UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

TEMA:

Creación de un modelo de minería de datos que identifica sentimientos, de los clientes de la banca privada con calificación de riesgo mayor a AA de la ciudad de Guayaquil con datos basados en Twitter.

AUTOR:

Roy Steven Mieles Romero

Trabajo de Titulación previo a la obtención del título de

INGENIERO EN SISTEMAS COMPUTACIONALES

TUTOR:

ING. ERAZO AYÓN, JOSÉ MIGUEL MGS

GUAYAQUIL-ECUADOR

2022

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

CERTIFICACIÓN

Certificamos que el presente trabajo de titulación fue realizado en su totalidad por el Sr. Mieles Romero, Roy Steven como requerimiento para la obtención del título de **INGENIERO EN SISTEMAS COMPUTACIONALES.**

TUTOR (A)

f. _____

Ing. Erazo Ayón, José Miguel MGS.

DIRECTORA DE CARRERA

f. _____

Ing. Ana Camacho Coronel, Mgs

Guayaquil, a los 21 del mes de septiembre del año 2022

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

DECLARACIÓN DE RESPONSABILIDAD

Yo, Mieles Romero Roy Steven

DECLARO QUE:

El Trabajo de Integración Curricular, "**Creación de un modelo de minería de datos que identifica sentimientos, de los clientes de la banca privada con calificación de riesgo mayor a AA de la ciudad de Guayaquil con datos basados en Twitter.''** previo a la obtención del título de **INGENIERO EN SISTEMAS COMPUTACIONALES** ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Integración Curricular referido.

Guayaquil, a los 21 del mes de septiembre del año 2022

_____

Mieles Romero Roy Steven

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE SISTEMAS COMPUTACIONALES

AUTORIZACIÓN

Yo, **Mieles Romero Roy Steven**

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Titulación, "**Creación de un modelo de minería de datos que identifica sentimientos, de los clientes de la banca privada con calificación de riesgo mayor a AA de la ciudad de Guayaquil con datos basados en Twitter."** cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los 21 del mes de septiembre del año 2022

EL AUTOR:

_____

Mieles Romero Roy Steven

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA INGENIERÍA EN SISTEMAS COMPUTACIONALES

TRIBUNAL DE SUSTENTACIÓN

f. _____

ING. ANA CAMACHO CORONEL, MGS

DIRECTORA DE CARRERA

f. _____

ING. CESAR SALAZAR TOVAR, MGS

DOCENTE DE LA CARRERA

f. _____

ING. BYRON YONG YONG, MGS

OPONENTE

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

REPORTE URKUND

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN SISTEMAS COMPUTACIONALES

REPORTE URKUND

URKUND

Documento	Tesis Roy Mieles Final (1).docx (D143572893)
Presentado	2022-09-04 07:16 (-05:00)
Presentado por	jose.erazo@cu.ucsg.edu.ec
Recibido	jose.erazo.ucsg@analysis.urkund.com
	3% de estas 32 páginas, se componen de texto presente en 7 fuentes.

Fecha de elaboración: 2 de septiembre del 2022

Firma:

Nombre del tutor: José Miguel Erazo Ayón
Tutor de Trabajo de Titulación
Carrera de Ingeniería en Sistemas Computacionales

ÍNDICE DE CONTENIDO

ÍNDICE DE CONTENIDO VII

ÍNDICE DE FIGURAS..... VIII

RESUMEN..... XI

ABSTRACT XII

INTRODUCCIÓN.....2

CAPÍTULO I.....4

EL PROBLEMA 4

PLANTEAMIENTO DEL PROBLEMA.....4

CAPÍTULO II.....8

MARCO TEÓRICO.....8

CAPÍTULO III.....31

METODOLOGÍA DE LA INVESTIGACIÓN 31

CAPÍTULO IV 37

PROPUESTA TECNOLOGICA.....37

CONCLUSIONES.....86

RECOMENDACIONES.....88

REFERENCIAS BIBLIOGRAFICAS.....89

ÍNDICE DE FIGURAS

Figura 1 19

Figura 2 43

Figura 3 45

Figura 4 46

Figura 5 48

Figura 6 49

Figura 7 50

Figura 8 51

Figura 9 53

Figura 10 54

Figura 11 55

Figura 12 60

Figura 13 61

Figura 14 62

Figura 15 63

Figura 16 64

Figura 17 66

Figura 18 68

Figura 19 70

Figura 20 72

Figura 21 74

Figura 22 84

INDICE DE TABLAS

Tabla 1 Comparación de lenguajes de programación para el desarrollo de minería de datos 38

Tabla 2 Comparación de herramientas de visualización de datos parte 1. ..40

Tabla 3 Comparación de herramientas de visualización de datos parte 2. ..41

Tabla 4 Comparación de herramientas de visualización de datos parte 3. ..42

Tabla 5 Reglas que se usaron para etiquetar los tweets en base a si son negativos o positivos del conjunto de datos obtenido lista 1 57

Tabla 6 Reglas que se usaron para etiquetar los tweets en base a si son negativos o positivos del conjunto de datos obtenido lista 2 58

Tabla 7 Reglas que se usaron para etiquetar los tweets en base a si son negativos o positivos del conjunto de datos obtenido lista 3 59

Tabla 8 Resultados de la precisión de los algoritmos ejecutados en la primera iteración.....76

Tabla 9 *Resultados de la exhaustividad de los algoritmos ejecutados en la primera iteración*.....77

Tabla 10 Resultados del F1-Score de los algoritmos ejecutados en la primera iteración.....77

Tabla 11 Resultados de la precisión de los algoritmos ejecutados en la segunda iteración 78

Tabla 12 Resultados de la exhaustividad de los algoritmos ejecutados en la segunda iteración 79

Tabla 13 Resultados del F1-Score de los algoritmos ejecutados en la segunda iteración.....79

Tabla 14 Resultados de la precisión de los algoritmos ejecutados en la tercera iteración.....80

Tabla 15 Resultados de la exhaustividad de los algoritmos ejecutados en la tercera iteración.....81

Tabla 16 Resultados del F1-Score de los algoritmos ejecutados en la tercera iteración.....81

RESUMEN

Dentro del presente proyecto de investigación se encuentra la creación de un modelo de minería de datos que tenga la capacidad de realizar un análisis de sentimientos basándose en datos acerca de la banca privada ecuatoriana el cual tendrá poder predictivo para asignar los valores de “Positivo”, “Negativo” o “Neutro” hacia los comentarios que sean procesados por este. Esto se hará haciendo uso de herramientas de programación con las que se pueda realizar una limpieza de datos, procesamiento del lenguaje natural, y por último en el que se puedan aplicar algoritmos de aprendizaje de máquina para que se realice el entrenamiento del modelo. Esto es con la finalidad de poder tener una medición de los niveles de satisfacción de los usuarios de la banca privada ecuatoriana de manera automática y que cuyos resultados puedan ser luego visualizados dentro de una herramienta de visualización de datos. Los objetivos del presente proyecto de investigación abarcan la creación del conjunto de datos que será utilizado, el diseño del modelo de minería de datos y la creación de este. El proyecto finaliza con la comparación de los rendimientos de los diferentes algoritmos de clasificación de aprendizaje de máquina, la selección del mejor de estos para el desarrollo del de modelo y la presentación de los resultados de la capacidad predictiva del modelo desarrollado.

Palabras Clave: *Minería de datos, Aprendizaje de máquina, Algoritmo de clasificación, procesamiento del lenguaje natural*

ABSTRACT

Within this research project is the creation of a data mining model that could perform a sentiment analysis based on data about Ecuadorian private banking which will have predictive power to assign values of "Positive", "Negative" or "Neutral" to the comments that are processed by it. This will be done by making use of programming tools with which data cleaning, natural language processing, and finally in which *machine learning* algorithms can be applied to perform the training of the model. This is in order to be able to measure the satisfaction levels of users of Ecuadorian private banks automatically and whose results can then be visualized within a data visualization tool. The objectives of this research project include the creation of the data set to be used, the design of the data mining model and its creation. The project ends with the comparison of the performances of the different *machine learning* classification algorithms, the selection of the best of these for the development of the model and the presentation of the results of the predictive capability of the developed model.

Key words: *Data mining, Machine learning, Classification algorithm, Natural language processing*

INTRODUCCIÓN

En la actualidad el alcance que las personas tienen hacia las redes sociales es instantáneo, teniendo así la oportunidad de expresarse libremente de cualquier tema con el mundo, ya sea sobre un producto, servicio, empresas o personas y a su vez existiendo la posibilidad de que la imagen de los antes mencionados se vea afectada tanto positiva como negativamente. Una de las redes sociales más populares para la libre expresión es Twitter con un total de 330 millones de usuarios activos según (Ahlgren, 2022), en esta se comparten opiniones, quejas y denuncias de manera pública o privada por parte de estos sin ningún tipo de restricción o censura.

En el ámbito bancario esta red social es utilizada por sus usuarios en su mayoría para realizar quejas públicas del mal funcionamiento de los servicios otorgados por estas entidades. Con esta información se puede hacer uso de herramientas de minería de datos para conocer más a profundidad cómo se llegan a sentir los clientes con los servicios que se les ofrece.

la minería de datos otorga la posibilidad de encontrar y clasificar patrones, tendencias o reglas dentro de grandes cantidades de información para de esta forma explicar el significado de los datos en un contexto específico, este significado no se puede hallar fácilmente por medio de una exploración típica de los datos debido a la complejidad de estos, o por la gran cantidad de datos existentes. (syloper, nd)

De esta forma se puede hacer uso de la herramienta para realizar un procesamiento del lenguaje natural, es decir la forma en la que las personas nos comunicamos, y obtener un análisis de sentimientos de los clientes sobre

la banca ecuatoriana encontrados dentro de los datos que pueden ser obtenidos mediante Twitter para generar un mayor panorama de cómo se sienten con respecto al servicio que se les ha otorgado, en donde se encuentran sus mayores dolencias o preocupaciones y cuál organización brinda una mejor experiencia hacia el cliente.

Para finalizar, dentro del capítulo 1.- Se describe el problema, su ubicación, sus causas y consecuencias, se definen el objetivo general y objetivos específicos, el alcance, la justificación y la importancia; Capítulo 2.- En este capítulo se presenta el marco teórico en el cual se argumentan los conceptos, normas, estándares, leyes y reglamentos que harán de soporte para el presente trabajo de investigación; Capítulo 3.- Se encontrará descrita la metodología de investigación, además se dimensiona la población y se especifican los instrumentos de recolección de datos; Capítulo 4.- En este capítulo se presentará la propuesta tecnológica del trabajo de titulación, en el que se detallarán las herramientas utilizadas, las técnicas del procesamiento de datos y otros aspectos utilizados en el desarrollo; Conclusiones y recomendaciones.- Por último, se muestran los resultados del trabajo de titulación y se dará respuesta a los objetivos y propósitos planeados.

CAPÍTULO I

EL PROBLEMA

PLANTEAMIENTO DEL PROBLEMA

1.1 Planteamiento del problema

Hasta la fecha de diciembre del 2021 existe un total de 24 instituciones bancarias privadas que proveen sus servicios a toda la población ecuatoriana, de los cuales 10 poseen una calificación de riesgo institucional AAA +/- según lo anuncia la (Superintendencia de Bancos, 2021), lo cual implica que estas son instituciones fuertes con una excelente trayectoria de rentabilidad y claras perspectivas de estabilidad (Jiménez, 2014), y así un cliente puede tomar esta calificación como una medida para decidir que banco elegir al momento de depositar su dinero.

Sin embargo, por medio de las redes sociales es posible encontrar múltiples quejas por parte de los clientes de las instituciones bancarias antes mencionadas con respecto a sus servicios tales como: aplicativo móvil, funcionamiento de los cajeros automáticos, redes de agencias, canales de atención al cliente, servicios de retiros de dinero, puntos de pago, servicios de tarjeta de crédito, gestión de cobros, entro otros.

Por lo cual únicamente con la calificación de riesgo no se puede saber a ciencia cierta si una institución bancaria puede ofrecer un correcto servicio a sus clientes o no, y así mismo no hay una herramienta a la que los nuevos clientes puedan dirigirse para que se les facilite la visualización de la reputación y niveles de satisfacción que actualmente posee las entidades bancarias en base a como se siente su clientela actual dentro de las redes

sociales y poder tomar una decisión adecuada a la hora de confiar su dinero a alguna de estas instituciones.

En el año 2017, en la universidad politécnica Salesiana de ecuador se realizó un trabajo de investigación realizado por (Ramírez, 2017) en el que se medían los niveles de satisfacción de los clientes de la banca privada de guayaquil, pero únicamente dirigidos hacia los canales de atención de reclamos; con esta investigación no se evidencia con claridad la visibilidad la banca dentro de las redes sociales, además este estudio se quedó estancado en el 2017, debido a que no tiene una actualización continua con información de los clientes, al contrario a lo que pasaría con un sistema en el que se esté descargando información de la opinión de los usuarios a diario, mensual o anual dependiendo de cómo se desee realizar el estudio.

Por lo cual se puede observar que no existe una herramienta con el poder de predecir de manera computacional el nivel de satisfacción sobre una opinión otorgada hacia las entidades bancarias de guayaquil que posean una calificación de riesgo mayor que AA, y así mismo tampoco un estudio que describa su desarrollo y compruebe su funcionamiento y eficacia.

1.2 Preguntas de Investigación

¿Es posible crear un modelo de minería de datos para realizar un análisis de sentimientos con información obtenida de Twitter?

1.3 Objetivos de la investigación

1.3.1 Objetivo General

Crear un modelo de minería de datos para hacer un análisis de sentimientos con respecto a entidades bancarias privadas con calificación de riesgo mayor que AA de la ciudad de Guayaquil basado en datos obtenidos de la red social Twitter.

1.3.2 Objetivos Específicos

- Obtener información con base a la red social Twitter en relación con las entidades bancarias con calificación de riesgo mayor que AA de la ciudad de Guayaquil.

- Diseñar una estrategia de minería de datos para identificar los sentimientos de los clientes de las entidades bancarias con calificación de riesgo mayor que AA de la ciudad de Guayaquil con datos recopilados desde Twitter.

- Construir un modelo de minería de datos aplicando algoritmos de aprendizaje de máquina para clasificar los tweets basado en los sentimientos de los clientes de las entidades bancarias privadas con calificación de riesgo mayor que AA de la ciudad de Guayaquil.

1.4 Justificación y alcance

El análisis de sentimiento de texto es una herramienta que permite extraer el tono emocional que existen dentro de una frase o párrafo, este con la finalidad de identificar si se trata de una opinión positiva, negativa o neutra de algún tema, producto, servicio o entidad en específico de la que se esté hablando en el texto. (Herrero, 2016)

Al momento de diseñar un modelo de análisis de sentimientos se combinan distintas áreas como lo son la minería de datos, algoritmos de clasificación, la selección de atributos, y el procesamiento de lenguaje natural, este último se encarga de convertir el texto lingüístico en un lenguaje que la máquina pueda comprender y así mismo el *big data* recauda grandes cantidades de información para obtener resultados más precisos, ajustando y mejorando el desempeño de los algoritmos. (Bannister, 2015)

Es así como se pretende crear un modelo de minería de datos sobre el análisis de sentimientos que otorgue la capacidad de comprender la opinión y analizar los niveles de satisfacción que poseen los usuarios de la red social Twitter con respecto a la banca ecuatoriana. A su vez para el desarrollo de este se desea recaudar información de los tweets con la opinión de los usuarios para el desarrollo del algoritmo de aprendizaje de máquina que realice el análisis de sentimientos deseado basado en minería de datos.

CAPÍTULO II
MARCO TEÓRICO

Para el desarrollo de este proyecto es importante comprender los conceptos, las definiciones, y las teorías existentes que se van a ser necesarios para la investigación en la cual se encuentran temas tales como la minería de datos, las redes sociales, procesamiento del lenguaje natural, el análisis de sentimientos, entre muchos otros más que estarán siendo desarrollados a lo largo de este capítulo.

2.1 Minería de datos

Uno de los pilares fundamentales de este proyecto de investigación es la minería de datos la cual es una ciencia que tiene sus inicios en los años sesenta y ha estado en constante estudio desde entonces, esta hace referencia a un conjunto de algoritmos que permiten la identificación de patrones útiles y novedosos que se encuentran “ocultos” en grandes bases de datos (Vallejo Ballesteros, Guevara Iñiguez , & Medina Velasco, 2017).

A su vez también se puede definir como el proceso de encontrar una estructura interesante en los datos, la estructura puede adoptar muchas formas, como un conjunto de reglas, un gráfico o una red, un árbol, una o varias ecuaciones, etc. (Roiger, 2017). Con esto podemos entender a la minería de datos como un método innovador para sacarle el mayor provecho posible a los datos de formas en las que den resultados que sean de alto valor para el consumidor de estos datos.

Otro termino con el que se conoce a la minería de datos es conocimiento descubierto desde la data, o por sus siglas en inglés KDD (knowledge

Discovery from data), esta se trata de una secuencia iterativa de los siguientes pasos (Han, Pei, & Kamber, 2011):

- **Limpieza de datos:** Proceso en que se remueven inconsistencias y ruido de la data.
- **Integración de los datos:** Proceso en el que múltiples fuentes de datos pueden ser combinados.
- **Selección de los datos:** Proceso en el que la data relevante para el análisis que se desea desarrollar es obtenida de las bases de datos.
- **Transformación de los datos:** Proceso en el que los datos se transforman y consolidan en formas apropiadas para la minería mediante la realización de operaciones de resumen o agregación.
- **Minería de datos:** Proceso esencial en el que métodos inteligentes son aplicados para extraer los patrones de los datos.
- **Evaluación de patrones:** Se identifican los patrones verdaderamente interesantes que representan el conocimiento basado en medidas de interés.
- **Presentación del conocimiento:** Proceso en el cual Las técnicas de visualización y representación del conocimiento se utilizan para representar el conocimiento extraído a los usuarios.

