tweets que fueron etiquetados con esta polaridad, ya que es muy probable que gran parte de esta clasificación, puedan hacer parte de las otras etiquetas de sentimientos positivos o negativos, esto debido al resultado de precisión y *recall* bajo en la evaluación.
Para dar por cumplido el cuarto y último objetivo correspondiente a la validación estática que busca evaluar si el modelo logra comprender el fenómeno estudiado desde la perspectiva biopsicosocial y cultural. Para esto se realizó una entrevista semiestructurada al representante Carlos Delgado del grupo de investigación de la universidad, en donde se plantearon preguntas relacionadas a los componentes del modelo BPSC, a la variable estudiada, a la técnica de modelado seleccionada y la utilidad del modelo para trabajos futuros.

A partir de esta entrevista, se obtuvieron resultados positivos ya que el beneficiario mostró interés y satisfacción por el modelo desarrollado. Esto se puede deducir gracias a las respuestas obtenidas en la entrevista realizada, y a la retroalimentación otorgada por el beneficiario. De las preguntas planteadas, la gran mayoría fueron respondidas positivamente, sin embargo, se obtuvieron observaciones y recomendaciones para realizar en trabajos futuros.

En relación con la utilidad del modelo el profesor mencionó que muestra ser preciso con respecto al etiquetado positivo y negativo, no obstante resaltó que sería indicado buscar la mejorar en relación al etiquetado neutral y de esta manera encontrar un incremento entre un 5 - 10 % en la exactitud general del modelo que mostro en un principio ser del 70% que para el fenómeno tratado al no ser de carácter ampliamente variable puede indicar ser un buen resultado con oportunidad de mejoras.

Como conclusión, el ingeniero Carlos resalta que una utilidad del modelo es la generación de nuevos métodos y herramientas de trabajos que podrían permitirle al grupo de investigación, estudiar situaciones propias de la comunidad universitaria. Una de estas situaciones es el plan estratégico de la universidad que se planea para los próximos cinco años, permitiendo tener en cuenta la búsqueda de nuevos fenómenos que pueden ser llamativos para trabajar en la universidad, con el fin de generar soluciones a un fenómeno desde las diferentes facultades de la universidad, a partir de los diferentes puntos de vista encontrados en las redes sociales.

# 7. CONCLUSIONES

El presente trabajo de grado tuvo como objetivo el desarrollo de un modelo descriptivo sobre un fenómeno político aplicando la minería de datos a una red social, que permita comprender los actores, flujos de información e interacciones que rigen el comportamiento del fenómeno desde la perspectiva biopsicosocial y cultural. Para esto fue necesario el cumplimiento de cuatro objetivos específicos para lograr con el resultado esperado del objetivo general.

El primer objetivo específico se cumplió satisfactoriamente luego de realizar una profunda investigación a trabajos antecedentes que dieron fundamentos bases para el desarrollo de este proyecto. Se obtuvo un corpus bibliográfico robusto y preciso de documentos importantes que tocaron temas similares tanto nacional como internacionalmente. Estos trabajos realizados, tuvieron un impacto importante para la elaboración del presente trabajo.

En cuanto al segundo objetivo, se pudo definir un fenómeno político adecuado capaz de adaptarse correctamente al problema de investigación del proyecto. Esto fue posible gracias al proceso de investigación que permitió realizar el objetivo anterior exitosamente. Además, a través del análisis de tendencias y noticias fue posible dar cumplimiento con el objetivo dos, cuyo fin era determinar un fenómeno político social posible de modelar aplicando minería de datos y técnicas de análisis de texto. A partir del problema definido, se pudo determinar la siguiente pregunta de investigación:

¿Cómo se expresa la aceptación o rechazo por parte de la población colombiana, respecto a la gestión política y social en relación con el fenómeno de inmigración venezolana que ocurre en el país?

Esta pregunta permitió iniciar las fases de la metodología, permitiendo obtener datos específicos y procesarlos para cumplir con el desarrollo del modelo descriptivo planteado en el objetivo general del trabajo.

El análisis de sentimientos fue la técnica de modelado escogida para dar cumplimiento con el tercer objetivo el cual consistió en el desarrollo de un proceso de minería de datos, guiado bajo la metodología CRISP-DM, en donde inicialmente se recopilaron los datos exitosamente gracias al aprendizaje y uso de la API de Twitter. Con esto se pudo desarrollar un modelo que permitió estudiar y evaluar el comportamiento del fenómeno político definido.
Para lograr el cumplimiento de este proceso de minería de datos, se clasificaron sentimientos expresados por parte de los usuarios colombianos en Twitter referente a la migración de los venezolanos. Esta fue la técnica escogida ya que se consideró la más pertinente para dar respuestas variables a medir (Aceptación o rechazo). El

modelo Vader fue escogido por su buen funcionamiento en cuanto a redes sociales se trata dio como resultado polaridades esperadas desde un principio en el desarrollo del proyecto.

Para el cumplimiento del último objetivo, se logró programar una entrevista con el beneficiario del desarrollo tecnológico construido en este proyecto, con el fin de obtener su punto de vista y su nivel de satisfacción positivo con respecto al artefacto desarrollado.

Gracias al desarrollo de los objetivos específicos, se pudo concluir que se logró cumplir con el objetivo general del proyecto. Pese a las limitaciones encontradas se pudo cumplir con el alcance planteado inicialmente en el proyecto.

# 8. RECOMENDACIONES Y TRABAJOS FUTUROS

## 8.1 Recomendaciones

Es de suma importancia mejorar la precisión del modelo realizado en relación al etiquetado neutral, para ello se recomienda agregar un conjunto de palabras relacionadas al fenómeno que se vaya a estudiar, esto con el fin de incrementar el conjunto de léxicos con el que trabaja el modelo, o en su defecto, considerar otras técnicas de análisis de texto y hacer comparativa en los resultados obtenidos.