Según el trabajo de investigación realizado por (Lagla, Moreano, Arequipa, & Quishpe, 2019) en el que se estudia a la minería de datos como herramienta

estratégica, se describe al KDD como los pasos a desarrollar para la aplicación de la minería de datos en la que se siguen los siguientes:

2.1.1 Extracción de datos

El proceso de extracción de datos consiste en crear un conjunto de datos de destino, es decir que se procederá a seleccionar una fuente principal de la que se extraerán los datos que se requieren utilizar, esto centrándose en un subconjunto de variables o muestras de datos, en el que hay que realizar los descubrimientos. (Lagla, Moreano, Arequipa, & Quishpe, 2019)

2.1.2 Preprocesamiento de los datos

El preprocesamiento de los datos se trata de la limpieza de estos; es decir, se realizarán operaciones básicas de limpieza de ruido y texto siempre que sea apropiado, recogiendo la información necesaria para modelar y definir estrategias para el manejo de campos faltantes de datos según lo indican (Lagla, Moreano, Arequipa, & Quishpe, 2019); después de la aplicación de la fase de preprocesamiento, el conjunto resultante puede ser visto como una fuente consistente y adecuada de datos de calidad para la aplicación de algoritmos de minería de datos. (García, Ramírez-Gallego, Luengo, & Herrera, 2016)

2.1.3 Aplicación de algoritmos

La aplicación de algoritmos de minería de datos, dentro del conjunto de datos obtenido en el primer paso de las técnicas de recolección de datos, se trata de la búsqueda de patrones de interés de una representación particular o en un conjunto de tales representaciones. Esto haciendo uso de algoritmos tales como reglas de clasificación, árboles de decisión, regresión, agrupación, entre otros. (Lagla, Moreano, Arequipa, & Quishpe, 2019)

2.1.4 Interpretación de resultados

La interpretación de los resultados consiste esencialmente en evaluar, en sí, al modelo de minería de datos desarrollado, enfocándose principalmente en los algoritmos implementados dentro del modelo, en el cual se tomarán en consideración criterios de evaluación en el que se aplican declaraciones cuantitativas, también llamados funciones de ajustes, sobre los objetivos del proceso de minería de datos y sus parámetros. (Lagla, Moreano, Arequipa, & Quishpe, 2019)

Por ejemplo, los modelos de predicción son evaluados por medio de la precisión de las predicciones empíricas que este realiza sobre conjunto de datos de prueba. Y los modelos descriptivos pueden ser evaluados, así mismo con los modelos de predicción, por medio de su precisión predictiva, pero también por su novedad, utilidad y comprensibilidad del modelo ajustado. (Lagla, Moreano, Arequipa, & Quishpe, 2019)

Para finalizar con el tema de la minería de datos se puede decir que hay muchas técnicas de minería de datos que se han desarrollado y utilizado en proyectos de esta índole, como la asociación, la clasificación, la agrupación,

el árbol de decisión, la predicción y las redes neuronales, etc. Estas son las que se encargan de encontrar patrones en los datos y darles un valor (Osman, 2019). Además de que se debe seguir un proceso en el que cada paso depende del anterior para poder aplicar la minería de datos en cualquier tipo de proyecto de esta índole.

2.2 *Machine Learning*

El *machine Learning*, por sus siglas ML, es una parte esencial de la minería de datos y este resuelve situaciones por sí solo a partir de un análisis de datos y cuantos más datos tengan mejores resultados, además, para realizar el análisis se utilizan algoritmos que diseñan otros datos según las necesidades. A través de los datos de entrada, ML ejecutar un algoritmo y como resultado, genera más información para el problema (Rojas, 2020).

Por otro lado (IBM, s.f.) indica que el ML es una forma de la IA que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. Sin embargo, ML no es un proceso sencillo, conforme el algoritmo ingiere datos de entrenamiento, es posible producir modelos más precisos basados en datos.

Existen distintos tipos de técnicas de ML y su uso se basa en el problema que queremos resolver, a continuación, se hablará de estas técnicas:

2.2.1 Aprendizaje Supervisado

Según (Rueda, 2022) Los algoritmos de aprendizaje supervisado basan su aprendizaje en un juego de datos de entrenamiento previamente etiquetados, es decir que para cada ocurrencia del juego de datos de entrenamiento conocemos el valor de su atributo objetivo. Esto le permitirá al algoritmo poder

aprender una función capaz de predecir el atributo objetivo para un juego de datos nuevo.

Es decir, un modelo de *machine learning* es supervisado cuando a este se le otorga una serie de características, que para el modelo serán preguntas, y a su vez el dato objetivo, o la respuesta de estas preguntas, de esta forma en el futuro el mismo algoritmo será capaz de predecir la respuesta conociendo únicamente las características (Sandoval, 2018), en este método se pueden encontrar 2 clasificaciones:

- **Algoritmos de clasificación:** En este algoritmo se espera que la predicción exprese a que grupo pertenecen las características que fueron utilizadas en él, encargándose de encontrar patrones en los datos y clasificándolos, Luego compara los nuevos datos y los ubica en uno de los grupos y es así como puede predecir de que se trata.

- **Algoritmos de regresión:** en este método lo que se espera es un número. No lo ubica en un grupo, sino que devuelve un valor específico.

2.2.2 Aprendizaje No Supervisado

Con este algoritmo de *machine learning,* lo que se introduce en el modelo son únicamente los campos llamados características del set de datos que se desea utilizar, pero nunca el campo objetivo. Lo que hará el algoritmo es realizar agrupaciones de las características introducidas por medio de similitudes entre ellas. (Sandoval, 2018)

2.2.3 Modelos de *Machine learning*

A estos algoritmos de *machine learning* antes mencionados se los puede clasificar en 3 tipos de modelos los cuales son según (Sandoval, 2018):

- **Modelos lineales:** Los modelos lineales se encargan de encontrar una línea que se ajuste a la serie de puntos que se disponen mediante las características que se les otorga; en el destacan modelos muy conocidos como la regresión lineal y la regresión logística. Aunque estos dos modelos se encuentran con el problema de que pueden llegar a estar sobre ajustados a los datos disponibles además al ser modelos muy simples, no ofrecen resultados favorables al momento de otorgarles comportamientos o características más complejas.

- **Modelos de árbol:** Se tratan de modelos precisos, estables y mucho más sencillos de interpretar, esto debido a que se encargan de construir una serie de reglas de decisiones que se pueden interpretar como un árbol y que, a diferencia de los modelos lineales, pueden hacer representaciones de modelos no lineales para resolver problemas.

- **Modelos Neuronales:** Se tratan de redes artificiales de neuronas, las cuales tratan de replicar el comportamiento del cerebro humano, en el cual se tienen redes de millones de neuronas para enviar mensajes entre ellas. Esta se trata de uno de los algoritmos de moda debido a las habilidades cognitivas de razonamiento que poseen tales como el reconocimiento de imágenes o videos. El mayor

problema que se encuentra en este tipo de modelo es que necesitan cantidades masivas de datos por lo cual entrenarlos es una tarea que toma más tiempo que el resto y a su vez necesitan gran capacidad de cómputo.

2.3 Análisis de sentimientos

El análisis de sentimientos es una minería contextual de texto que identifica y extrae información subjetiva en el material de origen, y que ayuda a una empresa a entender el sentimiento social de su marca, producto o servicio mientras monitoriza las conversaciones online. (Gupta, 2018).

Esta se utiliza para determinar si un texto dado contiene emociones negativas, positivas o neutras. Es una forma de análisis de texto que utiliza el procesamiento del lenguaje natural (PLN) y el aprendizaje automático. El análisis de sentimientos también se conoce como "minería de opinión" o "inteligencia artificial de las emociones" (Iglesias & Moreno, 2019).

Los algoritmos desarrollados en Python para realizar análisis de sentimiento tienen poder predictivo, la cual viene determinada por *machine learning*, este es una modalidad de inteligencia artificial que entrena a una máquina a través de *data mining* para automatizar procesos de análisis de datos (SAURA, REYES-MENENDEZ, & PALOS-SANCHEZ, 2018)

2.4 Procesamiento del lenguaje natural

El procesamiento del lenguaje natural, por sus siglas PLN, es el conjunto de métodos para hacer accesible el lenguaje humano a los ordenadores. El PLN se centra en el diseño y análisis de algoritmos y representaciones computacionales para el procesamiento del lenguaje, el objetivo de este es

proporcionar nuevas capacidades computacionales en torno al lenguaje humano. (Eisenstein, 2018).

También, el PLN investiga el uso de los ordenadores para entender el lenguaje humano con el fin de realizar tareas útiles. El PLN es un campo interdisciplinar que combina la lingüística computacional, la informática, la ciencia cognitiva y la inteligencia artificial. (Li Deng, 2018). Este es un área de investigación y aplicación de la inteligencia artificial, la cual engloba múltiples disciplinas tales como la traducción automática de texto, síntesis automática de texto, análisis de sentimientos, dictado por voz a texto y viceversa, etc. (Montarroso, 2019)

Hoy en día, existe una infinidad de recursos abiertos para implementar las técnicas de Procesamiento de Lenguaje Natural, por ejemplo, el Natural Language Toolkit de Python es una plataforma para construir programas que permiten manipular lenguaje natural según (Collaguazo, 2017).

Además, hoy en día con el avance de la tecnología y el arduo trabajo de *Google* se tiene al *Transformer,* el cual se trata de una arquitectura de redes neuronales que se considera en estado de arte de modelos secuenciales, el *transformer* en PLN resuelve tareas de secuencia a secuencia sin problemas de dependencia que presentan redes neuronales recurrentes, La idea clave de esta arquitectura es gestionar las dependencias entre la entrada y la salida con atención y recurrencia, así lo dice (The Machine Learners, 2022)

2.5 Base de datos

(Valverde, Portalanza, & Mora, 2019) Indican que las bases de datos son grandes cantidades de información almacenadas en registros para lograr una

mejor eficiencia al momento de ingresar, buscar, actualizar o eliminar la información. Esta información debe estar interrelacionada para evitar la duplicidad de información y mejorar su organización.

De esta forma una base de datos SQL es aquella base de datos relacional que está escrita en lenguaje SQL (lenguaje de consulta estructurado), también pronunciado “sequel”. Este lenguaje se considera el lenguaje estándar para las bases de datos según el ANSI (Instituto Nacional Americano de Estándares). Para hacer una base de datos se necesitan casi únicamente los comandos básicos de SQL como: “Seleccionar”, “Insertar”, “Actualizar”, “Eliminar”, “Crear” y “Eliminar”, así lo informa la (European Knowledge Center for Information Technology, 2019).

Con estos conceptos se puede entender fácilmente que una base de datos serán grandes cantidades de información que pueden mantener relaciones entre sí y se encontrarán almacenadas dentro de un sistema en el que cual es posible la ejecución de diferentes tipos de comandos para su visualización y manipulación.

2.6 Visualización de datos

El proceso de la visualización de datos, como su nombre lo indica, es crear una representación visual de la información que tenemos recopilada en la cual se representen tendencias, patrones, y perspectivas críticas de esta. Existen muchos tipos de gráficos de visualización como, por ejemplo: gráficos circulares, gráfico de barras, tablas, histogramas, diagramas de Grant, entre otros, así lo indica (Duò, 2022)

En la propia página de (Tableu, s.f.), una de las herramientas más utilizadas de visualización de datos, se describe que la visualización de datos es otra forma de arte visual que capta el interés de los usuarios y mantiene sus ojos en el mensaje. Cuando se ve un gráfico, se observa rápidamente las tendencias y los valores atípicos. Al ver la información de forma gráfica esta de interioriza de forma rápida. Es contar historias con un propósito.

¿Pero cómo es posible realizar gráficos para la visualización de datos?, existe los bien llamados herramientas de visualización de datos, las cuales son de uso para esta tarea como otras de análisis de datos, en el mercado, según el cuadrante mágico de Gartner de inicios del 2022 que podemos observar en la figura 3, las empresas líderes en visualización de datos son Microsoft con su herramienta PowerBi, Salesforce con la herramienta Tableu y Qlik con QlikView.

Figura 1

Cuadrante mágico de gartner en que se muestran herramientas de visualización de datos

Nota. En este gráfico se describe la posición en la que se encuentran las distintas herramientas de visualización de datos dentro del mercado, gráfico extraído de (Gartner, 2022)

2.7 Lenguajes de programación

Según (Mendoza, 2020) los lenguajes de programación son un conjunto de instrucciones con los cuales el humano interactúa con la computadora, estos nos permiten interactuar con las computadoras mediante algoritmos e

instrucciones escritas en una sintaxis que permita a la computadora entender lo que queremos realizar y lo pueda ejecutar mediante un compilador.

De acuerdo con Mendoza (2020) se pueden categorizar a los lenguajes de programación en tres tipos como:

- **Lenguaje de máquina:** Se trata del más primitivo de los códigos y se basa únicamente en numeración binaria y es utilizado directamente por las computadoras o máquinas.
- **Lenguajes de bajo nivel:** Este puede variar según el computador o máquina que lo esté utilizando
- **Lenguaje de alto nivel:** Se caracteriza en utilizar comandos y palabras normalmente escritas en inglés las cuales son de fácil entendimiento para el ser humano y las puede utilizar para dar las instrucciones al computador en el que desee usar este lenguaje.

2.8 Sistema financiero

Según el (Banco Internacional, 2021), el principal objetivo del Sistema Financiero es canalizar el ahorro de las personas y contribuir de forma directa en el sano desarrollo económico del país, está conformado por todas las instituciones bancarias públicas o privadas, mutualistas, o cooperativas, legalmente constituidas en el país. Es decir, el rol de las entidades financieras es transformar el ahorro de unas personas, en la inversión de otras, administrando adecuadamente los riesgos correspondientes.

Así mismo el (Banco Mundial, s.f.) en su página acerca de los sectores financieros indica que estos sustentan el crecimiento económico y el

desarrollo, un buen acceso a esta mejora el bienestar general de un país, debido a que permite a su población prosperar y gestionar de mejor manera sus necesidades financieras como por ejemplo facilitando las gestiones de consumo, pagos y ahorro de dinero.

Según la (Corporación Financiera Nacional, 2017) entre los productos y servicios que el sistema financiero nacional ofrece a la población del ecuador se encuentran los siguientes:

- **Crédito de consumo:** Este se trata de un préstamo a corto o mediano plazo en el cual se otorga dinero de libre disposición el cual permite la adquisición de bienes de consumo, por lo general mantiene una tasa alta debido al riesgo que implica.

- **Crédito comercial:** Su objetivo es satisfacer las necesidades de efectivo de empresas de cualquier tamaño el cual suele ser pactado para ser pagado a corto o mediano plazo.

- **Crédito de vivienda:** También conocido como crédito hipotecario, este es otorgado para la adquisición de una vivienda o para reparaciones y remodelaciones del hogar propio, este crédito se otorga con pagos a largo plazo.

- **Microcrédito:** Este crédito es otorgado a prestatarios ya sean personas naturales o jurídicas y tiene como finalidad financiar actividades en pequeña escala de producción, comercialización o servicios.

- **Cuentas corrientes:** Al obtener una cuenta corriente se establece un contrato entre el usuario acreedor de la cuenta y la institución financiera en el que se acuerda que dicha entidad haga efectiva las órdenes de pago emitidas por el usuario, es decir los cheques.

- **Cuentas de ahorro:** Estas permiten al usuario acreedor de la cuenta generar ahorros con seguridad y rentabilidad, además facilita la planificación financiera y el manejo de los recursos, por lo general es el producto más solicitado por los usuarios.

- **Depósito a plazo:** Se trata de una inversión financiera en el cual el cliente entrega un monto de dinero a la entidad a un plazo determinado con una tasa de retorno fija, al concluir este plazo la institución le devuelve al cliente el monto inicial más los intereses generados.

- **Fondo de garantía:** Se trata de una herramienta que tiene como objetivo facilitar el acceso al crédito a micro, pequeños y medianos empresarios que no cuentan con los colaterales suficientes para respaldar una operación crediticia en el sistema financiero ecuatoriano.

- **Fondo de inversiones:** Es una alternativa de inversión diversificada que busca reducir el riesgo, esta se encuentra constituida por las aportaciones de diversas personas a las cuales se les denomina participes del fondo y es administrado por una sociedad gestora responsable de su gestión.

- **Negocios fiduciarios:** La legislación ecuatoriana contempla la existencia de 2 tipos de negocios fiduciarios los cuales son:

    a. **Encargo fiduciario:** Se trata de un contrato en el cual el constituyente, ya sea una persona natural o jurídica, le da instrucciones de administración a la fiduciaria para que cumpla las condiciones pactadas en la que no existe transferencias de propiedades.

    b. **Fideicomisos mercantiles:** Se trata de un contrato en el cual una o más personas llamadas constituyentes o fideicomitentes transfieren de manera temporal o irrevocable la propiedad de bienes muebles o inmuebles a otra persona, en este caso fiduciaria, para que ésta la administre o invierta los bienes en beneficios propios o de un tercero.

- **Tasas:** La tasa de interés es el precio del dinero que un inversionista debe recibir por el tiempo que hace uso de este dinero. Al igual que el precio de cualquier producto, cuando existe mayor liquidez la tasa baja y viceversa. Existen 3 tipos de tasas: Tasa de colocación o activa, tasa de captación o pasiva y margen de intermediación.

- **Medios de pago:** En esta categoría podemos encontrar 2 medios de pago los cuales son:

    a. **Tarjetas de crédito:** Es un instrumento que permite al titular de la tarjeta adquirir bienes o servicios en establecimientos afiliados al correspondiente sistema, este es un medio de

financiamiento y permite realizar pagos sin desembolsar dinero en el acto.

b. **Tarjetas de débito:** Se trata de un instrumento de pago en la red de establecimientos afiliados al sistema que cuenten con dispositivos electrónicos. Los montos son directamente debitados de la cuenta del titular de la tarjeta y acreditados en la cuenta del beneficiario, previa autorización y validación de los fondos existentes.

- **Canales:** las instituciones financieras proveen distintos servicios mediante una variedad de medios llamados canales, entre los cuales tenemos:

  a. **Sucursales bancarias:** Son oficinas que las instituciones financieras crean con el fin de abarcar todas las zonas geográficas estratégicas posibles en las que se encuentren sus clientes para satisfacer sus necesidades financieras.

  b. **Cajeros automáticos:** Se trata de un dispositivo electrónico que opera en línea con acceso en tiempo real a información que permite a los usuarios autorizados retirar dinero en efecto. Adicional se ofrecen otros servicios como: consultas de saldo, transferencias de fondo, depósitos, pagos, bloqueos de tarjeta, y consultas generales.

  c. **Banca electrónica:** Se trata de portales webs o móviles que las instituciones financieras ponen a la disposición de sus clientes para realizar distinto tipo de consultas u hacer uso de

otros servicios como transferencias, pago de servicios, pago de tarjeta, entre otros.

d. **Banca telefónica:** Al igual que la banca electrónica es un servicio que se pone a disposición de los clientes por el cual este puede realizar: transferencias, consultas, pagos, inversiones, solicitudes de productos, atención a empresas, atención a clientes de tarjeta de crédito, soporte de la banca electrónica, afiliación a servicios automáticos y emergencias bancarias.

e. **Corresponsales no bancarios:** Se tratan de vías de acceso no tradicionales que los bancos tienen hacia sus clientes, estos son establecimientos que representan un punto de atención de las instituciones financieras en poblaciones de bajo ingreso o lugares remotos, estos se encuentran en pequeños locales como por ejemplo tiendas barriales.

Con todo lo anterior explicado se debe entender que uno de los principales objetivos de una institución financiera es salvaguardar la seguridad financiera de sus clientes y así mismo otorgar el mejor soporte posible con respecto a los productos y servicios que otorgan hacia sus clientes sin ningún error debido a la delicadez de su naturaleza.

2.8.1 Instituciones bancarias de Guayaquil

Además, como punto importante para la realización de este proyecto de investigación se debe tener en cuenta que, dentro de Guayaquil, existen un total de tres bancos privados con calificación de riesgo mayor que AA a corte

de diciembre del 2021, que mantienen su casa matriz en esta ciudad, los cuales son:

2.8.1.1 Banco de Guayaquil

Según (Banco de Guayaquil, s.f.), indica que la institución financiera inició sus actividades el día 20 de septiembre de 1923 como un banco extranjero italiano, con un capital inicial de 20’000.000 de sucres, denominado como Sociedad Anónima Banco Italiano.

Para el año de 1941 el país de Italia se convierte en eje central dentro de la segunda guerra mundial y una reforma del 14 de agosto de ese año cambia los estatus y la denominación por Banco Nacional del Ecuador, pero a finales de septiembre del mismo año, una nueva escritura pública lo denomina Banco Guayaquil, obteniendo así su actual nombre.

Hasta diciembre del 2021 la institución mantiene una calificación de AAA con una calificación relativa menos, es decir AAA-, según lo indica la (Superintendencia de Bancos, 2021); Esta institución, a finales del año 2021, se ha posicionado como la segunda institución bancaria más rentable del ecuador, siendo uno de los bancos que mayores utilidades ha generado y cuyo representante a la fecha, era el actual presidente de la república, Guillermo Lasso (Espinosa, 2021).

2.8.1.2 Banco Del Pacífico

Dentro de la propia página web del banco del Pacífico, (Banco del Pacífico, s.f.), se indica que este fue fundado el día 10 de abril de 1072 por Marcel J. Laniado de Wind; el banco inició sus actividades con un capital de 40’000.000

de sucres, aportados por un total de 447 accionistas de las ciudades de Guayaquil, Quito, Cuenca, Machala, Manta y Babahoyo.

En el año de 1998, su fundador era considerado como uno de los banqueros con mayor conciencia social del país debido a que logró que la institución mantenga una filosofía de brindar mayor acceso al crédito a todos los sectores de la economía, incluidos artesanos y microempresarios y convertirse en catalizador del desarrollo del país, revolucionando así el sistema financiero del Ecuador.

Sin embargo, en este mismo año, Laniado falleció a sus 71 años, mientras recibía tratamiento médico en el hospital Anderson, de Houston Texas; en el año de 1999 la institución enfrentó el momento más crítico de su historia llegando a estar al borde de la quiebra.

No obstante gracias a la aparición de un nuevo accionista, la fidelidad de los clientes y el manejo prudente y profesional de su administración permitieron su recuperación en tiempo récord y el relanzamiento comercial de la entidad con una imagen renovada y moderna.

Sin embargo, a pesar de todos los esfuerzos que la institución ha realizado para mantenerse como uno de los mejores bancos del ecuador, a partir del inicio del proceso de su venta, dentro del periodo del actual presidente Guillermo Lasso, su posición dentro del ranking de los bancos más rentables del ecuador, pasó de estar en el 3er puesto en el 2020 a bajar hasta el 10mo puesto a finales del 2021 según lo indica (Ramos, 2021).

Aun así, sigue entre los más grandes del país, ya que conserva el segundo puesto en activos, con $ 6.970,3 millones, solo superado por Banco Pichincha.

Además, que, para finales del 2021, esta institución mantiene una calificación de riesgo de AAA- al igual que el banco de Guayaquil según se muestra en la página de la (Superintendencia de Bancos, 2021).

2.8.1.3 Banco Bolivariano

El (Banco Bolivariano C.A., 2017) indica que la entidad bancaria, Banco Bolivariano, fue constituida en la ciudad de guayaquil en el día 19 de abril de 1979, sin embargo, fue hasta el 13 de diciembre de 1980 que inició sus operaciones como entidad financiera. Esta tuvo un capital inicial de 150’000.000 de sucres, siendo el más grande de la república del Ecuador gracias al apoyo de sectores de la industria, las finanzas, el agro y el comercio.

A partir del inicio del nuevo milenio, se conformó el Grupo Financiero Bolivariano, de esta manera, los servicios financieros se extendieron a la administración de fondos y negocios fiduciarios; presencia en la Bolsa de Valores, mediante la administración de portafolios y negocios bursátiles, junto con finanzas corporativas en estructuraciones de deuda en el Mercado de Capitales. (Banco Bolivariano C.A., 2017)

Para el 2021, la institución bancaria, ocupó el 3er dentro del ranking de los bancos más rentables del ecuador con un valor de utilidad de $38′044.977 según (Espinosa, 2021); además este presenta una calificación de riesgo de AAA para diciembre del 2021 (Superintendencia de Bancos, 2021).

2.9 Redes sociales virtuales

Tradicionalmente, una red social se ha definido como un conjunto de personas que tienen vínculos entre sí, sea por temas comerciales, amistad, trabajo, parentesco, etc. El internet permitió que esos conjuntos de personas

se encontraran en un entorno virtual, convirtiéndose en sitios web conformados por comunidades de personas que tienen cosas en común. (rockcontent, 2019).

El sitio SixDegrees.com, creado en 1997, es considerado por muchos como la primera red social moderna, ya que permitía a los usuarios tener un perfil y agregar a otros participantes en un formato parecido a lo que conocemos hoy. (Cruz, 2020).

El número de usuarios de estas hoy en día es equivalente al 58% de la población total del mundo, ocupando la mayor parte del tiempo de los medios conectados en 2021 y se reporta que los usuarios pasan más tiempo en los canales sociales cada día más que el año anterior, 2 horas y 27 minutos según lo indica (Hall, 2022).

2.10 Twitter

Twitter es un sitio de noticias y redes sociales en línea donde la gente se comunica en mensajes cortos llamados tweets, se conoce como tuitear a la acción de publicar mensajes cortos para cualquiera que te siga en Twitter, con la esperanza de que tus palabras sean útiles e interesantes para alguien de tu audiencia. (Gil, 2021)

Según un estudio realizado por (Kemp, 2022) indica Twitter tenía 1,45 millones de usuarios en Ecuador a principios de 2022 lo cual equivale al 8,1% de la población total en ese momento. Sin embargo, Twitter restringe el uso de su plataforma a personas mayores de 13 años, por lo que puede ser útil saber que el 10.5% de la audiencia "elegible" en Ecuador utiliza Twitter en 2022.

Así queda evidenciado que Twitter es una de las redes sociales en línea más grandes del mundo en la que se expresan distintos tipos de opinión pública, de esta es posible la obtención de información con la cual se pueden realizar varios tipos de estudios basados en el procesamiento del lenguaje natural.

CAPÍTULO III
METODOLOGÍA DE LA INVESTIGACIÓN

En el presente capitulo se procederá a detallar las metodologías utilizadas para la realización de la investigación, además de las técnicas e instrumentos de recolección, procesamiento y análisis de los datos.

3.1 Enfoque metodológico

En el actual trabajo de titulación la metodología de investigación aplicada es la observación directa aplicando la investigación exploratoria con un enfoque cuantitativo; se dice que es de observación directa debido a que para la realización de recolección de datos se realiza de manera en que el investigador se encuentra en el lugar de los hechos y registra lo que sucede en el momento. Para el contexto del desarrollo de la investigación esto significa que los datos se obtienen directamente desde una fuente específica, en este caso, Twitter.

Además, el proyecto se utiliza la investigación exploratoria en el ámbito de buscar las fuentes de información que permitan conocer la situación actual del tema en estudio y definir su factibilidad de ejecución, elaborar el diseño de investigación más apropiado y seleccionar o elaborar las técnicas necesarias para la obtención de los datos. (Rolando Alfredo Hernández León, 2014)

Por último, el enfoque que se le da al presente trabajo de titulación es de índole cuantitativo porque según (López) la metodología cuantitativa utiliza la recolección y el análisis de datos para contestar preguntas de investigación y probar hipótesis establecidas previamente, y confía en la medición numérica, el conteo y frecuentemente el uso de estadística para establecer con exactitud

patrones de comportamiento en una población. Es decir que, dentro de la investigación a desarrollar, se evaluarán los distintos algoritmos de ML que existen para la realización del análisis de sentimientos y se tomará el que presente un mejor rendimiento.

3.2 Modelo de minería de datos

Dentro del desarrollo del presente proyecto de investigación se presente el modelo de minería de datos que consiste en 7 módulos, todos basados en el método KDD presentado en el marco teórico, los cuales son necesarios de desarrollar de manera continua para la obtención de las predicciones de los algoritmos de clasificación. El desarrollo de estos módulos será presentado de manera detallada dentro del capítulo 4, el cual trata sobre la propuesta tecnológica, mientras que en el presente capítulo serán mencionados de una manera más superficial a continuación:

3.2.1 Construcción del conjunto de datos

En este primero módulo entra en desarrollo la extracción de datos, el cual se hará mediante el uso del lenguaje de programación Python y una librería que permita realizar una conexión con la fuente de datos, en este caso la red social Twitter, para posteriormente empezar a descargarlos haciendo uso de reglas, o también pudiéndose llamar parámetros, que ayudarán a limitar la cantidad de datos descargados según un rango de tiempo y menciones que se hagan dentro de los datos de las entidades bancarias que serán utilizadas para el presente estudio.

Lo antes mencionado dará como resultado el conjunto de datos en un formato de archivo de texto el cual puede ser procesado como tabla, con todas

las variables seleccionadas que serán necesarias para el desarrollo de los siguientes módulos en el proceso del desarrollo del modelo de minería de datos.

3.2.2 Preprocesamiento y limpieza de datos

Una vez construido nuestro conjunto de datos, es necesario eliminar toda clase de datos que pueda generar ruido dentro del modelo que se desea desarrollar, por ende, se deberá hacer una eliminación de todos los datos que se encuentren repetidos, en blanco, o que no tengan una correlación con lo que se requiere analizar, es decir que no se trate de una opinión hacia el servicio otorgado por medio de la banca ecuatoriana. Así también se buscará eliminar menciones, expresiones regulares como signos de puntuación, y palabras tales como conectores de texto.

Este módulo se basa en el proceso del método KDD llamado preprocesamiento de los datos, y en el cual participarán distintas herramientas para su realización, con estas se buscará eliminar todos los puntos antes mencionados por medio de código dentro de Python para la limpieza del texto y eliminación de datos repetidos, y posteriormente el uso de Excel para la búsqueda de palabras o frases que se identifiquen como datos no correlacionados con el estudio a desarrollar y su posterior eliminación.

De esta forma se busca disminuir el conjunto de datos construido con la finalidad de que el modelo de minería de datos a desarrollar sea preciso a la hora de valorar publicaciones, comentarios u opiniones acerca de la banca ecuatoriana.

3.2.3 Etiquetado de Datos

Este módulo también forma parte del preprocesamiento de datos en el método KDD, en el cual se busca etiquetar los datos obtenidos con su resultante, es decir que se va a asignar el valor de positivo, negativo, o neutro a todos los campos del conjunto de datos obtenidos en base a lo que diga el texto descargado.

Esto con la finalidad de tener un resultado preliminar de las opiniones acerca de los servicios y productos de la banca ecuatoriana, para poder desarrollar el entrenamiento de los algoritmos para el modelo de minería de datos. Esto se realiza mediante 2 iteraciones, la primera es de manera autónoma junto a una librería de redes neuronales llamado *transformer* la cual consta con un procesador de lenguaje natural desarrollado por Google llamado sentiment-analysis el cual otorga el valor de negativo o positivo según el texto ingresado a esta red neuronal. Junto a este resultado preliminar se realiza la búsqueda de palabras, frases o expresiones dentro del texto descargado realizando una revisión manual y exploratoria de los datos obtenidos y por consiguiente el etiquetado o corrección de etiquetas otorgadas por la red neuronal antes mencionada.

3.2.4 Procesamiento del lenguaje natural

Como paso final del preprocesamiento de datos desarrollado en el presente proyecto de investigación para la construcción de un modelo de minería de datos que realice un análisis de sentimientos es necesario realizar un procesamiento del lenguaje natural en el que todas las palabras de los textos puestos a analizar por el modelo sean transformadas a una equivalencia

numérica para que estas puedan ser entendidas por la computadora. Esto se realiza mediante algoritmos desarrollados para Python en el que a todas las palabras se les asigna un peso según su frecuencia dentro del conjunto de datos. Gracias a este último paso, será posible entrenar los algoritmos de clasificación seleccionados para el desarrollo del modelo de minería de datos.

3.2.5 Separación de los datos

Para iniciar con el proceso de la aplicación de algoritmos del método KDD, es en primer lugar necesario dividir nuestro conjunto de datos en dos, uno que será utilizado para el entrenamiento del algoritmo y el otro que es utilizado para ser puesto a prueba. Esto se realiza mediante una selección aleatoria de datos en el que el conjunto de entrenamiento debe ser mucho mayor que el de pruebas para entrenar de manera óptima el algoritmo. Además, para la comparación del rendimiento de los algoritmos se realizarán distintas iteraciones de entrenamiento y prueba en los cuales al conjunto de datos se los separará en tres distintos conjuntos de datos.

3.2.6 Entrenamiento y clasificación

Como segundo paso en el proceso de la aplicación de algoritmos del método KDD, se deberá elegir el o los algoritmos que se desean poner a prueba para el desarrollo del modelo de minería de datos. Estos algoritmos serán entrenados para el reconocimiento de patrones y características dentro del conjunto de datos obtenido previamente para que, junto al sentimiento otorgado en el módulo de etiquetado de datos, pueda saber cuál etiqueta otorgarle y puestos a prueba con el conjunto de datos antes separado y como resultado se obtendrá tanto la predicción automática que estos hagan, así

como su rendimiento. Para el presente estudio se hará la selección de hasta 5 distintos algoritmos de clasificación que se alimentarán con el conjunto de datos ya preparado para la realización de esta tarea.

3.2.7 Comparación de rendimiento

Como último módulo en el modelo de minería de datos desarrollado, y a su vez parte del último proceso en el método KDD llamado interpretación de resultados, se obtendrán los valores que demuestren la precisión, exhaustividad y la unión de estos dos últimos llamado F1-Score, de los resultados otorgados por los algoritmos puestos a prueba en las 3 iteraciones, estos resultados serán comparados entre sí con la finalidad de decidir cuál de todos los algoritmos es el mejor al momento de otorgar predicciones de las etiquetas otorgadas a las distintas opiniones que se tienen en el conjunto de datos utilizado.

3.3 Población y muestra

La población total del presente trabajo de investigación está constituida por todos los tweets que sean obtenidos mediante el proceso de extracción de datos en el cual serán todos los tweets que se hayan escrito dentro del periodo del 1 de enero del 2022 hasta el 27 de Julio del 2022 en los que se mencione al menos a una de las siguientes instituciones financieras, Banco de Guayaquil, Banco del Pacífico, o Banco Bolivariano.