En cuanto a la minería de datos, en esta puede ser frecuente encontrar resultados no esperados, no necesariamente por mala ejecución de la metodología o mala implementación de los algoritmos, a veces es consecuencia de una mala limpieza de los datos o unos datos iniciales carentes de calidad. Por ello como recomendación es de vital importancia extraer o contar con un dataset limpio y de calidad, o en su defecto realizar una limpieza exhaustiva a los datos con el fin de realizar un mejor análisis e interpretación de la información.

## 8.2 Trabajos futuros

Como sugerencia para trabajo futuro y con el fin de obtener un análisis de mayor alcance, se recomienda estudiar el efecto que tiene la aceptación o rechazo expresado en los tweets sobre la población a la cual son dirigidas estas expresiones. Adicionalmente, se puede tener en cuenta la posibilidad de estudiar otras nuevas variables, que se puedan generar a partir de los resultados obtenidos en este proyecto.

El modelo resultante de este proyecto puede ser adaptado a nuevas propuestas de estudios para distintos fenómenos políticos o sociales, que puedan ser similares al estudiado en este trabajo.

# 9. LECCIONES APRENDIDAS

Durante el desarrollo y gestión del proyecto se adquirieron los nuevos conocimientos a partir de las siguientes lecciones aprendidas:

- Seguir un cronograma de reuniones periódicas con el director de la tesis y el cliente permitió entender las necesidades que el cliente tenía referente al modelo; permitió a su vez aclarar dudas y corregir errores en el tiempo necesario para que no ocurrieran retrasos significativos.
- El proceso de búsqueda del fenómeno elegido a través una serie de tareas (Búsqueda de bibliografía, Matriz de estado del arte, categorización de referencias, análisis de relevancia y evaluación de calidad) permitió seleccionar el tema ideal que sería objeto de investigación para el desarrollo de nuestro modelo de análisis de sentimiento.
- Fue necesario tener reuniones con profesores como: Jorge Duarte para definir la mejor metodología a seguir para el proyecto realizado, Jaime Montaña para definir cómo debían ser estructurado el modelo biopsicosocial y cultural que surgiría del modelo, Frank Romero para establecer cuáles eran los diferentes métodos que teníamos para realizar análisis de texto y de opiniones.

En cuanto a la implementación del artefacto, se adquirieron las siguientes lecciones referentes al proceso de ingeniería:

- El trabajo realizado siguiendo una metodología dirigida a proyectos de minería de datos *CRISP-DM* permitió tener un conocimiento concreto de cuáles eran las tareas para realizar y el orden en que estas debían ser ejecutadas para tener éxito dentro del proyecto, adicionalmente permitió devolverse entre sus fases para corregir errores presentados sin la necesidad de volver al principio y retrasar los tiempos estimados en la ejecución.
- Obtener rápido el acceso a las llaves para el uso de la API de Twitter que, mediante *postman* permitió la fácil extracción que conformaría el corpus de datos que serían utilizados en el modelo. Esta librería permite realizar diferentes acciones entre las que más destacan: Búsqueda de contenido, contenido en tiempo real, mensajes directos, embeber tweets entre otras.
- Python es un lenguaje robusto en lo que a datos se trata por ello fue de gran utilidad sin necesidad de llegar a usar softwares externos para realizar el análisis de sentimientos. Proporciona una gran cantidad de librerías y

modelos como son: *Pandas, NumPy, NLTK, TensorFlow, Re, Sentiment-analysis-spanish* perfectos para el análisis y manejo de textos y números.

- Si se desea acotar tiempo en la construcción del corpus de datos es necesario desde un principio definir el método de extracción más adecuado que permita establecer las mejores técnicas de limpieza según el formato en el que se obtienen, entre las técnicas que se utilizaron resaltan: *Web Scrapping en páginas web, API news, TAGexplorer, Twitter API v2*.

## REFERENCIAS BIBLIOGRÁFICAS

[1] M. Fachrie, “Machine Learning for Data Classification in Indonesia Regional Elections Based on Political Parties Support,” J. Ilmu Komput. dan Inf., vol. 13, no. 2, pp. 89–96, 2020, doi: 10.21609/jiki.v13i2.860.

[2] M. Hoferer, L. Böttcher, H. J. Herrmann, and H. Gersbach, “The impact of technologies in political campaigns,” Phys. A Stat. Mech. its Appl., vol. 538, no. September, 2020, doi: 10.1016/j.physa.2019.122795.

[3] A. Khan et al., “Predicting Politician ’ s Supporters ’ Network on Twitter Using Social Network Predicting Politician ’ s Supporters ’ Network on Twitter Using Social Network Analysis and Semantic Analysis,” no. September, 2020, doi: 10.1155/2020/9353120.

[4] A. A. Arifiyanti, E. D. Wahyuni, and A. Kurniawan, “Emotion Mining of Indonesia Presidential Political Campaign 2019 using Twitter Data,” J. Phys. Conf. Ser., vol. 1569, p. 022035, 2020, doi: 10.1088/1742-6596/1569/2/022035.

[5] S. Bradshaw and P. N. Howard, “The Global Disinformation Order 2019 Global Inventory of Organised Social Media Manipulation,” Univ. Oxford, p. 25, 2019, [Online]. Available: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf.

[6] E. L. Boldyreva, “Cambridge Analytica: Ethics And Online Manipulation With Decision-Making Process,” no. December 2018, pp. 91–102, 2018, doi: 10.15405/epsbs.2018.12.02.10.

[7] I. Majeed and S. Naaz, “Current State of Art of Academic Data Mining and Future Vision,” Indian J. Comput. Sci. Eng., vol. 9, no. 2, pp. 49–56, 2018, doi: 10.21817/indjcse/2018/v9i2/180902026.

[8] S. Mathew, J. T. Abraham, and S. J. Kalayathankal, “Data mining techniques and methodologies,” Int. J. Civ. Eng. Technol., vol. 9, no. 7, pp. 246–252, 2018.

[9] O. Papakyriakopoulos, S. Hegelich, M. Shahrezaye, and J. C. M. Serrano, “Social media and microtargeting: Political data processing and the consequences for Germany,” Big Data Soc., vol. 5, no. 2, pp. 1–15, 2018, doi: 10.1177/2053951718811844.