CAPÍTULO IV
PROPUESTA TECNOLOGICA

4.1 Herramientas de desarrollo

Como herramientas utilizadas para el desarrollo de la solución informática para la creación de un modelo de minería de datos que realice un análisis de sentimientos sobre las opiniones obtenidas mediante Twitter con respecto a la banca privada ecuatoriana, se realizó una comparación de las siguientes herramientas:

4.1.1 *Benchmark* de herramientas de desarrollo

Para poder realizar un modelo de minería de datos, es necesario un lenguaje de programación que tenga las facilidades para la realización de esta tarea, entre todos los lenguajes de programación existentes, entre los mejores para la ciencia de datos están Python, Java y R los cuales son comparados a continuación:

Tabla 1

Comparación de lenguajes de programación para el desarrollo de minería de datos

Lenguaje	Paradigma	Licencia	Ventajas	Desventajas	Operatividad
java	Orientado a objetos	Gratuita	-Excelente para la escritura de códigos de producción de ETL y algoritmos de *Machine Learning* muy intensivo computacionalmente	-En comparación con los lenguajes específicos de dominio como R, no dispone de muchas librerías disponibles para métodos estadísticos avanzados.	Multiplataformas
Python	Multiparadigma	Gratuita	-Python es un lenguaje fácil de aprender. -Paquetes como pandas, scikit-learn y Tensorflow hacen de Python una opción sólida para aplicaciones avanzadas de aprendizaje automático. -Python es escalable al funcionar más rápido que R, esto permite crecer y escalar junto con los proyectos.	-Python no dispone de buena documentación, -Python es un lenguaje de tipo dinámico, lo que significa que debemos ser muy cuidadosos.	Multiplataformas
R	Funcional y Orientado a Objetos	Gratuita	-Excelente gama de paquetes de código abierto y de alta calidad. R tiene un paquete para casi todas las aplicaciones cuantitativas y estadísticas	-R no es bueno para programaciones de propósito general. -R tiene algunas características poco frecuentes que pueden atrapar a los programadores con experiencia en otros idiomas.	Multiplataformas

Nota: La tabla muestra características importantes de los diferentes lenguajes de programación existentes para la creación de modelos de minería de datos. Cuadro comparativo desarrollado por el autor.

Después de realizar esta comparación, el lenguaje seleccionado para la realización del modelo de minería de datos es Python debido a sus múltiples librerías para la ciencia de datos y su gran velocidad y versatilidad, este es utilizado como instrumento principal de programación. En el caso de la solución planteada, se lo utiliza para la recolección, limpieza y etiquetado autónomo de un gran volumen de datos, recolectados en Twitter, para posteriormente ser procesados con la ayuda de los siguientes algoritmos de *machine learning* que vienen incluidos en su paquete de "Scikit-learning" o también llamado "SKlearn":

- *Naive Bayes*
- Árbol de decisión
- Vecinos cercanos
- Random Forest
- Máquina de soporte vectorial

4.1.2 *Benchmark* de herramientas de visualización

Para poder mostrar los resultados de las predicciones realizadas por el modelo desarrollado, es necesario una herramienta de visualización de datos con la que se pueda expresar por medio de gráficas los resultados obtenidos. Así como fue visto en el capítulo 2, dentro del apartado de herramientas de visualización de datos, en el cuadrante mágico de Gartner, las mejores herramientas para esta tarea son PowerBi, Tableau, y QlickView los cuales son comparados a continuación:

Tabla 2

Comparación de herramientas de visualización de datos parte 1.

Herramientas de Visualización de datos	Tipos de Licencias	Ventajas	Desventajas	Versiones disponibles
Power BI	**Licencia Gratuita** ✓ **Licencia de paga** PowerBi premium ($13.70 mensual) PowerBi pro ($27.50 mensual por usuario)	-Integración con todo el ecosistema de Microsoft y de Azure (Excel, Teams, etc.)	-Las versiones empresariales tienen un precio elevado	-Power BI (Power BI Pro, Power BI Premium)
		-Capacidades de inteligencia artificial	-Difícil de migrar fuera del ecosistema Microsoft	-Power BI Mobile
		-No es necesaria experiencia técnica, es muy fácil de aprender		-Power BI Report Server
		-Automatización de reportes y limpieza de datos		

Nota. Esta tabla muestra características importantes de las diferentes herramientas de visualización de datos para la demostración de los resultados de las predicciones realizadas por el modelo desarrollado. Cuadro comparativo desarrollado por el autor.

Tabla 3

Comparación de herramientas de visualización de datos parte 2.

Herramientas de Visualización de datos	Tipos de Licencias	Ventajas	Desventajas	Versiones disponibles
Tableau	**Licencia Gratuita** **X** **Licencia de paga** Tableau creator ($70 mensual) Tableau explorer ($42 mensual) Tableau Viewer ($15 mensual)	-Integración con servicios cloud -Fácil de usar (Drag and Drop) -Extensible con addons para gobierno	-Enfocado solamente en tareas de reporting, sin capacidad de ETL y transformación de datos	-Tableau Cloud -Tableau Desktop -Tableau Prep

Nota. Esta tabla muestra características importantes de las diferentes herramientas de visualización de datos para la demostración de los resultados de las predicciones realizadas por el modelo desarrollado. Cuadro comparativo desarrollado por el autor.

Tabla 4

Comparación de herramientas de visualización de datos parte 3.

Herramientas de Visualización de datos	Tipos de Licencias	Ventajas	Desventajas	Versiones disponibles
QlinkView	**Licencia Gratuita** **X** **Licencia de paga** Documental CALs ($360 por documento) User CALs ($1350 por usuario)	-Dashboards interactivos -Despliegue en cloud (QlinkSense) -Exportación de datos y dashboards a formatos de imagen, excel o PDF	-Curva de aprendizaje elevada para nuevos usuarios -Soporte empresarial	-Qlink Sense Enterprise -Qlink Sense Analytics Platform -Qlink Sense Core -Qlink Sense Cloud

Nota. Esta tabla muestra características importantes de las diferentes herramientas de visualización de datos para la demostración de los resultados de las predicciones realizadas por el modelo desarrollado. Cuadro comparativo desarrollado por el autor.

Para esta tarea, debido a su licencia gratuita y su gran facilidad de uso, la herramienta seleccionada fue Power BI, la cual se utilizó para mostrar los resultados de las predicciones realizadas por el modelo elegido después del análisis de rendimiento de los algoritmos seleccionados.

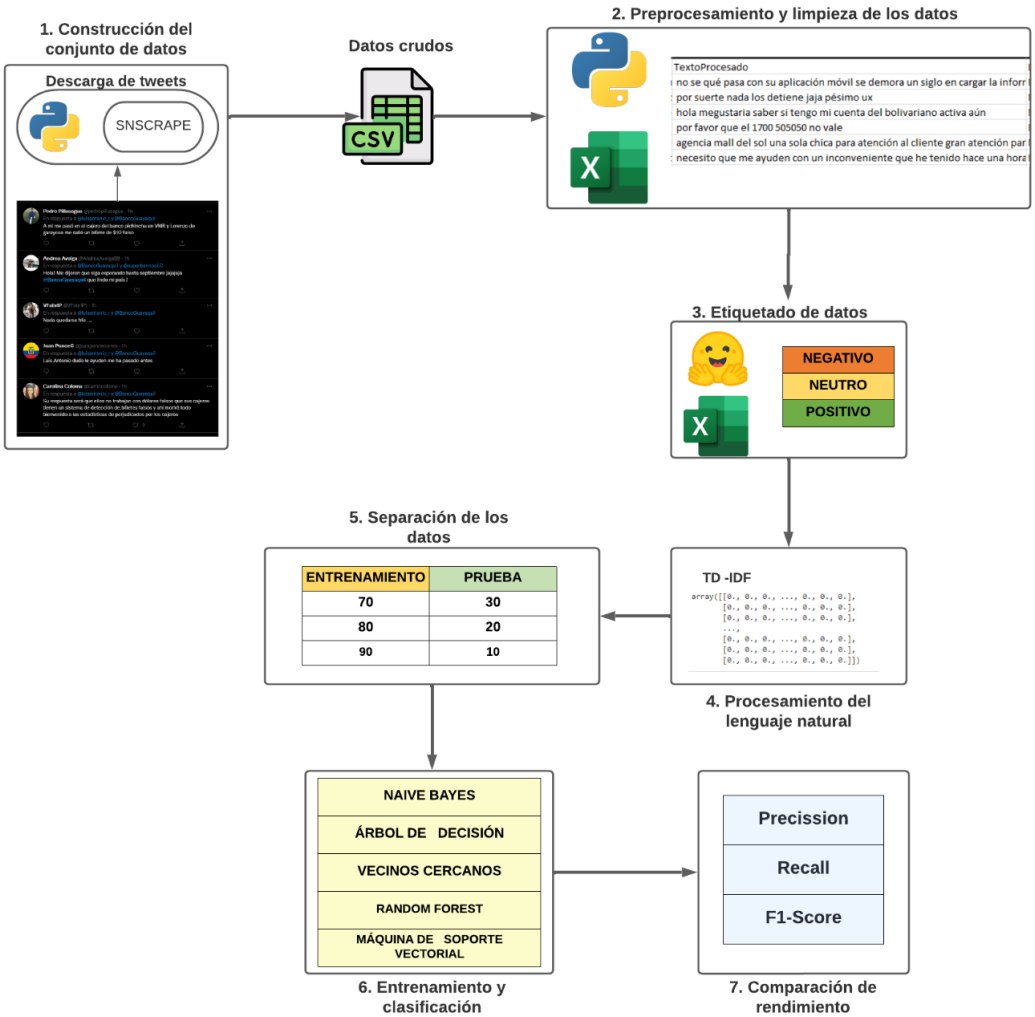
Por último, es de recalcar que la herramienta Excel fue utilizado para realizar una segunda limpieza de los datos recolectados a través de filtros, para posteriormente elaborar la etiqueta de estos (negativo, neutro, positivo).

4.2 Desarrollo del modelo de minería de datos

Para poder llevar a cabo la realización del modelo de minería de datos que realice un análisis de sentimientos se implementan una serie de módulos en la que cada una depende de la realización del anterior, para esto se aplica el modelo de minería de datos mencionado en el capítulo 3 del presente proyecto de investigación que se esquematiza de la siguiente forma:

Figura 2

Diagrama de los diferentes módulos a realizar para la creación del modelo de minería de datos



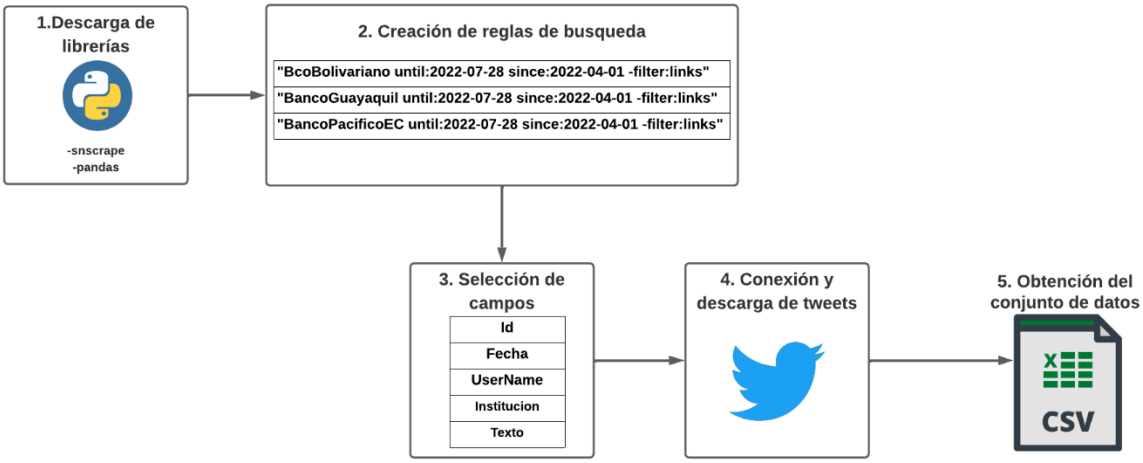
Nota. Diagrama desarrollado por el autor.

4.2.1 Construcción del conjunto de datos

La construcción del conjunto de datos utilizados para el desarrollo del modelo de minería de datos para realizar análisis de sentimientos fue hecha mediante el lenguaje de programación Python en conjunto a la librería snscrape con su módulo especializado para Twitter. Esta librería fue seleccionada debido a que permite realizar una extracción de tweets directamente desde la página de Twitter sin restricciones de claves de autorización, ni límites de cantidad máxima de tweets descargados o del periodo de tiempo de descarga de estos, a diferencia de la propia API de Twitter que si opone estas limitantes. Además, se hace uso de la librería Pandas, la cual es especializada para el manejo de datos dentro de Python y es utilizado en el desarrollo del modelo de minería de datos del presente proyecto de investigación para la creación del conjunto de datos. Es así como podemos observar a este módulo que consta de una serie de procesos los cuales son posibles visualizar de la siguiente manera:

Figura 3

Diagrama del proceso de construcción del conjunto de datos



Nota. Diagrama desarrollado por el autor.

De esta forma se realizó la extracción de datos realizando 3 consultas con la librería snscrape, en las cuales se solicitan todos los tweets en los que se mencionen a las cuentas de los bancos Banco Bolivariano, Banco de Guayaquil y Banco del Pacífico, en un periodo correspondiente al primero de enero del 2022 hasta el primero de Julio del 2022 y gracias a esto se logró la obtención de un total de 26.955 Tweets para construir el conjunto de datos en el cual podemos encontrar los siguientes campos:

- **Id:** Es el código que se le asignó al registro, este es incremental, lo que quiere decir que inicia en 0 y va sumando más uno por cada registro encontrado hasta llegar al 26954.
- **Fecha:** Muestra la fecha y hora en la que se escribió el tweet descargado.

- **UserName:** Muestra el nombre de usuario que escribió el tweet descargado.
- **Institución:** Muestra la institución que es nombrada en el tweet descargado.
- **Texto:** Muestra el texto escrito en el tweet descargado sin modificar.

Para poder realizar esta tarea dentro del lenguaje de programación python se implementó el siguiente bloque de código:

Figura 4

Bloque de código usado para la descarga del conjunto de datos

```
import snscrape.modules.twitter as sntwitter
import pandas as pd


query = ["BcoBolivariano until:2022-07-01 since:2022-01-01 -filter:links",
         "BancoGuayaquil until:2022-07-01 since:2022-01-01 -filter:links",
        "BancoPacificoEC until:2022-07-01 since:2022-01-01 -filter:links"]

column = ['Id','Fecha','UserName','Institucion','Texto']

data = []

id = 0

for banco in query:
    print(banco)
    for tweet in sntwitter.TwitterSearchScraper(banco).get_items():

        if(tweet.user.username != 'BcoBolivariano' and tweet.user.username != 'superbancosEC' and tweet.user.username != 'BancoGuayaquil' and tweet.user.username != 'Banc

            data.append([id, tweet.date, tweet.user.username, banco.split()[0], tweet.content])

            id += 1

            TweetsDf = pd.DataFrame(data, columns=column)

TweetsDf.to_csv('OpinionesBancarias_df.csv', sep = ";", encoding= 'UTF-8')
```

Nota. Captura obtenida del código fuente desarrollado por el autor.

Como se puede apreciar, el primer paso para la obtención del conjunto de datos es la carga de las librerías pandas y la librería snscrape con su módulo de twitter. Como paso se siguiente se procede a crear una variable que contiene el conjunto de las reglas de búsqueda que serán utilizadas para la

descarga de los tweets, en este caso se buscan a los tweets que mencionen a las entidades bancarias Banco Bolivariano, Banco de Guayaquil, y Banco del Pacífico en un periodo del 1 de enero del 2022 al 27 de Julio del 2022 por medio de un *query* de búsqueda en que primero se coloca en nombre de la cuenta a buscar, seguido de la última y primera fecha de búsqueda y por último se agregó como filtro extra no traer los tweets que contengan enlaces con la cláusula “Filter: -links”.

Después del paso anterior, se definen las columnas pertenecientes al conjunto de datos que se desea obtener por medio de un conjunto de palabras llamado “Column”. Además, se crea la variable “data” de tipo arreglo en que se almacenaran temporalmente todos los tweets descargados. Por último, antes de iniciar la conexión con *twitter* y la descarga de los Tweets, se crea la variable “id” de tipo numérico la cual va a ser usada para ser el identificador de cada tweet y se va a registrar de manera incremental.

Para empezar la descarga de los tweets se crean un ciclo repetitivo con el comando FOR el cual va a leer cada uno de los querys registrados en la variable con este mismo nombre para que como paso siguiente sea crear otro ciclo FOR dentro del anterior el cual leerá cada uno de los tweets encontrados con el query buscado y se crea una condición con el comando IF para quitar todos los tweets que sean escritos por las mismas entidades bancarias que son buscadas. Haciendo uso del comando TwitterSearchScrape y su función getItem se obtienen todos los datos obtenidos de la búsqueda de los cuales se seleccionan y registran en la variable “data” los siguientes:

- **Date:** Contiene la fecha en la que se escribió el tweet.

- **UserName:** Contiene el nombre de usuario de quien escribió el tweet.

- **Content:** Contiene el texto escrito en el tweet.

También son registrados la variable “id” que se usa como identificador de cada registro guardado, y se guarda el nombre de la entidad bancaria que es mencionada en el *tweet* el cual se obtiene seleccionando la variable que contiene el query buscado y usando la función Split, la cual divide cada palabra de un texto en un conjunto de palabras, para sacar únicamente la primera palabra del texto. Luego de esto la variable “Id” es incrementada en uno para registrar el siguiente *tweet* que se va a guardar, y se almacena la información obtenida en un data *frame* llamado TweetsDf con la función de pandas DataFrame en el que se guardan los datos y se asigna a cada columna respectivamente de la variable “column”.

Por último, al terminar el ciclo FOR principal se obtiene el conjunto de datos deseado que se presenta en la figura 5, y se guarda en un archivo csv haciendo uso del comando to_csv de la librería pandas.

Figura 5

Este gráfico muestra el resultado obtenido de la ejecución del segmento de código.

```
>>> Tweets_FinalDf.head()
   Id                     Fecha         UserName     Institucion                                              Texto                                      TextoProcesado
0   0 2022-07-26 20:11:52+00:00          Jsantosdm  BcoBolivariano  Una vez más tarjeta clonada y débitos realizad...  una vez más tarjeta clonada débitos realizados...
1   1 2022-07-26 19:12:45+00:00  Stephanoherrer2  BcoBolivariano  @BcoBolivariano Quiero diferir el consumo de m...   quiero diferir el consumo de mis tarjetas 24 ...
2   2 2022-07-26 16:53:47+00:00  Stephanoherrer2  BcoBolivariano  @BcoBolivariano no se para que tanta gente en ...   no se para que tanta gente en el banco si uno...
3   3 2022-07-26 16:25:55+00:00     EmmanuelVliz  BcoBolivariano  @BcoBolivariano @superbancosEC Buenos dias. Al...   buenos dias alguna info del requerimiento que...
4   4 2022-07-26 03:01:37+00:00  stalinalfonzo01  BcoBolivariano  @BcoBolivariano Buenas noches en el banco me p...   buenas noches en el banco me promocionaron as...
```

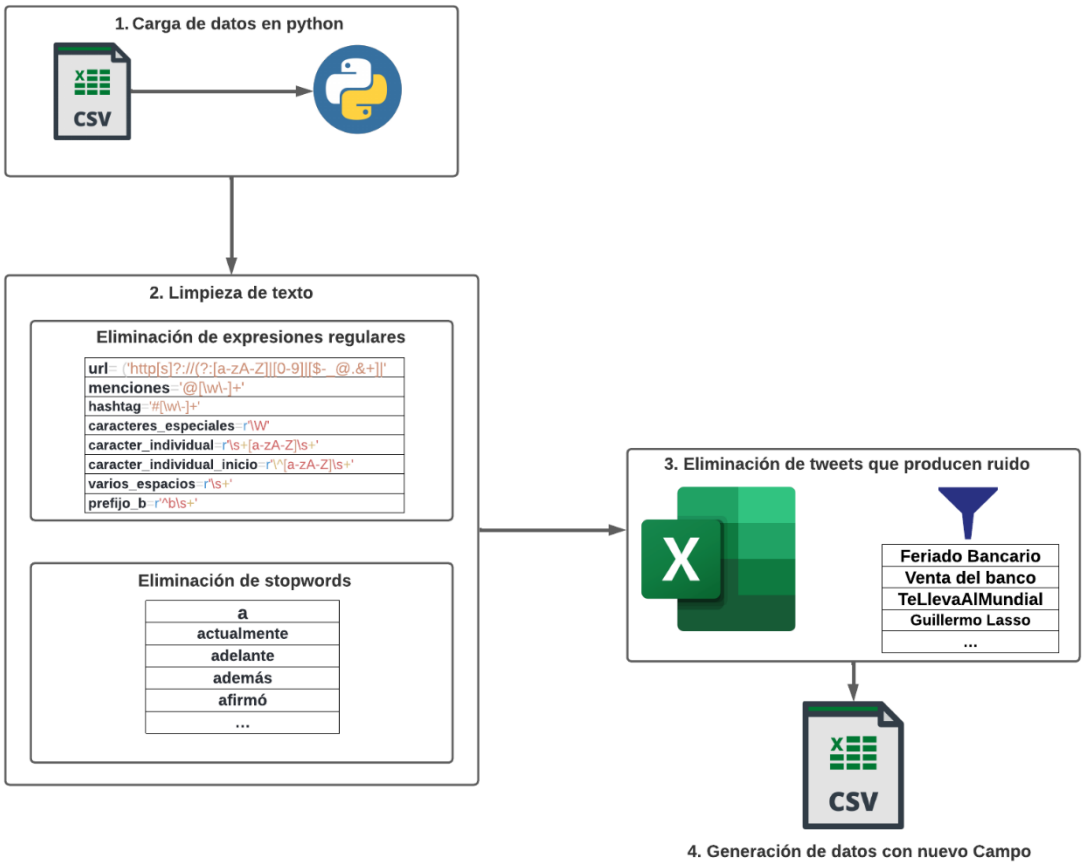
Nota. Gráfico obtenido directamente de la ejecución del segmento de código anterior desarrollado por el autor.

4.2.2 Preprocesamiento y limpieza de datos

El módulo de preprocesamiento y limpieza de datos es graficado de la siguiente manera:

Figura 6

Diagrama del proceso de preprocesamiento y limpieza de datos



Nota. Diagrama desarrollado por el autor.

En este se cargó el archivo generado en la etapa anterior dentro de Python y se procedió a la creación de un nuevo campo llamado “TextoProcesado” en el cual se almacena el texto escrito por el usuario después de que se le realizara un proceso de limpieza en el cual se eliminaron todos los enlaces

que direccionaban a otras páginas web, las menciones a otras cuentas de Twitter (ej: @RoyMieles), los hashtags (ej: #VamosAlMundial), caracteres especiales tales como signos de puntuación (ej: “.”, “;”, “,”, “:”), así como signos de exclamación, interrogación, tildes y emojis, también se eliminaron los espacios innecesarios encontrados dentro del texto, y por último los conectores, también llamados “stopwords” los cuales contemplan palabras tales como “de”, “la”, “y”, “por”, entre otras. A continuación se muestra el segmento de código que fue utilizado para el desarrollo de este proceso:

Figura 7

Primer parte del segmento de código utilizado para la limpieza de datos.

```
#preprocesamiento  de los tweets
import re

#re.sub("cadena a buscar", "con la que se reemplaza", cadena_leida)
url = ('http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|'
       '[!*\(\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+')
menciones = '@[\w\-]+'
hashtag = '#[\w\-]+'
caracteres_especiales = r'\W'
caracter_individual=r'\s+[a-zA-Z]\s+'
caracter_individual_inicio= r'\^[a-zA-Z]\s+'
varios_espacios= r'\s+'
prefijo_b = r'^b\s+'
numeros = '[0-9]+'

processed_tweets = []
n = 0
for tweet in TweetsDf['Texto']:

    tweet_procesado = tweet.lower()  #Convertir a minúsculas
    tweet_procesado = re.sub(menciones, ' ', tweet_procesado)
    #tweet_procesado = re.sub(hashtag, ' ', tweet_procesado)
    tweet_procesado = re.sub(url, ' ', tweet_procesado)
    tweet_procesado = re.sub(caracteres_especiales, ' ', tweet_procesado)
    tweet_procesado = re.sub(caracter_individual, ' ', tweet_procesado)
    tweet_procesado = re.sub(caracter_individual_inicio, ' ', tweet_procesado)
    tweet_procesado = re.sub(prefijo_b, '', tweet_procesado)
    #tweet_procesado = re.sub(numeros, ' ', tweet_procesado)
    tweet_procesado = re.sub(" cta ", ' cuenta ', tweet_procesado)
    tweet_procesado = re.sub(" dlrs ", ' dolares ', tweet_procesado)
    tweet_procesado = re.sub(" q ", ' que ', tweet_procesado)
    tweet_procesado = re.sub(" sr ", ' señor ', tweet_procesado)
    tweet_procesado = re.sub(" srs ", ' señores ', tweet_procesado)
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Figura 8

Segundo segmento de código utilizado para la limpieza de datos

```
for tweet in TweetsDf['Texto']:

    tweet_procesado = tweet.lower()  #Convertir a minúsculas
    tweet_procesado = re.sub(menciones, ' ', tweet_procesado)
    #tweet_procesado = re.sub(hashtag, ' ', tweet_procesado)
    tweet_procesado = re.sub(url, ' ', tweet_procesado)
    tweet_procesado = re.sub(caracteres_especiales, ' ', tweet_procesado)
    tweet_procesado = re.sub(caracter_individual, ' ', tweet_procesado)
    tweet_procesado = re.sub(caracter_individual_inicio, ' ', tweet_procesado)
    tweet_procesado = re.sub(prefijo_b, '', tweet_procesado)
    #tweet_procesado = re.sub(numeros, ' ', tweet_procesado)
    tweet_procesado = re.sub(" cta ", ' cuenta ', tweet_procesado)
    tweet_procesado = re.sub(" dlrs ", ' dolares ', tweet_procesado)
    tweet_procesado = re.sub(" q ", ' que ', tweet_procesado)
    tweet_procesado = re.sub(" sr ", ' señor ', tweet_procesado)
    tweet_procesado = re.sub(" srs ", ' señores ', tweet_procesado)
    tweet_procesado = re.sub(" x ", ' por ', tweet_procesado)
    tweet_procesado = re.sub(" d ", ' de ', tweet_procesado)
    tweet_procesado = re.sub(" xq ", ' por que ', tweet_procesado)
    tweet_procesado = re.sub(varios_espacios, ' ', tweet_procesado, flags=re.I)

    processed_tweets.append([n, tweet_procesado])   #agregar a la lista de tweets procesados
    n += 1


TweetsProcesados = pd.DataFrame(processed_tweets, columns=['Id', 'TextoProcesado'])

Tweets_FinalDf = pd.merge(TweetsDf, TweetsProcesados, on = 'Id')

Tweets_FinalDf.to_csv('OpinionesBancarias_df.csv', sep = ";", encoding= 'UTF-8')
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

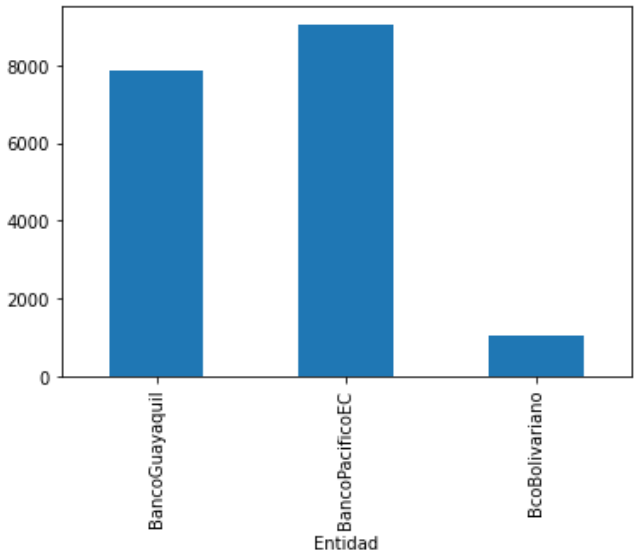
Como se puede observar se importa la librería re la cual es una librería especializada en expresiones regulares, las cuales fueron mencionadas anteriormente, luego se procede a crear una variable por cada tipo de texto que se desea eliminar y luego se crea un ciclo repetitivo con el comando FOR para procesar cada registro de texto, iniciando por convertirlo a minúsculas a todo el texto con el comando lower y por consiguiente se usa el comando sub para sustraer todos los tipos de texto antes creados y este nuevo texto generado se guarda en un arreglado llamado “processed_tweets” el cual se

va a unir con el data frame principal haciendo uso del comando merge() uniéndolos en base a su identificador.

Como paso siguiente, se guardó un nuevo archivo en formato de archivo de texto separado por comas CSV, con el comando to_csv de la librería pandas, en el que se contenía todos los datos antes mencionados con el nuevo campo generado, para ser tratado como archivo de Excel y aplicar un filtro en el que se buscó y eliminó todo tweet que sea perteneciente a campañas publicitarias, campañas políticas, y además, cualquier tweet que trate temas que no sean correspondientes comentarios, o publicaciones que hablen directamente sobre la banca ecuatoriana que se está analizando como por ejemplo comentarios acerca del presidente Guillermo Lasso, sobre denuncias a la fiscalía por robos a clientes, sobre campañas de *footbal*l, entre otros. De este modo el set de datos inicial de 26.955 fue reducido a un total de 17.940 Tweets separados de la siguiente forma por institución bancaria:

Figura 9

Gráfico de barras de la cantidad de tweets existentes en el conjunto de datos separados por institución bancaria



Nota. Gráfico desarrollado por el autor.

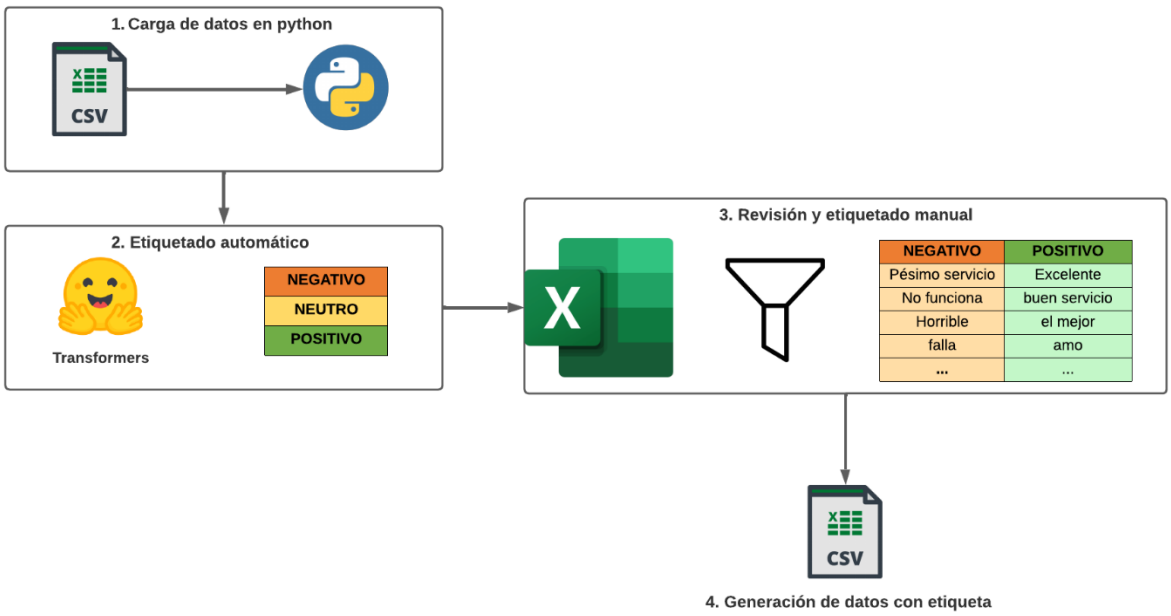
Así como se puede ver en el gráfico anterior, el Banco del Pacífico representa la mayor cantidad de tweets con un total de 9.055, seguido del Banco de Guayaquil con 7.847 tweets, y por último está en Banco Bolivariano con un total de 1.038 tweets.

4.2.3 Etiquetado de Datos

El módulo de etiquetado de datos es diagramado de la siguiente forma:

Figura 10

Diagrama del proceso de etiquetado de datos



Nota. Diagrama desarrollado por el autor.

Dentro del cual se realizaron dos instancias, en el que a cada tweet descargado se le asignará una categoría entre positivo, neutro o negativo, dependiendo de su contexto. En la primera instancia se cargaron todos los datos a Twitter y se utilizó la librería Transformers, la cual consiste en un conjunto de redes neuronales pre-entrenadas para la realización de diferentes tipos de tareas, y con su módulo de pipeline se hace llamamiento a un modelo pre-entrenado llamado "sentiment-analysis" para realizar un etiquetado automático de los tweets obtenidos, con esto se crea el campo “Sentimiento” dentro del set de datos del modelo en el cual se encuentra descrito a que

clasificación pertenece. Por consiguiente, de crea un segmento de código mostrado a continuación:

Figura 11

Segmento de código utilizado para el etiquetado automático de los datos

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification, pipeline
sentiment_pipeline = pipeline("sentiment-analysis")

Tweets_FinalDf = pd.read_csv('OpinionesBancarias_df.csv', sep = ";")

Tweets_Analisis = Tweets_FinalDf['TextoProcesado']

columna_label = []

n = 0

for tweets in Tweets_Analisis:
    sentimiento = sentiment_pipeline(tweets)

    columna_label.append([n, sentimiento[0]['label'], sentimiento[0]['score']])
    n+=1

sentimientos = pd.DataFrame(columna_label, columns=['Id', 'Label', 'score'])

Tweets_FinalDf = pd.merge(Tweets_FinalDf, sentimientos, on = 'Id')

Tweets_FinalDf.to_csv('OpinionesBancarias_ConLabel_df.csv', sep = ";", encoding= 'UTF-8')
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Para el desarrollo de esta tarea en python el primer paso es carga la librería Transformers e importar su módulo pipeline, paso seguido se inicializa una instancia del modelo “sentiment-analysis” dentro de la variable “sentiment_pipeline”, luego de esto se carga el conjunto de datos que contiene todos los tweets que serán procesados por el modelo de análisis de sentimientos y se crea la variable Tweets_Analisis que los contendrá, por último se crea un ciclo repetitivo con el comando FOR en el que por cada tweet será procesado por el modelo de análisis de sentimientos de Transformers, cuyos resultados serán registrados en un *data frame* que contendrá un identificador para poder unirlo con el conjunto de datos principal llamado id, también contendrá dentro del campo “Label” el sentimiento

encontrado por el modelo y por último el score de precisión con el que predijo el sentimiento. Para finalizar estos resultados son añadidos al conjunto de datos principal haciendo uso de el comando merge() nuevamente y uniéndolos por sus “id”.

Gracias a este primer paso, se obtiene una etiqueta con la que, junto a Excel y su herramienta de filtros, se buscaron palabras claves, frases o expresiones realizadas por los mismos usuarios con los que se crearon reglas de búsqueda con las que se procedió a la realización de una revisión manual de los datos ya clasificados, y etiquetándolos en base a estas reglas creadas

Tabla 5

Reglas que se usaron para etiquetar los tweets en base a si son negativos o positivos del conjunto de datos obtenido lista 1

NEGATIVO	POSITIVO
se demora	buena atención
pésimo	buen servicio
mal servicio	gran servicio
malo	excelente
horrible	el mejor
no vale	gran banco
no sirve	fue resuelto
no contesta	resolvieron
no atiende	muchas gracias
porquería	muchísimas gracias
colmo	nada mejor
irrespeto	mejor banco
desorden	ya funcionó
inconveniente	gracias
no me sirve	amo
perder tiempo	van adelante
perder el dinero	innovadores
no responde	innovador
incompetente	felicitaciones
no dan solución	felicidad
inútil	soy fiel
indignado	feliz
indignante	yo si pude
indignada	sí puedo
no solucionan	vamos por más
tiene problemas	mejor
no es funcional	muy agradecido
no es lo más funcional	agradecido
no funciona	pude ingresar
sigo esperando respuesta	agradecimiento
basura	cariño

Nota: Esta tabla muestra todas las palabras, frases o expresiones que identifican si un tweet posee un sentimiento negativo o positivo, tabla desarrollada por el autor.