[10] O. D. Apuke, “The role of social media and computational propaganda in political communication campaign,” J. Lang. Commun., vol. 5, no. 2, 2018, [Online]. Available: http://www.journalfbmk.upm.edu.my/index.php/jlc/article/view/28.

[11] J. Ndavula, “Social Media adoption as alternative platforms for political marketing in,” no. June, 2017.

[12] P. Pawełczyk and J. Jakubowski, “Political marketing in the times of big data,” Przegląd Politol., vol. 0, no. 3, pp. 34–44, 2017, doi: 10.14746/pp.2017.22.3.3.

[13] G. M. Fulgoni, A. Lipsman, and C. Davidsen, “The power of political advertising: Lessons for practitioners: How data analytics, social media, and creative strategies shape U.S. presidential election campaigns,” J. Advert. Res., vol. 56, no. 3, pp. 239–244, 2016, doi: 10.2501/JAR-2016-034.

[14] A. C. Carrillo Chappe, J. C. Urueña García, F. A. Julián David, and L. E. Caicedo Ortiz, “Análisis del sentimiento político mediante la aplicación de herramientas de minería de datos a través del uso de redes sociales,” p. 98, 2015.

[15] J. Uribe, “Twitter, la nueva era de la comunicación política - Un análisis del manejo de la red social en las elecciones a la alcaldía de Bogotá 2015-,” Tesis Pregr., 2015.

[16] V. Galán, “Crisp-Dm a Un Proyecto De Minería,” p. 120, 2015.

[17] D. A. Solano Oviedo, “Aplicación de Minería de Datos para la Identificación de Patrones de comportamiento en las organizaciones enfocado en Prácticas de Impresión: Caso de Estudio,” pp. 1–93, 2014.

[18] D. W. Nickerson and T. Rogers, “Political campaigns and big data,” J. Econ. Perspect., vol. 28, no. 2, pp. 51–74, 2014, doi: 10.1257/jep.28.2.51.

[19] UIAF, “Técnicas de minería de datos para la detección y prevención del lavado de activos y la financiación del terrorismo (LA/FT),” Minhacienda, p. 35, 2014.

[20] C. Monti, A. Rozza, G. Zappella, M. Zignani, A. Arvidsson, and E. Colleoni, “Modelling political disaffection from Twitter data,” Proc. 2nd Int. Work. Issues Sentim. Discov. Opin. Mining, WISDOM 2013 - Held Conjunction with SIGKDD 2013, no. August, 2013, doi: 10.1145/2502069.2502072.

[21] D. W. Nickerson and T. Rogers, “Political Campaigns and Big Data Faculty Research Working Paper Series,” p. 28, 2013, [Online]. Available: http://web.hks.harvard.edu/publications%0A.

[22] L. Castro, “El marketing político en Estados Unidos: el caso Obama,” Norteamérica, vol. 7, no. 1, pp. 209–222, 2012, [Online]. Available:

http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-35502012000100008.

[23] S. Ansolabehere and E. Hersh, “Validation: What big data reveal about survey misreporting and the real electorate,” Polit. Anal., vol. 20, no. 4, pp. 437–459, 2012, doi: 10.1093/pan/mps023.

[24] J. O. Olujide, S. L. Adeyemi, and R. A. Gbadeyan, “Nigerian Electorates’ Perception of Political Advertising and Election Campaign,” J. Soc. Sci., vol. 27, no. 3, pp. 179–185, 2011, doi: 10.1080/09718923.2011.11892918.

[25] G. R. Murray and A. Scime, Microtargeting and electorate segmentation: Data mining the American National Election Studies, vol. 9, no. 3. 2010.

[26] H. M. Gomes, J. Read, and J. P. Barddal, “Machine learning for streaming data : state of the art , challenges , and opportunities.”

[27] R. Eduardo, Z. Díaz, and K. Andrés, “Uso de la red social ‘ Facebook ’ en la política cartagenera , basado en el Modelo de la recepción selectiva . Caso del alcalde Dionisio Vélez , periodo Director :,” 2015.

[28] C. I. García Osorio, “Metodologías de desarrollo de proyectos de minería de datos - Una visión centrada en CRISP-DM,” Univ. Burgos, no. May, 2019, doi: 10.13140/RG.2.2.34208.02566.

[29] C. Pete et al., “Crisp-Dm 1.0,” Cris. Consort., p. 76, 2000.

[30] D. Corzo, “Modelo conceptual para determinar el próximo presidente de estado a través de datamining y twitter”, 2018.

[31] M. Fachrie “Machine learning for data classification in indonesia regional elections based on political parties support”, 2020.

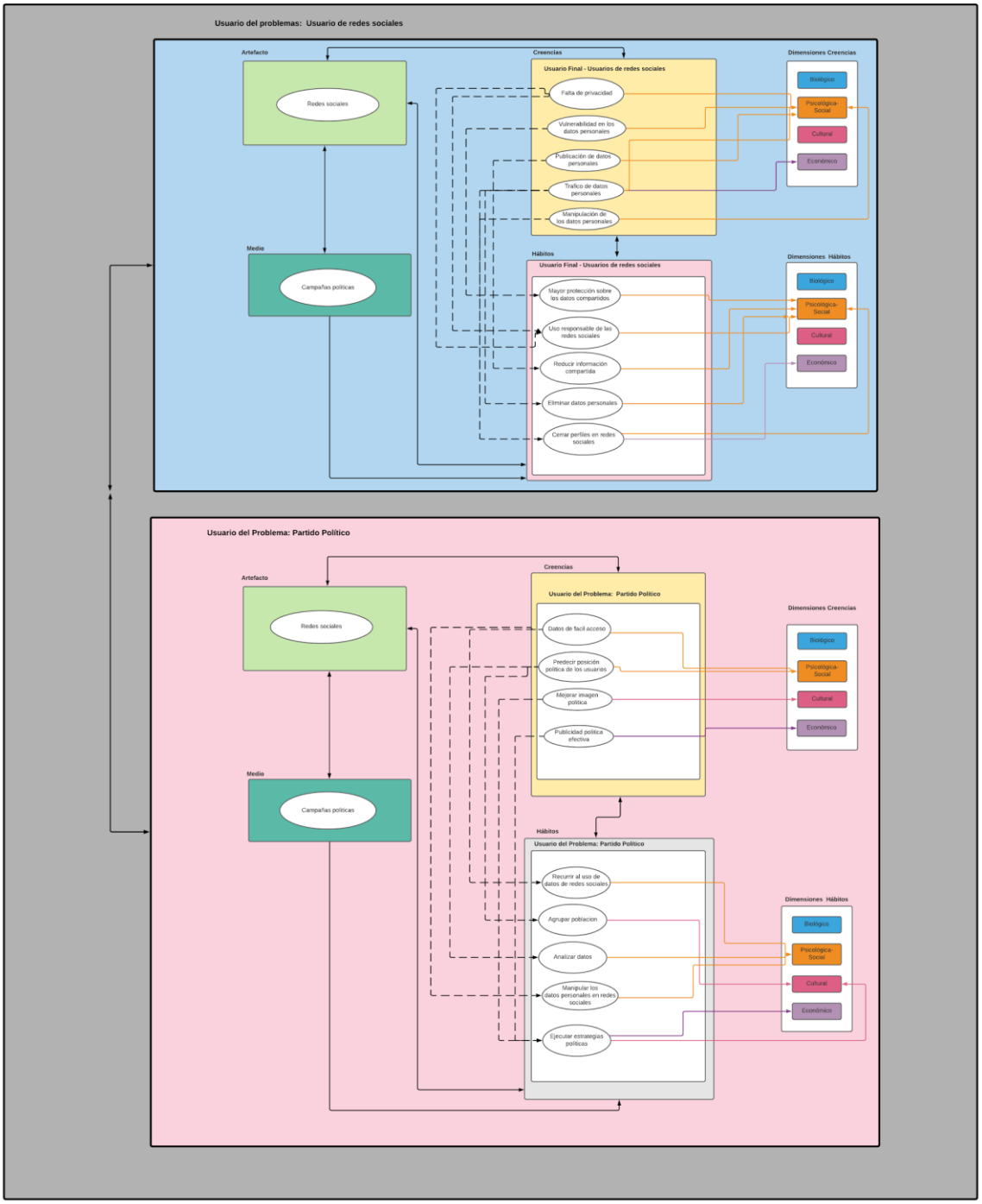
[32] P. Casas, “Libro vivo de ciencias de datos ,” 2019.

[33] Gilbert, E “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text” 2013

# ANEXOS

## ANEXO A: ANÁLISIS DEL CONTEXTO DESDE EL MODELO BIOPSICOSOCIAL Y CULTURAL

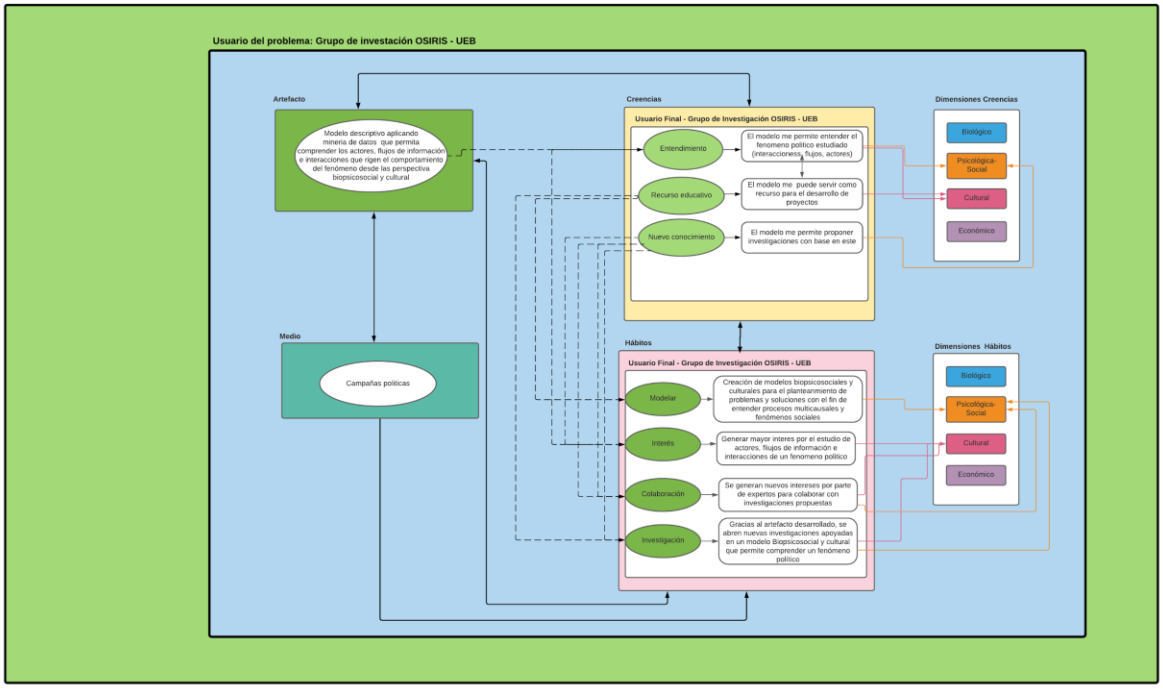
**Análisis del Modelo BPSC (Análisis del Contexto Actual y Problemática/Necesidad/Oportunidad)**