Tabla 6

Reglas que se usaron para etiquetar los tweets en base a si son negativos o positivos del conjunto de datos obtenido lista 2

NEGATIVO	POSITIVO
no abre	mejor opción
nunca abre	te devuelven
mejoren la calidad	son eficientes
arreglen	es eficiente
el peor	las mejores
ratas	banco con visión
error	fácil
intentando entrar	simplifica
pésimo	super rápido
mismo problema	muy buen
horas para atención	resolvió
no deja entrar	es estable
no puedo ingresar	maravilla
no funciona	sin problemas
falla	súper atención
no contestan	buena atención
nunca contestan	los mejores
no llega	yo me quedo con
nadie contesta	huelen muy bien
causa problema	enhorabuena
peor servicio	bien banco
no es verdad	en buena hora
estoy intentando	agradezco
decepcionante	gracias
sinsentido	muy amable
terrible	facilito
no se puede	orgullosa
pésima app	orgulloso
pésima	positivo
busquen otro banco	felicitar
terrible experiencia	increíble
cayéndose	si te reembolsa
nadie ayuda	sí reembolsa
no la puedo usar	amable

Nota: Esta tabla muestra todas las palabras, frases o expresiones que identifican si un tweet posee un sentimiento negativo o positivo, tabla desarrollada por el autor.

Tabla 7

Reglas que se usaron para etiquetar los tweets en base a si son negativos o positivos del conjunto de datos obtenido lista 3

NEGATIVO	POSITIVO
engaño	felicidades
mentira	vamos
chulquero	grande
ladrón	escucha
robo	lo devuelven
evasor	le devuelven
vergüenza	apoyo
no me dan respuesta	correcto
cobro indebido	encontré ayuda
mala imagen	
retrasado	
ladrones	
incumplimiento	
abuso	
corrupto	
no me deja ingresar	
estafa	
grosero	
grosera	
denuncia	
mala práctica	
juegan con los clientes	

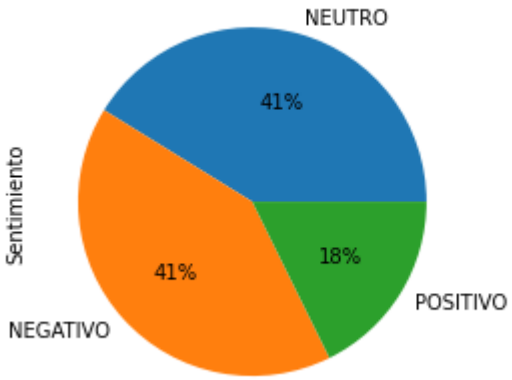
Nota: Esta tabla muestra todas las palabras, frases o expresiones que identifican si un tweet posee un sentimiento negativo o positivo, tabla desarrollada por el autor.

Con la creación de estas reglas se procedió a corregir o re-etiquetar los tweets en negativo o positivo en base a estas, y a todos los demás comentarios que no posean estas reglas se los consideró como neutros debido a que pueden tratarse tanto como de preguntas hacia la institución, sugerencias, solicitudes de ayuda, entre otro tipo de comentario que no

representa tanto una emoción positiva como negativa, quedando de esta forma una distribución de sentimientos de la siguiente forma:

Figura 12

Gráfico de pastel en el que se muestran los porcentajes de clasificación manual de la data cargada para la realización del modelo de machine Learning



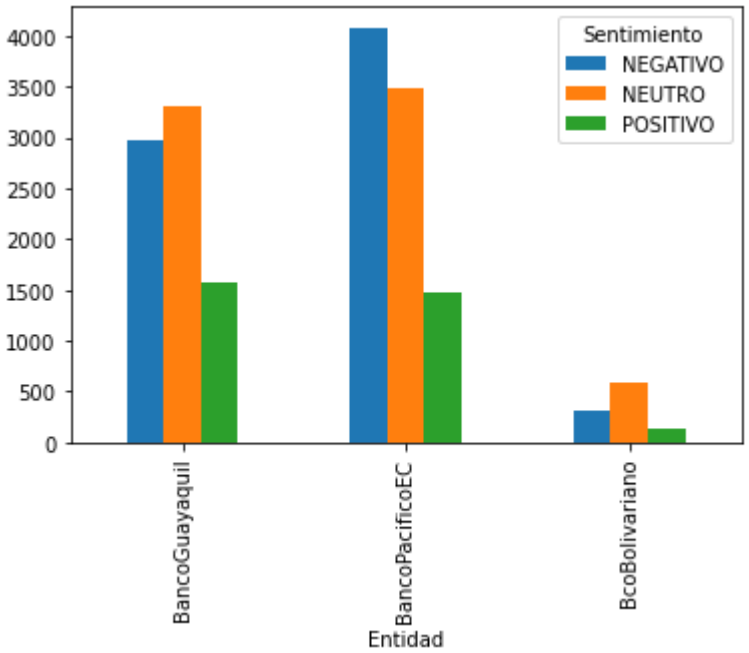
Nota. Gráfico desarrollado por el autor.

Es decir que, al finalizar este proceso, se logró identificar que el 41% de los datos pertenecían a la clasificación de “Negativo”, otro 41% pertenece a la clasificación de “Neutro”, y por último el 18% de los datos pertenecen a la clasificación de “Positivo”.

Así mismo, se realizó un gráfico comparativo de las clasificaciones realizadas distinguidas por la entidad bancaria al a que pertenece y como resultado preliminar se denota que la entidad bancaria con mayor cantidad de comentarios negativos es el Banco del Pacífico, así mismo se demuestra que de esta entidad es de la cual se poseen mayor cantidad de registros, como se muestra a continuación:

Figura 13

Cantidad de comentarios positivos, negativos y neutros separado por institución bancaria



Nota. Gráfico de barras en el que se compara la cantidad de comentarios positivos, neutros o negativos que posee cada entidad bancaria como resultado preliminar después de la realización del etiquetado manual. Gráfico desarrollado por el autor.

4.2.4 Procesamiento del lenguaje natural

El procesamiento del lenguaje natural consiste en otorgarle capacidad a la máquina para entender las palabras escritas en un texto y que de este modo puedan ser procesadas de manera computacional, el primer paso para llegar a esto es la tokenización de los datos, lo cual no es otra cosa más que separar cada palabra formando un arreglo de palabras por medio de una función definida como “tokenize” en la que se ingresa una variable de tipo texto y esta

usa la función Split() para separar cada palabra dentro del texto y aplicándolo al conjunto de datos por medio de la función apply(), haciendo uso de lambda para que este proceso se realice por cada uno de los registros dentro del *data frame*. Esto se logra mediante código realizado en Python que se muestra a continuación:

Figura 14

Segmento de código utilizado para la tokenización del texto

```
def tokenize(txt):
    tokens = re.split('\W+', txt)
    return tokens


Tweets_FinalDf['TextoTokenizado'] = Tweets_FinalDf['TextoProcesado'].apply(lambda x: tokenize(str(x).lower()))

Tweets_FinalDf.head()
```

	Id	Fecha	UserName	Entidad	Texto	TextoProcesado	Sentimiento	TextoTokenizado
0	0	31/5/2022 16:25	ykita21	BcoBolivariano	@BcoBolivariano no se qué pasa con su aplicaci...	no se qué pasa con su aplicación móvil se dem...	NEGATIVO	[, no, se, qué, pasa, con, su, aplicación, móv...
1	1	31/5/2022 14:41	gildamaria	BcoBolivariano	@samanthadarko0 @BcoBolivariano Por suerte “na...	por suerte nada los detiene jaja pésimo ux	NEGATIVO	[, por, suerte, nada, los, detiene, jaja, pési...
2	3	30/5/2022 16:06	BazanDanilo	BcoBolivariano	@BcoBolivariano Hola Megustaria saber si tengo...	hola megustaria saber si tengo mi cuenta del ...	NEUTRO	[, hola, megustaria, saber, si, tengo, mi, cue...
3	4	30/5/2022 16:00	Ruiz12Evelyn	BcoBolivariano	@BcoBolivariano Por favor que el 1700 505050 n...	por favor que el 1700 505050 no vale	NEGATIVO	[, por, favor, que, el, 1700, 505050, no, vale]
4	5	30/5/2022 15:31	alyPe24	BcoBolivariano	@BcoBolivariano agencia mall del sol una sola ...	agencia mall del sol una sola chica para aten...	POSITIVO	[, agencia, mall, del, sol, una, sola, chica, ...

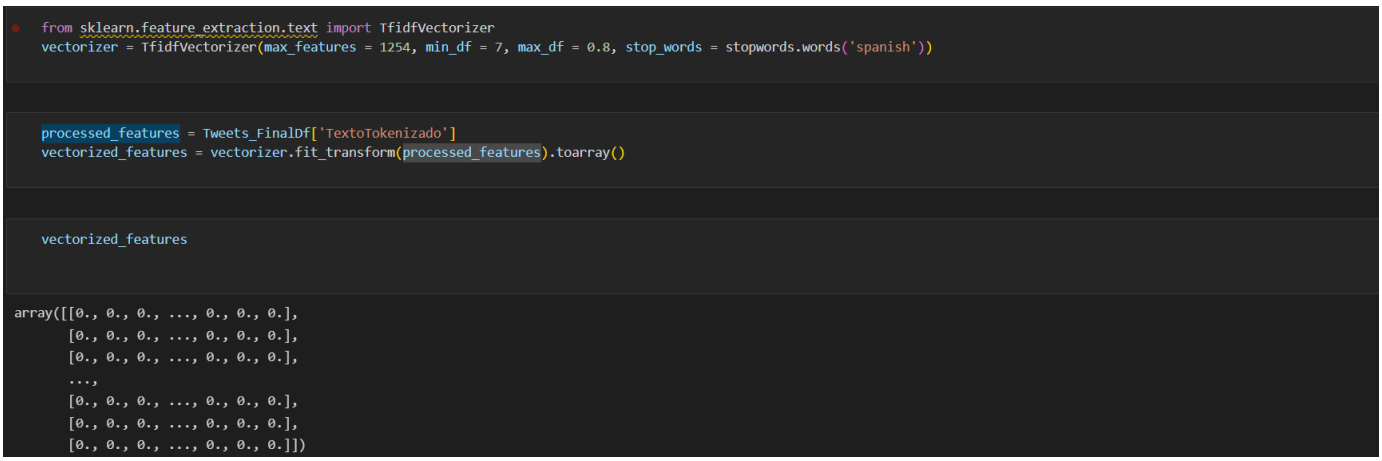
Nota. Imagen extraída del código fuente desarrollado por el autor.

Paso seguido a esto se procede a otorgarles a estas palabras un valor numérico para que puedan ser entendidas por la máquina, esto eso es gracias al algoritmo TF-IDF, que significa Frecuencia de término - Frecuencia de documento inversa, el cual es un método para medir la relevancia de las palabras en los documentos de una colección de frases o palabras. TF-IDF tiene muchos usos, como la recuperación de información, el análisis de texto, la extracción de palabras clave y la obtención de características numéricas de texto para algoritmos de aprendizaje automático y este último es para lo que fue utilizado, mediante Python haciendo uso de la librería SKlearn,

especializada para ciencia de datos, con su módulo TFidfVectorizer creando así la variable “processed_features” la cual contiene un arreglo de todos los pesos obtenidos para cada palabra como se muestra a continuación:

Figura 15

Segmento de código en el que se utiliza el algoritmo TF-IDF para procesar el lenguaje natural en datos numéricos.



Nota. Imagen obtenida del código fuente desarrollado por el autor.

4.2.5 Separación de los datos

Esta etapa consiste en dos procesos, el primero es separar los datos que será utilizada por el modelo de *machine learning* para realizar las predicciones de los datos resultantes, en este caso se separan el campo “processed_features”, el cual será utilizado para las predicciones del algoritmo de *machine learning*, y el campo “Sentimiento”, el cual será el resultado deseado de la predicción realizada por el algoritmo de *machine learning*.

El segundo proceso consiste en muestrear estos dos datos en conjuntos de entrenamiento del modelo y testeo del modelo, para eso se realizaron 3 iteraciones, en las cuales se realiza una muestra aleatoria simple de los datos

en 3 conjuntos, en el primero se escoge un 70% de los datos para realizar el entrenamiento del algoritmo y un 30% para testeo del mismo; en la segunda iteración se escogió un 80% de los datos para entrenamiento y un 20% para testeo; y, por último, para la tercera iteración se escogió un 90% de los datos para entrenamiento y un 10% para testeo. Todo esto se logra mediante la librería SciKit-learn importando su módulo de train_test_split y haciendo uso de este para separar los datos en datos de entrenamiento y datos de prueba para cada una de las variables como se muestra a continuación:

Figura 16

Segmento de código utilizado para dividir los datos.

```
label = Tweets_FinalDf['Sentimiento']


from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(vectorized_features, label, test_size = 0.3, random_state=1, shuffle=True)
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Como se puede ver en la imagen anterior, se separan los datos en un conjunto de prueba del 30% y automáticamente se creará un conjunto de entrenamiento con el 70% restante de los datos, esta separación se realiza de manera aleatoria dentro del conjunto de datos. Este proceso debe ser realizado para cada instancia cambiando el parámetro test_size dependiendo de que porcentaje de datos se desean utilizar para entrenamiento y prueba de los algoritmos de *machine learning*.

4.2.6 Entrenamiento y clasificación

Para esta penúltima etapa, se va a hacer uso de la librería SKlearn la cual contiene un conjunto muy útil de algoritmos de *machine learning* tanto supervisados como no supervisados, en este caso se hará uso de 5 algoritmos de *machine learning* supervisados de clasificación.

4.2.6.1 Naive Bayes

El primer algoritmo de clasificación utilizado para el desarrollo del proyecto es el llamado Naive Bayes, el cual es una colección de algoritmos de clasificación basados en el teorema de Bayes, siendo una familia de algoritmos de comparten un principio común en el cada par de características que se clasifican son independientes entre sí. (Loukas, 2020)

Este teorema se basa en la probabilidad de que un acontecimiento suceda dada la probabilidad de un acontecimiento que ya sucedió, esto matemáticamente se puede escribir como:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

En donde A y B son los eventos; P(A|B) es la probabilidad de que el evento A suceda dado el evento B; P(A) es la probabilidad individual de que el evento A suceda; y P(B|A) es la probabilidad de que el evento B suceda después de ver la evidencia A. (Loukas, 2020)

Para el caso del presente proyecto de investigación se hace uso del módulo “naive_bayes” de la librería SKlearn para aplicar un algoritmo Naive bayes multinomial al set de datos, el algoritmo multinomial es adecuado para la clasificación de características discretas, como es el caso del recuento de

palabras para la clasificación de texto. A continuación, se muestra el segmento de código y su aplicación para el desarrollo de esta tarea:

Figura 17

Segmento de código utilizado para el entrenamiento y prueba del algoritmo de Naive Bayes

```
from sklearn import naive_bayes

NB_text_classifier = naive_bayes.MultinomialNB()
NB_model = NB_text_classifier.fit(X_train, Y_train)

NB_predictions = NB_model.predict(X_test)
```

```
print(confusion_matrix(Y_test,NB_predictions))
print(classification_report(Y_test,NB_predictions))
print(accuracy_score(Y_test, NB_predictions))
```

```
[[1096  391   43]
 [ 333 1057   28]
 [ 185  179  276]]
              precision    recall  f1-score   support

    NEGATIVO       0.68      0.72      0.70      1530
      NEUTRO       0.65      0.75      0.69      1418
    POSITIVO       0.80      0.43      0.56       640

    accuracy                           0.68      3588
   macro avg       0.71      0.63      0.65      3588
weighted avg       0.69      0.68      0.67      3588

0.6769788182831661
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Como se puede observar, en primer lugar se hace llamado al módulo “naive_bayes” de la librería sklearn como se mencionó anteriormente para

paso siguiente se crea una instancia, dentro de la variable “NB_text_classifier”, del algoritmo de naive bayes que se va a utilizar, en este caso naive bayes multinomial, luego de esto se crea la variable “NB_Model” en el cual se van a ingresar los datos de entrenamiento dentro de la instancia anteriormente creada haciendo uso de la función fit() y después de esto se obtienen las predicciones del modelo con los datos de prueba. Por último, gracias a la librería sklearn es posible obtener la matriz de confusión de los resultados del modelo, la cual muestra las predicciones correctas e incorrectas realizadas por el algoritmo, así como sus valores de precisión, exhaustividad y f1-score de los cuales se hablará en la comparación de los resultados.

4.2.6.2 Árboles de decisión

Los árboles de decisión son algoritmos de *machine learning* que sirven tanto para la construcción de modelos de clasificación como de modelos de regresión, para el caso de la presente investigación se usa para construir un modelo de clasificación en el cual se pretende predecir en valor de un variable clasificando la información dada en función de otras variables. (UNIR, 2021)

Este se construye por medio de nodos internos, los cuales representan las características que se tendrán en cuenta para la toma de decisión; las ramas las cuales representan la decisión tomada en base a una condición; y los nodos finales el cual es el resultado final de la decisión tomada por el algoritmo.