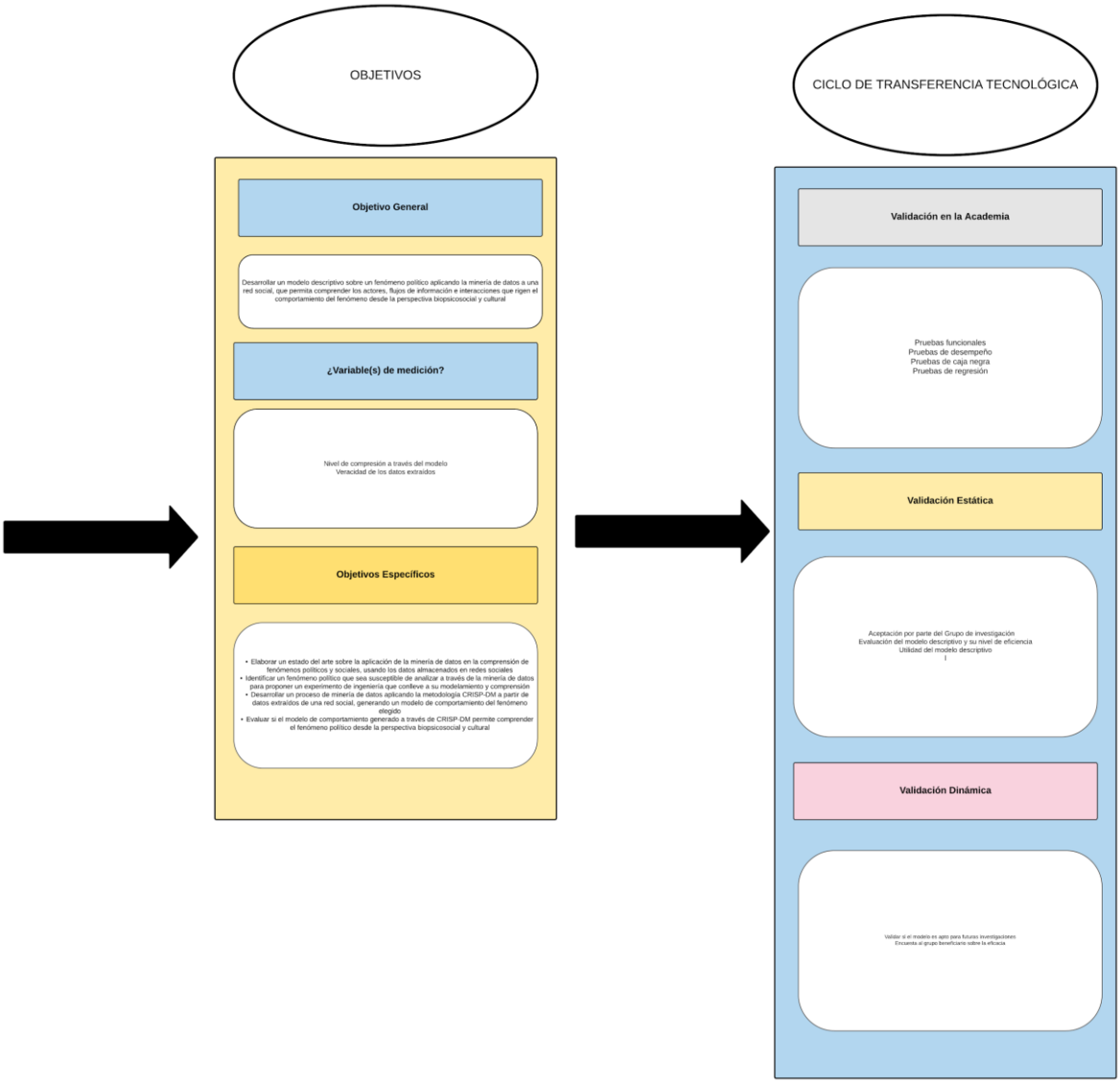


# ANEXO B: ANÁLISIS DE LA SOLUCIÓN DESDE EL MODELO BPSC

ANÁLISIS DEL CONTEXTO Y DE LA SOLUCIÓN

**Análisis del Modelo BPSC (Análisis de la Solución)**




OBJETIVOS
Objetivo General
Desarrollar un modelo descriptivo sobre un fenómeno político aplicando la minería de datos a una red social, que permita comprender los actores, flujos de información e interacciones que rigen el comportamiento del fenómeno desde la perspectiva biopsicosocial y cultural
¿Variable(s) de medición?
Nivel de compresión a través del modelo
Veracidad de los datos extraídos
Objetivos Específicos
• Elaborar un estado del arte sobre la aplicación de la minería de datos en la comprensión de fenómenos políticos y sociales, usando los datos almacenados en redes sociales
• Identificar un fenómeno político que sea susceptible de analizar a través de la minería de datos para proponer un experimento de ingeniería que conlleve a su modelamiento y comprensión
• Desarrollar un proceso de minería de datos aplicando la metodología CRISP-DM a partir de datos extraídos de una red social, generando un modelo de comportamiento del fenómeno elegido
• Evaluar si el modelo de comportamiento generado a través de CRISP-DM permite comprender el fenómeno político desde la perspectiva biopsicosocial y cultural
CICLO DE TRANSFERENCIA TECNOLÓGICA
Validación en la Academia
Pruebas funcionales
Pruebas de desempeño
Pruebas de caja negra
Pruebas de regresión
Validación Estática
Aceptación por parte del Grupo de investigación
Evaluación del modelo descriptivo y su nivel de eficiencia
Utilidad del modelo descriptivo
Validación Dinámica
Validar si el modelo es apto para futuras investigaciones
Encuesta al grupo beneficiario sobre la eficacia

## ANEXO C. ECUACIÓN DE BÚSQUEDA

### PREGUNTA ORIENTADORA DE LA BÚSQUEDA:

¿Qué trabajos o investigaciones hay sobre el uso de datos personales relacionados con fines políticos, campañas o elecciones donde se haga uso de herramientas como minería de datos, aprendizaje de máquina o técnicas estadísticas?

### Núcleos temáticos

- **Personal information**
  *"Personal data"*
  *"Personal information"*
  *"Mass communication psychology"*

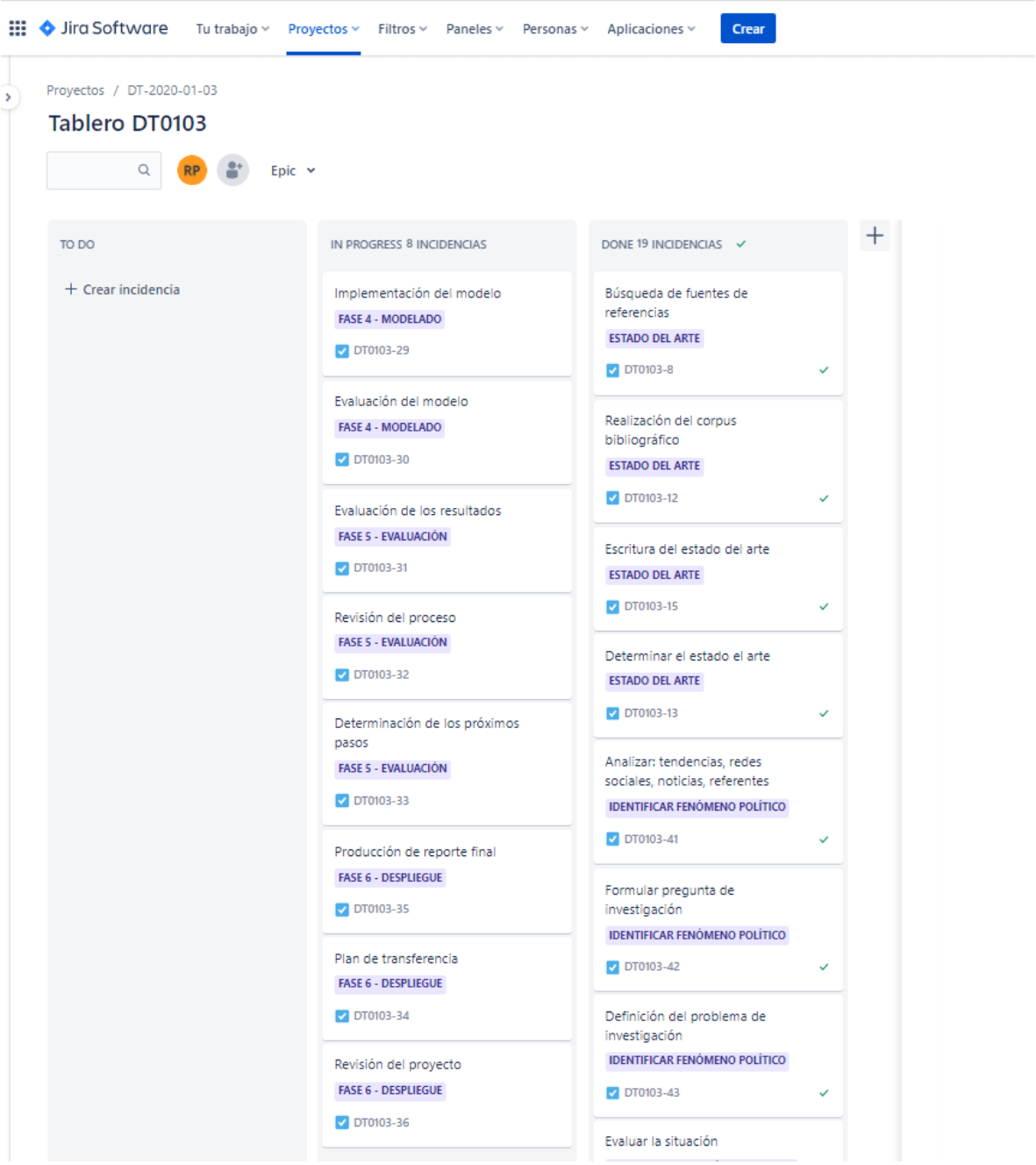
- **Politics**
  *"Political phenomena"*
  *"Political campaigns"*
  *"Campaigns"*
  *"Political Strategies"*
  *"Presidential campaigns"*
  *"Political Elections"*
  *"Political Marketing"*

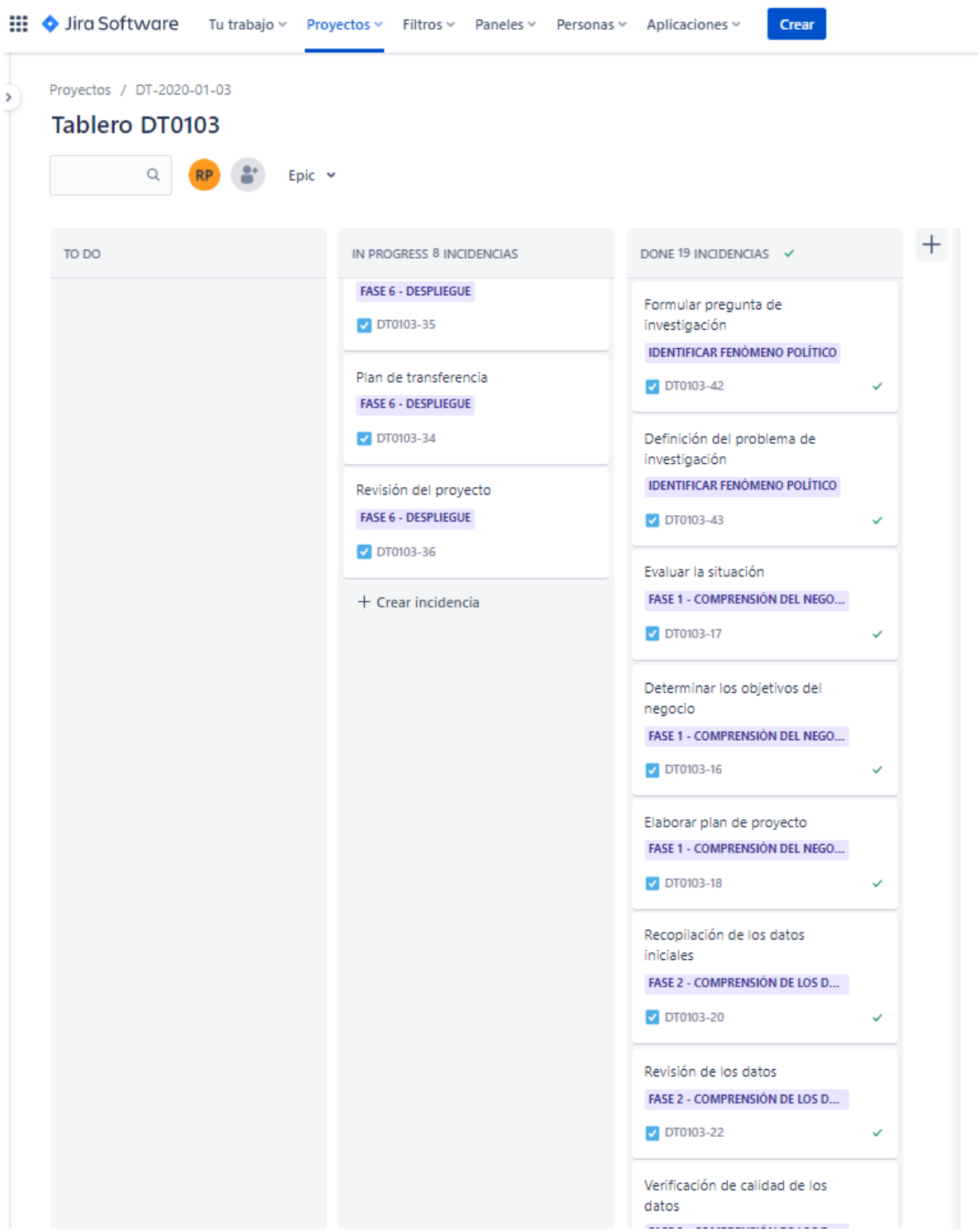
- **Minería de datos, aprendizaje de máquina, técnicas estadísticas**
  "Data mining"
  "Machine learning"
  "Statistics"
  "Social Networks"