Para aplicar este algoritmo dentro del proyecto desarrollado, se hace uso del módulo “DecissionTreeClassifier” de la librería SKlearn con el cual se

puede aplicar el algoritmo de árbol de decisión como se muestra a continuación:

Figura 18

Segmento de código utilizado para el entrenamiento y prueba del algoritmo de árbol de decisión

```
from sklearn.tree import DecisionTreeClassifier

DT_text_classifier = DecisionTreeClassifier()
DT_model = DT_text_classifier.fit(X_train, Y_train)
DT_predictions = DT_model.predict(X_test)
```

```
print(confusion_matrix(Y_test,DT_predictions))
print(classification_report(Y_test,DT_predictions))
print(accuracy_score(Y_test, DT_predictions))
```

```
[[956 467 107]
 [385 954  79]
 [126 139 375]]
              precision    recall  f1-score   support

    NEGATIVO       0.65      0.62      0.64      1530
      NEUTRO       0.61      0.67      0.64      1418
    POSITIVO       0.67      0.59      0.62       640

    accuracy                           0.64      3588
   macro avg       0.64      0.63      0.63      3588
weighted avg       0.64      0.64      0.64      3588

0.6368450390189521
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Este proceso es similar al de naive bayes, en el que primero se crea una instancia del algoritmo dentro de la variable “DT_text_classifier” y dentro del cual se utilizan los datos de entrenamiento creando así la variable “DT_model” que contiene al algoritmo entrenado, por consiguiente, es posible obtener sus predicciones haciendo uso de la función predict() junto al conjunto de datos de prueba y obtenido sus resultados posteriormente.

4.2.6.3 K vecinos cercanos

El K-nearest neighbors, o por su traducción en español K vecinos cercanos, es un algoritmo de *machine learning* de tipo supervisado, el cual tiene una aplicación de fácil uso y es utilizado para problemas de clasificación y regresión.

Este se puede dividir en 5 etapas, en la primera etapa se selecciona el número de K vecinos; en la segunda etapa se calcula la distancia entre los vecinos; en la tercera etapa se toman los K vecinos más cercanos según la distancia calculada; en la cuarta etapa se cuenta el número de vecinos cercanos de cada categoría entre todos los K vecinos; y por último, en la quinta etapa se les atribuye un nuevo punto a la categoría que se encuentre más presente entre los K vecinos (datascientest.com, 2021). Y a continuación se muestra su implementación:

Figura 19

Segmento de código utilizado para el entrenamiento y prueba del algoritmo de k vecinos cercanos

```
from sklearn.neighbors import KNeighborsClassifier

KNN_text_classifier = KNeighborsClassifier()

KNN_model = KNN_text_classifier.fit(X_train, Y_train)
KNN_predictions = KNN_model.predict(X_test)


print(confusion_matrix(Y_test,KNN_predictions))
print(classification_report(Y_test,KNN_predictions))
print(accuracy_score(Y_test, KNN_predictions))
```

```
[[843 574 113]
 [341 988  89]
 [158 228 254]]
              precision    recall  f1-score   support

    NEGATIVO       0.63      0.55      0.59      1530
      NEUTRO       0.55      0.70      0.62      1418
    POSITIVO       0.56      0.40      0.46       640

    accuracy                           0.58      3588
   macro avg       0.58      0.55      0.56      3588
weighted avg       0.59      0.58      0.58      3588

0.5811036789297659
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Su uso dentro del proyecto es similar a los algoritmos anteriormente mencionados, se debe importar el módulo correspondiente al algoritmo llamado “KNeighborsClassifier” de la librería sklearn y por consiguiente se debe crear una instancia de este, haciéndolo dentro de la variable llamada “KNN_text_classifier”. El paso para seguir, igual que antes, es el de entrenar al algoritmo con los datos de entrenamiento haciendo uso de la función fit() y

guardando al algoritmo entrenado dentro de la variable “KNN_Model” y por último poder obtener sus predicciones por medio de la función predict() con el conjunto de datos de prueba.

4.2.6.4 Random Forest

El algoritmo de *machine learning* llamado random forest consiste la creación y combinación de múltiples árboles de decisión generados de forma aleatoria, en el cual se ajustan estos árboles clasificadores y hace uso de su promedio de precisión para mejorar en sí mismo sus predicciones. Para la realización de este se deben especificar el número de árboles de decisión que se van a general, para el presente trabajo, durante las 3 iteraciones de ejecución del modelo, se generarán 250 árboles de decisión. A continuación, se muestra su implementación:

Figura 20

Segmento de código utilizado para el entrenamiento y prueba del algoritmo de random forest

```
from sklearn.ensemble import RandomForestClassifier

RF_text_classifier = RandomForestClassifier(n_estimators = 250, random_state = 0)
RF_model = RF_text_classifier.fit(X_train, Y_train)
```

```
RF_predictions = RF_model.predict(X_test)
```

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

print(confusion_matrix(Y_test,RF_predictions))
print(classification_report(Y_test,RF_predictions))
print(accuracy_score(Y_test, RF_predictions))
```

```
[[1129  343   58]
 [ 345 1034   39]
 [ 107  140  393]]
              precision    recall  f1-score   support

    NEGATIVO       0.71      0.74      0.73      1530
      NEUTRO       0.68      0.73      0.70      1418
    POSITIVO       0.80      0.61      0.70       640

    accuracy                           0.71      3588
   macro avg       0.73      0.69      0.71      3588
weighted avg       0.72      0.71      0.71      3588

0.7123745819397993
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Al igual que los algoritmos anteriores, los pasos para la implementación del algoritmo de random Forrest se repiten iniciando por importar el módulo de la librería SKlearn perteneciente al algoritmo llamado “RandoForestClassifier”. La única diferencia con respecto al proceso de aplicación de los algoritmos anteriormente mencionados, es la de que al momento de crear su instancia dentro de la variable RF_text_classifier es necesario definir la cantidad de

árboles de decisión que se ejecutarán y la semilla con la que se generarán aleatoriamente los árboles de decisión dentro del Random Forest. Después de esto los pasos a seguir son los mismos que los mencionados anteriormente en los que se usa la instancia anteriormente creada para entrenar el algoritmo junto a los datos de entrenamiento y almacenarlo dentro de la variable "RF_model". Por último, se obtienen y almacenan sus predicciones dentro de la variable "RF_predictions" y se obtienen sus resultados de rendimiento junto a su matriz de confusión.

4.2.6.5 Máquina de soporte vectorial

El algoritmo de *machine learning* máquina de soporte vectorial, por sus siglas en inglés SVM, es un algoritmo supervisado utilizado tanto para clasificación como para regresión, este determina cual es el mejor límite de decisión entre vectores que pertenecen a una categoría determinada y los vectores que no. Esto se puede aplicar a cualquier tipo de vector que codifique cualquier tipo de dato, esto quiere decir que para aprovecha el poder de clasificación de texto de SVM, estos deben ser convertidos a vectores.

Para su implementación, dentro de la creación de la instancia es necesario definir el tipo de máquina de soporte vectorial que se utiliza para el proceso de clasificación, en este caso se define como una máquina de soporte vectorial lineal como se muestra a continuación:

Figura 21

Segmento de código utilizado para el entrenamiento y prueba del algoritmo de máquina de soporte vectorial

```
from sklearn import svm

#Clasificando usando soporte de máquina vectorial
SVM_text_classifier = svm.SVC(kernel = 'linear', probability = True )

#Fit the model
SVM_model = SVM_text_classifier.fit(X_train, Y_train)

#hacemos la clasificación y predicción con los datos de prueba
SVM_predictions = SVM_model.predict(X_test)
```

```
print(confusion_matrix(Y_test,SVM_predictions))
print(classification_report(Y_test,SVM_predictions))
print(accuracy_score(Y_test, SVM_predictions))
```

```
[[1119  371   40]
 [ 329 1086    3]
 [ 132  145  363]]
              precision    recall  f1-score   support

    NEGATIVO       0.71      0.73      0.72      1530
      NEUTRO       0.68      0.77      0.72      1418
    POSITIVO       0.89      0.57      0.69       640

    accuracy                           0.72      3588
   macro avg       0.76      0.69      0.71      3588
weighted avg       0.73      0.72      0.71      3588

0.7157190635451505
```

Nota. Imagen obtenida del código fuente desarrollado por el autor.

Una vez importado el módulo svm de la librería SKlearn, que corresponde al algoritmo de máquina de soporte vectorial, se crea su instancia haciendo un llamado a la función SVC() y definiendo el Kernel de ese algoritmo como “linear” para que, como se mencionó anteriormente, el algoritmo sea de tipo lineal. Los pasos a seguir, al igual que los casos anteriores, son los de entrenar al algoritmo haciendo uso de la función fit() y almacenándolo dentro de la variable SVM_model, par que al finalizar esto poder obtener sus predicciones haciendo uso del conjunto de datos de prueba y almacenándolas

dentro de la variable llamada “SMV_predictions” para que de esta forma sea posible obtener los resultados de su rendimiento los cuales serán comparados con los resultados de rendimiento de los demás algoritmos implementados en el siguiente punto presentado a continuación.

4.2.7 Comparación de rendimiento

Como última etapa en el proceso de la realización de un modelo basado en minería de datos para la realización de un análisis de sentimientos, se procede a realizar la comparación de los resultados de rendimiento obtenido tras la ejecución del entrenamiento y prueba de los algoritmos de *machine learning* seleccionados para este caso en cada una de sus iteraciones según el muestreo de los datos.

Para esto se tendrá en cuenta 3 valores que se pueden obtener mediante la librería de ciencia de datos de Python SKlearn los cuales son los siguientes (Heras, 2020):

- **Precisión:** Con esta se es capaz de medir la calidad del algoritmo implementado en tareas de clasificación, en sí como su nombre lo indica defina la precisión porcentual con la que responde a la pregunta que se está formulando al momento de realizar la clasificación, como por ejemplo ¿Qué porcentaje de usuarios van a comprar un producto?

- **Recall:** Este valor indica la exhaustividad de nuestro algoritmo implementado, es decir que va a dar información porcentual de la

cantidad de resultados que nuestro algoritmo es capaz de identificar correctamente.

- **F1-Score:** También conocido como valor F1, se trata de la combinación de las dos medidas antes mencionadas con la finalidad de poder facilitar la comparación del rendimiento de los algoritmos.

En la primera iteración, así como se mencionó anteriormente, tenemos la partición de los datos en un 70% para entrenamiento del de los algoritmos y un 30% para el testeo de estos, en este se obtuvieron los siguientes resultados de rendimiento:

Tabla 8

Resultados de la precisión de los algoritmos ejecutados en la primera iteración

	PRECISIÓN		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	65%	65%	79%
ÁRBOL DE DECISIÓN	63%	61%	66%
VECINOS CERCANOS	61%	56%	55%
RANDOM FOREST	69%	68%	79%
MÁQUINA DE SOPORTE VECTORIAL	68%	67%	91%

Nota. Esta tabla muestra los valores porcentuales de la precisión de clasificación de los algoritmos ejecutados dentro de la primera iteración dentro del modelo de minería de datos, tabla creada por el autor.

Tabla 9

Resultados de la exhaustividad de los algoritmos ejecutados en la primera iteración

	RECALL		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	71%	72%	42%
ÁRBOL DE DECISIÓN	62%	64%	58%
VECINOS CERCANOS	58%	68%	37%
RANDOM FOREST	73%	71%	61%
MÁQUINA DE SOPORTE VECTORIAL	72%	74%	56%

Nota. Esta tabla muestra los valores porcentuales de la precisión de clasificación de los algoritmos ejecutados en la primera iteración dentro del modelo de minería de datos, tabla creada por el autor

Tabla 10

Resultados del F1-Score de los algoritmos ejecutados en la primera iteración

	F1-SCORE		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	68%	68%	55%
ÁRBOL DE DECISIÓN	63%	62%	62%
VECINOS CERCANOS	59%	62%	44%
RANDOM FOREST	71%	69%	69%
MÁQUINA DE SOPORTE VECTORIAL	70%	70%	69%

Nota. Esta tabla muestra los valores porcentuales del F1-Score de clasificación de los algoritmos ejecutados dentro del modelo de minería de datos, tabla creada por el autor.

Con estos resultados podemos concluir que el peor algoritmo en esta iteración es vecinos cercanos debido a que tiene la menor cantidad de predicciones acertadas hacia todos los casos de clasificación, destacando su baja exhaustividad para la clasificación de “Positivo” por lo cual indica que para cualquier comentario que realmente pertenezca a esta clasificación se le hará difícil identificarlo. Así mismo, los mejores algoritmos de clasificación en esta primera iteración son el Random Forest y la máquina de soporte vectorial

debido a que mantienen un balance entre sus predicciones y ambos mantienen resultados bastante similares.

En la segunda iteración, tenemos la partición de los datos en un 80% para entrenamiento del de los algoritmos y un 20% para el testeo de estos, en el cual se obtuvieron los siguientes resultados:

Tabla 11

Resultados de la precisión de los algoritmos ejecutados en la segunda iteración

	PRECISIÓN		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	68%	65%	80%
ÁRBOL DE DECISIÓN	65%	61%	67%
VECINOS CERCANOS	63%	55%	56%
RANDOM FOREST	71%	68%	80%
MÁQUINA DE SOPORTE VECTORIAL	71%	68%	89%

Nota. Esta tabla muestra los valores porcentuales de la precisión de clasificación de los algoritmos ejecutados dentro de la segunda iteración dentro del modelo de minería de datos, tabla creada por el autor.

Tabla 12

Resultados de la exhaustividad de los algoritmos ejecutados en la segunda iteración

	RECALL		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	72%	75%	43%
ÁRBOL DE DECISIÓN	62%	67%	59%
VECINOS CERCANOS	55%	70%	40%
RANDOM FOREST	74%	73%	61%
MÁQUINA DE SOPORTE VECTORIAL	73%	77%	57%

Nota. Esta tabla muestra los valores porcentuales de la precisión de clasificación de los algoritmos ejecutados en la segunda iteración dentro del modelo de minería de datos, tabla creada por el autor.

Tabla 13

Resultados del F1-Score de los algoritmos ejecutados en la segunda iteración

	F1-SCORE		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	70%	69%	56%
ÁRBOL DE DECISIÓN	64%	64%	62%
VECINOS CERCANOS	59%	62%	46%
RANDOM FOREST	73%	70%	70%
MÁQUINA DE SOPORTE VECTORIAL	72%	72%	69%

Nota. Esta tabla muestra los valores porcentuales del F1-Score de clasificación de los algoritmos ejecutados en la segunda iteración dentro del modelo de minería de datos, tabla creada por el autor.

Así mismo como en el caso anterior, en esta iteración el peor algoritmo aplicado es el de vecinos cercanos nuevamente debido sus valores de precisión que en las clasificaciones de “Neutro” y “Positivo” no alcanzan ni el 60%. Y, al contrario, el mejor vuelve a ser el Random Forest gracias a que

sigue manteniendo su balance al momento de obtener sus resultados, y a pesar de que en la precisión de la clasificación de “Positivo” la máquina de soporte vectorial es superior, este es mejor en su exhaustividad y en su F1-Score

Y para finalizar con la comparación de rendimiento de los algoritmos aplicados, se hace la ejecución de la tercera iteración en la que se utiliza un 90% de los datos totales para entrenamiento del algoritmo y un 10% de estos para el testeo.

Tabla 14

Resultados de la precisión de los algoritmos ejecutados en la tercera iteración

	PRECISIÓN		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	65%	67%	80%
ÁRBOL DE DECISIÓN	64%	63%	67%
VECINOS CERCANOS	60%	57%	59%
RANDOM FOREST	70%	70%	81%
MÁQUINA DE SOPORTE VECTORIAL	69%	70%	89%

Nota. Esta tabla muestra los valores porcentuales de la precisión de clasificación de los algoritmos ejecutados dentro de la tercera iteración dentro del modelo de minería de datos, tabla creada por el autor.

Tabla 15

Resultados de la exhaustividad de los algoritmos ejecutados en la tercera iteración

	RECALL		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	73%	73%	42%
ÁRBOL DE DECISIÓN	64%	66%	58%
VECINOS CERCANOS	57%	68%	39%
RANDOM FOREST	75%	73%	60%
MÁQUINA DE SOPORTE VECTORIAL	73%	77%	57%

Nota. Esta tabla muestra los valores porcentuales de la precisión de clasificación de los algoritmos ejecutados en la tercera iteración dentro del modelo de minería de datos, tabla creada por el autor.

Tabla 16

Resultados del F1-Score de los algoritmos ejecutados en la tercera iteración

	F1-SCORE		
	NEGATIVO	**NEUTRO**	**POSITIVO**
NAIVE BAYES	69%	70%	55%
ÁRBOL DE DECISIÓN	64%	64%	62%
VECINOS CERCANOS	59%	62%	47%
RANDOM FOREST	73%	72%	69%
MÁQUINA DE SOPORTE VECTORIAL	71%	73%	69%

Nota. Esta tabla muestra los valores porcentuales del F1-Score de clasificación de los algoritmos ejecutados en la tercera iteración dentro del modelo de minería de datos, tabla creada por el autor.

En esta última iteración, los resultados del rendimiento de los algoritmos aplicados son similares a los casos anteriores en el que el peor de estos algoritmos, una vez más, vuelve a ser vecinos cercanos debido a su baja

exhaustividad. Así mismo se repite el caso de los mejores algoritmos aplicados, los cuales nuevamente son el Random Forrest y la máquina de soporte vectorial en el que se vuelven a apreciar resultados muy similares en los que se puede encontrar una alta precisión a la hora de valorar los comentarios Positivos. No obstante, en este caso también puede se puede resaltar al algoritmo *naive bayes* que también muestra un alto índice de precisión, pero se queda al momento de obtener una exhaustividad de los comentarios positivos.

Se puede concluir que bajo de cualquiera de los 3 modos de entrenamiento y prueba que fueron analizados en el presente trabajo de investigación, se puede elegir a la máquina de soporte vectorial o al random Forrest como algoritmo de clasificación para el modelo de minería de datos debido a que ambos mantienen resultados muy similares destacando su precisión a diferencia de los otros algoritmos puestos a prueba.