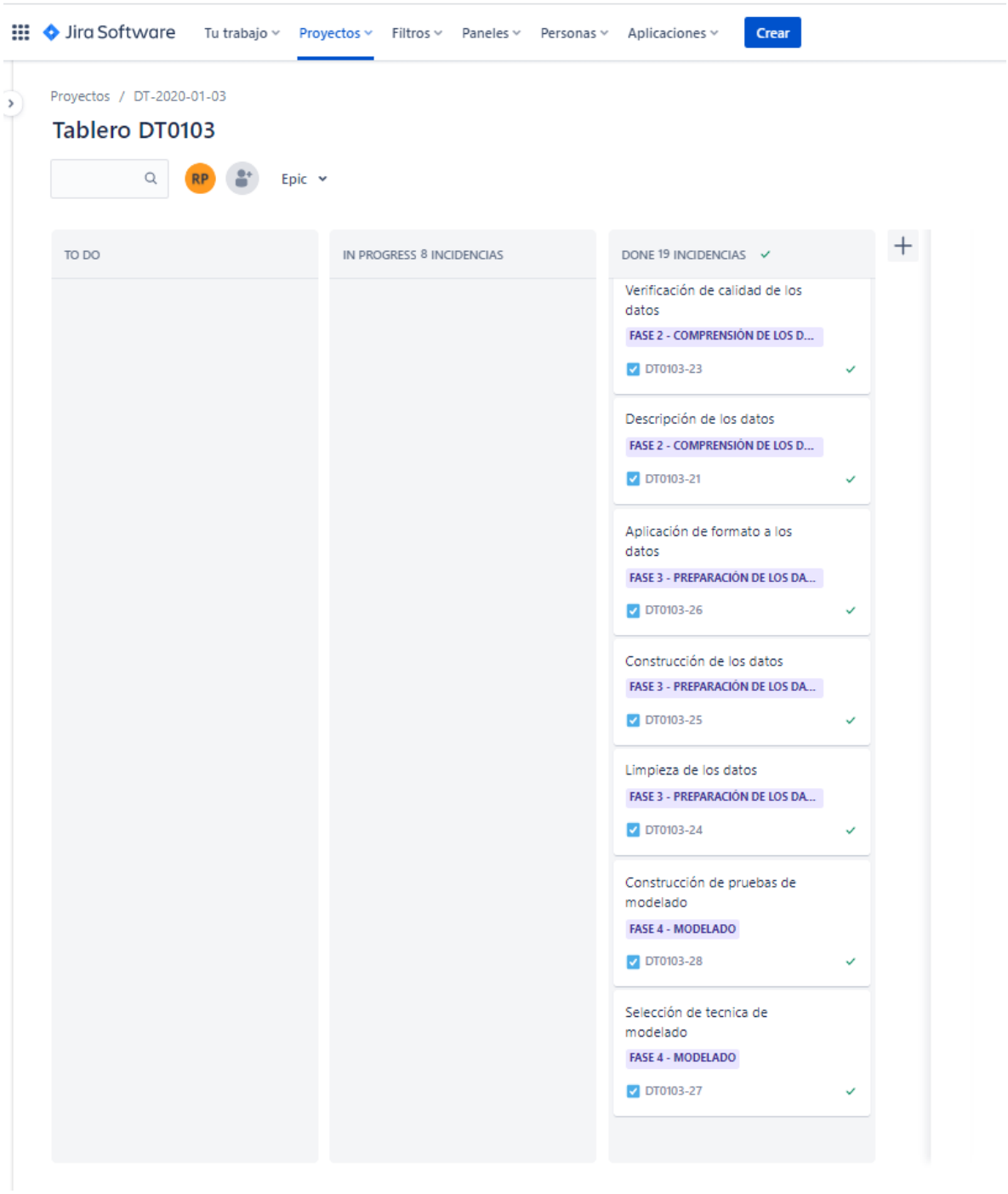
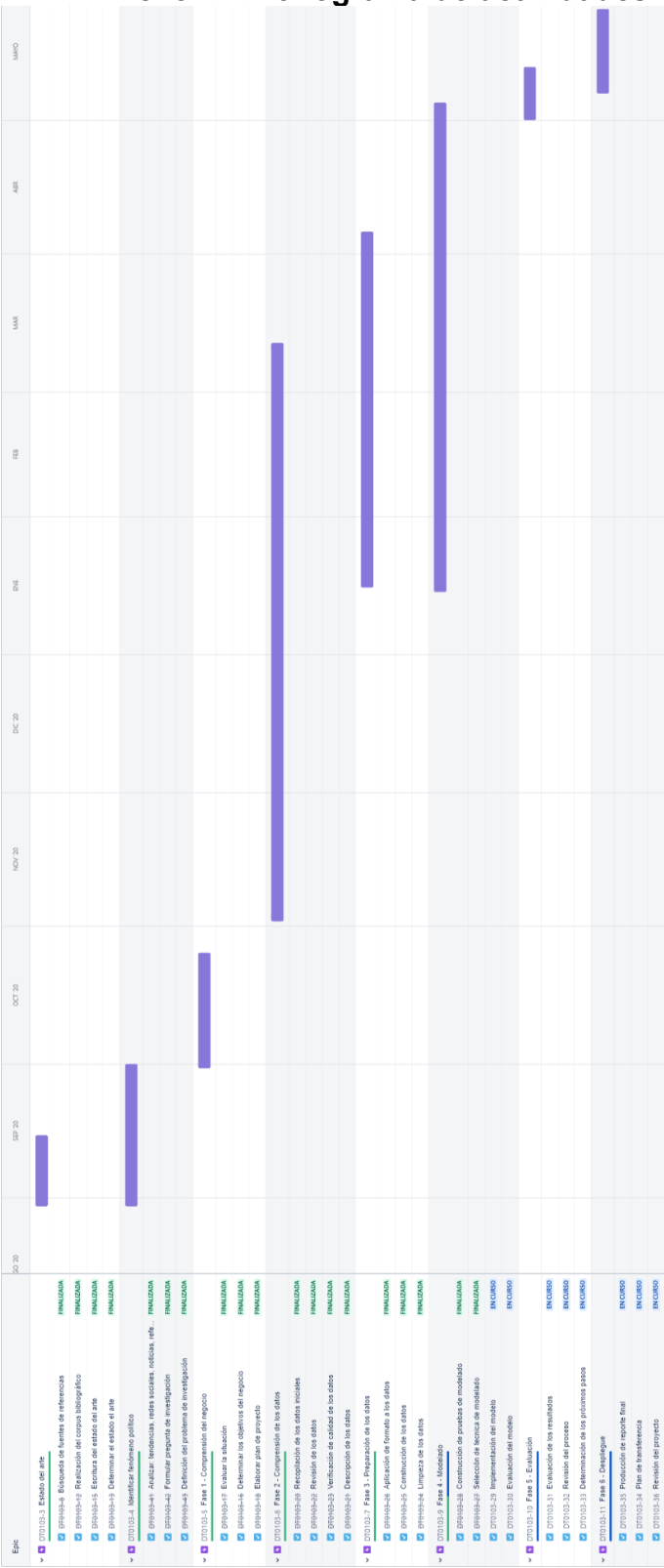
### Ecuación de búsqueda

("political phenomena" OR "political campaign" OR "presidential campaigns" OR "campaigns" OR "political strategies" OR "political elections" OR *"Political Marketing"*) AND ("personal data" OR "personal information" OR "social networks data" OR "mass communication psychology") AND ("data mining" OR "machine learning" OR statistics)

## ANEXO D: TABLERO DE CONTROL




Jira Software
Tu trabajo
Proyectos
Filtros
Paneles
Personas
Aplicaciones
Crear
Proyectos / DT-2020-01-03
Tablero DT0103
RP
Epic
TO DO
IN PROGRESS 8 INCIDENCIAS
DONE 19 INCIDENCIAS
FASE 6 - DESPLIEGUE
DT0103-35
Plan de transferencia
FASE 6 - DESPLIEGUE
DT0103-34
Revisión del proyecto
FASE 6 - DESPLIEGUE
DT0103-36
Crear incidencia
Formular pregunta de investigación
IDENTIFICAR FENÓMENO POLÍTICO
DT0103-42
Definición del problema de investigación
IDENTIFICAR FENÓMENO POLÍTICO
DT0103-43
Evaluar la situación
FASE 1 - COMPRENSIÓN DEL NEGO...
DT0103-17
Determinar los objetivos del negocio
FASE 1 - COMPRENSIÓN DEL NEGO...
DT0103-16
Elaborar plan de proyecto
FASE 1 - COMPRENSIÓN DEL NEGO...
DT0103-18
Recopilación de los datos iniciales
FASE 2 - COMPRENSIÓN DE LOS D...
DT0103-20
Revisión de los datos
FASE 2 - COMPRENSIÓN DE LOS D...
DT0103-22
Verificación de calidad de los datos


Jira Software
Tu trabajo
Proyectos
Filtros
Paneles
Personas
Aplicaciones
Crear
Proyectos / DT-2020-01-03
Tablero DT0103
RP
Epic
TO DO
IN PROGRESS 8 INCIDENCIAS
DONE 19 INCIDENCIAS
Verificación de calidad de los datos
FASE 2 - COMPRENSIÓN DE LOS D...
DT0103-23
Descripción de los datos
FASE 2 - COMPRENSIÓN DE LOS D...
DT0103-21
Aplicación de formato a los datos
FASE 3 - PREPARACIÓN DE LOS DA...
DT0103-26
Construcción de los datos
FASE 3 - PREPARACIÓN DE LOS DA...
DT0103-25
Limpieza de los datos
FASE 3 - PREPARACIÓN DE LOS DA...
DT0103-24
Construcción de pruebas de modelado
FASE 4 - MODELADO
DT0103-28
Selección de tecnica de modelado
FASE 4 - MODELADO
DT0103-27

# Anexo E: Cronograma de actividades