4.3 Análisis de predicciones del modelo

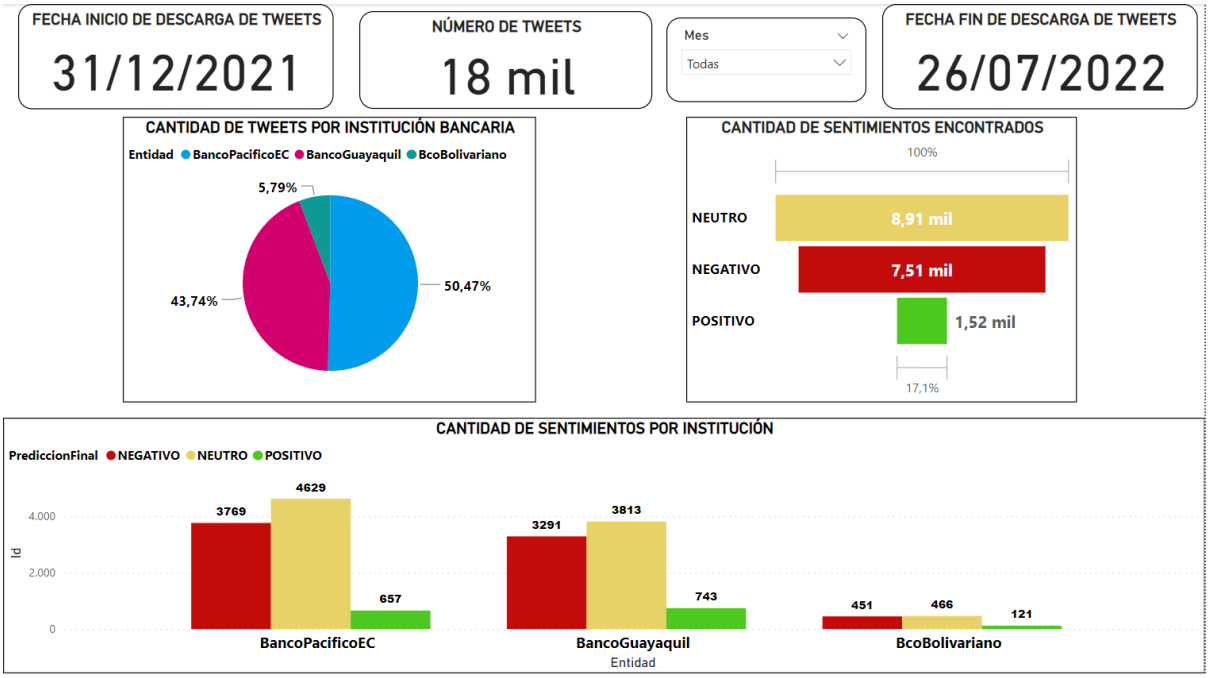
Una vez obtenido el diseño y de haber desarrollado el modelo de minería de datos, es posible la obtención de predicciones sobre un nuevo conjunto de datos que no necesitan de un etiquetado preliminar, dejando a la máquina realizar por completo este etiquetado de manera automática. Aun así, se pondrá a prueba el modelo de minería de datos con el mismo set de datos con el que se realizó su entrenamiento para poder realizar una comparación de las predicciones autónomas obtenidas por el modelo con las etiquetas que ya existían en el set de datos.

Finalmente, el algoritmo seleccionado para funcionar dentro del modelo de minería de datos desarrollado fue el de la máquina de soporte vectorial con base a la tercera iteración en el que se usó un 90% de los datos anteriores para su entrenamiento y un 10% para su testeo, esto debido a que nos otorga una mayor capacidad de predicción hacia la categoría de “Positivo”.

Y gracias a la herramienta PowerBI se pudo desarrollar un tablero en el que se visualizan las predicciones finales del modelo desarrollado, tablero que se muestra continuación:

Figura 22

Dashboard desarrollado en PowerBI con las predicciones del modelo de minería de datos desarrollado



Nota. En este gráfico podemos presenciar un resumen gerencial de lo que fueron las predicciones otorgadas por el modelo de minería de datos desarrollado en el presente caso de investigación, gráfico realizado por el auto

Con esto, junto a las predicciones del modelo, podemos observar que, en general, los clientes de la entidad bancaria Banco del Pacífico son los que poseen una mayor participación al momento de realizar opiniones dentro de la red social Twitter, y las predicciones nos indican que la gran mayoría de estas opiniones son comentarios neutros seguido de una gran cantidad de comentarios negativos, lo mismo se repite dentro del Banco de Guayaquil y dentro del banco Bolivariano, así como resumen general es posible observar que de los 18.000 tweets obtenidos, 8.910 pertenecen a opiniones neutras,

7.510 pertenecen a opiniones negativas y únicamente 1.520 pertenecen a opiniones positivas.

Es posible notar que los valores de las predicciones no coindicen por completo con los valores propuestos anteriormente en los que se dan a mostrar la cantidad de sentimientos y tweets que existen dentro del conjunto de datos, esto es debido a que ahora el etiquetado de datos es realizado por completo por la máquina gracias al poder de predicción del modelo desarrollado utilizando un algoritmo de máquina de soporte vectorial que otorga un 71% de precisión para sus predicciones. Es decir que existe un 29% de posibilidades de fallar al momento de predecir a que sentimiento pertenece cada tweet.

CONCLUSIONES

Para obtener la información de opiniones de los usuarios en relación con las entidades bancarias con calificación de riesgo mayor que AA de la ciudad de Guayaquil extraídas de la red social Twitter, se realizó una extracción mediante Python utilizando la librería *snscrape* para poder descargar tweets en los que se mencionasen a las cuentas del Banco Bolivariano, Banco de Guayaquil o al Banco del Pacífico en un periodo total del primero de enero del 2022 hasta el primero de Julio del 2022 en el cual se pudieron descargar un total de 26.955 registros.

Para lograr el objetivo de diseñar una estrategia de minería de datos para identificar los sentimientos de los clientes de las entidades bancarias con calificación de riesgo mayor que AA de la ciudad de Guayaquil con datos recopilados desde Twitter se buscaron y analizaron tanto fuentes bibliográficas como otros proyectos de índole similar sobre algoritmos de aprendizaje de máquina para poder aplicar las mejores metodologías y se logró graficar la arquitectura a seguir del modelo de minería de datos en base a la metodología KDD, en el que se encuentran 7 módulos de los cuales cada uno depende de la realización del anterior.

Por último, el objetivo de construir un modelo de minería de datos aplicando algoritmos de aprendizaje de máquina para clasificar los tweets basado en los sentimientos de los clientes de las entidades bancarias privadas con calificación de riesgo mayor que AA de la ciudad de Guayaquil se consiguió aplicando cada uno de los módulos graficados en el diseño del modelo. Con esto, al conjunto de datos creado se le extrajo el campo de texto obtenido y se le realizó una limpieza eliminando las menciones, los hashtags, los

enlaces, emojis y símbolos especiales. Además, se eliminaron palabras de conexión tales como “de”, “la”, “y”, “por”, entre otras, todo esto con la finalidad de realizar un procesamiento del lenguaje natural otorgándolas de un peso numérico a cada palabra por medio del algoritmo de vectorización TF-IDF haciendo posible que la máquina entienda estas palabras. Además, se aplicaron 5 algoritmos de clasificación: *naive bayes*, vecinos cercanos, árboles de decisión, random forest, y máquina de soporte vectorial. Con ellos se realizaron 3 iteraciones en los que se separó al conjunto de datos en un set de datos de entrenamiento y un set de datos de testeo en los que se escogió para la primera iteración un 70% de los datos para entrenamiento y un 30% para testeo; en la segunda iteración se escogió un 80% de los datos para entrenamiento y un 20% para testeo; y, por último, para la tercera iteración se escogió un 90% de los datos para entrenamiento y un 10% para testeo. Por último, se procedió a realizar una comparación de los resultados de rendimiento de cada uno de los modelos por medio de sus valores de precisión, exhaustividad llamada recall y la combinación de estos dos últimos llamado F1-Score con lo cual se pudo realizar la elección del mejor modelo realizado para empezar a predecir con nueva información.

RECOMENDACIONES

Como recomendaciones a futuras investigaciones de la misma índole se podría sugerir que:

- Para el proceso de construcción del conjunto de datos se descarguen información de un periodo más amplio.
- Se podría hacer uso de los módulos de *snscrape* de Facebook e Instagram para obtener datos de otras fuentes.
- Se debería realizar más iteraciones con nuevas particiones de datos para entrenar a los algoritmos de *machine learning*.
- Como la tecnología avanza constantemente, se debería hacer una búsqueda de nuevos algoritmos de clasificación y ponerlos a prueba junto a los otros presentados en el trabajo, para obtener nuevos resultados y posiblemente un mejor modelo de clasificación
- Se podría hacer uno de otro lenguaje de programación, tal como R, especializado en análisis de datos para la comparación de estos y obtener cual es el mejor para desarrollar modelos de minería de datos de clasificación

REFERENCIAS BIBLIOGRAFICAS

Banco Bolivariano C.A. (2017). *CÓDIGO DE BUEN GOBIERNO.* Obtenido de bolivariano: https://www.bolivariano.com/docs/default-source/general-pdf/gobierno-corporativo/codigobuengobierno_2017.pdf

Banco de Guayaquil. (s.f.). *Nuestra Historia*. Obtenido de bancoguayaquil: https://www.bancoguayaquil.com/conocenos/

Banco del Pacífico. (s.f.). *Nuestra historia*. Obtenido de Grupo Banco del Pacífico: https://www.bancodelpacifico.com/grupo-bdp/grupo-banco-del-pacifico/menu/nuestra-historia

Banco Internacional. (05 de Febrero de 2021). *¿Qué es y cómo funciona el sistema financiero ecuatoriano?* Obtenido de bancointernacional: https://www.bancointernacional.com.ec/que-es-y-como-funciona-el-sistema-financiero-ecuatoriano/

Banco Mundial. (s.f.). *Sector financiero*. Obtenido de BancoMundial: https://www.bancomundial.org/es/topic/financialsector/overview

Bannister, K. (2015). *Entendiendo el análisis de sentimiento: qué es y para qué se usa*. Obtenido de brandwatch: https://www.brandwatch.com/es/blog/analisis-de-sentimiento/#:~:text=El%20análisis%20de%20sentimiento%20es,pública%20general%20sobre%20ciertos%20temas.&text=La%20habilidad%20de%20extraer%20información,adoptando%20organizaciones%20a%20nivel%20mundial.

Coba, G. (11 de Febrero de 2021). *Se duplica el número de usuarios del sistema financiero en Ecuador*. Obtenido de Primicias: https://www.primicias.ec/noticias/economia/poblacion-adulta-sistema-financiero-ecuador-acceso/

Collaguazo, D. (20 de Junio de 2017). *¿Qué es el Procesamiento de Lenguaje Natural y cómo ponerlo en práctica con recursos abiertos?* Obtenido de IADB: https://blogs.iadb.org/conocimiento-abierto/es/que-es-el-procesamiento-de-lenguaje-natural-y-como-ponerlo-en-practica-con-recursos-abiertos/#:~:text=2%20Algunas%20aplicaciones%20del%20Procesamiento%20de%20Lenguaje%20Natural&text=También%20se%20le%20atribu

Corporación Financiera Nacional. (2017). *MODULO-III-PRODUCTOS-Y-SERVICIOS-DEL-SISTEMA-FINANCIERO-ECUATORIANO .*

Cruz, A. (2020). *Redes Sociales*. Obtenido de rdstation: https://www.rdstation.com/es/redes-sociales/

datascientest.com. (28 de diciembre de 2021). *¿Qué es el algoritmo KNN?* Obtenido de datascientest: https://datascientest.com/es/que-es-el-algoritmo-knn

Dávalos, N. (1 de Febrero de 2021). *En Ecuador, el 78,7% de los ciudadanos usa redes sociales*. Obtenido de https://www.primicias.ec/noticias/tecnologia/14-millones-ecuatorianos-usuarios-redes-sociales/:

https://www.primicias.ec/noticias/tecnologia/14-millones-ecuatorianos-usuarios-redes-sociales/

Duò, M. (26 de Enero de 2022). *Tu Guía de las Mejores Herramientas de Visualización de Datos en 2022*. Obtenido de kinsta: https://kinsta.com/es/blog/herramientas-de-visualizacion-de-datos/

Eisenstein, J. (2018). *Natural Language Processing.*

Espinosa, B. (27 de diciembre de 2021). *Utilidades de bancos privados crecieron un 65% hasta noviembre de 2021*. Obtenido de Radio Pichincha: https://www.pichinchacomunicaciones.com.ec/utilidades-de-bancos-privados-crecieron-un-65-hasta-noviembre-de-2021/

European Knowledge Center for Information Technology. (16 de septiembre de 2019). *Base de datos SQL.* Obtenido de TIC Portal. : https://www.ticportal.es/glosario-tic/base-datos-sql

García, S., Ramírez-Gallego, S., Luengo, J., & Herrera, F. (2016). *Big Data: Preprocesamiento.* Obtenido de Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de: https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf

Gartner. (2022). *Gartner Magic Quadrant for Analytics and Business Intelligence Platforms 2022.* Obtenido de https://www.bitec.es/noticias-bitec/cuadrante-magico-de-gartner-2022-analisis-y-business-intelligence/

Gil, P. (2021). *What Is Twitter & How Does It Work?* Obtenido de lifewire: https://www.lifewire.com/what-exactly-is-twitter-2483331

Gupta, S. (7 de Enero de 2018). *Sentiment Analysis: Concept, Analysis and Applications*. Obtenido de Towards Data Science: https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Hall, S. (26 de enero de 2022). *DIGITAL REPORT 2022: EL INFORME SOBRE LAS TENDENCIAS DIGITALES, REDES SOCIALES Y MOBILE.* Obtenido de wearesocial: https://wearesocial.com/es/blog/2022/01/digital-report-2022-el-informe-sobre-las-tendencias-digitales-redes-sociales-y-mobile/

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques.*

Heras, J. M. (Octubre de 2020). *Precision, Recall, F1, Accuracy en clasificación*. Obtenido de IArtificial.net: https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

Herrero, I. (2016). *El análisis de sentimiento de texto en las redes sociales*. Obtenido de BiblogTecarios.

IBM. (s.f.). *¿Qué es Machine Learning?* Obtenido de ibm: https://www.ibm.com/cl-es/analytics/machine-learning#:~:text=Un%20modelo%20de%20machine%20learning,predictivo%20creará%20un%20modelo%20predictivo.

Iglesias, C. A., & Moreno, A. (2019). Sentiment Analysis for Social Media. *Applied science*. Obtenido de https://getthematic.com/sentiment-analysis/

Jiménez, L. R. (2014). *METODOLOGÍA DE CALIFICACIÓN DE RIESGO DE INSTITUCIONES FINANCIERAS Y BANCOS.*

Kemp, S. (2022). *DIGITAL 2022: ECUADOR*. Obtenido de datareportal: https://datareportal.com/reports/digital-2022-ecuador

Lagla, G. A., Moreano, J. A., Arequipa, E. E., & Quishpe, M. W. (2019). Minería de datos como herramienta estratégica. *Revista Científica Mundo de la Investigación y el Conocimiento*, 955-970. Obtenido de https://recimundo.com/index.php/es/article/view/400/599#:~:text=La%20minería%20de%20datos%20es,ser%20aplicado%20en%20las%20empresas.

Li Deng, Y. L. (2018). *Deep Learning in Natural Language Processing.*

López, E. A. (s.f.). *POLITICA FISCAL Y ESTRATEGIA COMO FACTOR DE DESARROLLO DE LA MEDIANA EMPRESA COMERCIAL SINALOENSE. UN ESTUDIO DE CASO.* Obtenido de eumed: https://www.eumed.net/tesis-doctorales/2012/eal/metodologia_cuantitativa.html#:~:text=La%20metodología%20cuantitativa%20de%20acuerdo,o%20fenómeno%20objeto%20de%20estudio.

Loukas, S. (Octubre de 2020). *Text Classification Using Naive Bayes: Theory & A Working Example*. Obtenido de towardsdatascience:

https://towardsdatascience.com/text-classification-using-naive-bayes-theory-a-working-example-2ef4b7eb7d5a#:~:text=2.-,The%20Naive%20Bayes%20algorithm,is%20independent%20of%20each%20other.

Mendoza, M. L. (2020). *Qué es un lenguaje de programación*. Obtenido de openwebinars: https://openwebinars.net/blog/que-es-un-lenguaje-de-programacion/

Montarroso, A. M. (2019). *Análisis de sentimientos para la prevensió de mensajes de odio en las Redes Sociales.* Castilla.

Osman, A. S. (2019). Data Mining Techniques: Review. *INTERNATIONAL JOURNAL OF DATA SCIENCE RESEARCH* .

Ramírez, R. F. (2017). *Nivel de satisfacción de clientes de la banca privada de Guayaquil, respecto a los canales de atención de reclamos.* Obtenido de Universidad Politécnica Salesiana: https://dspace.ups.edu.ec/bitstream/123456789/14885/1/UPS-GT002014.pdf

Ramos, X. (27 de diciembre de 2021). *En el ranking de los bancos de Ecuador que más ganaron durante el 2021, Banco del Pacífico cayó del tercero al décimo lugar respecto al 2020*. Obtenido de https://www.eluniverso.com/noticias/informes/en-el-ranking-de-los-bancos-de-ecuador-que-mas-ganaron-durante-el-2021-banco-del-pacifico-cayo-del-tercero-al-decimo-lugar-respecto-al-2020-nota/

rockcontent. (2019). *Redes sociales: qué son, cómo funcionan, qué tipos existen y cómo influyen en las estrategias de Marketing*. Obtenido de rockcontent: https://rockcontent.com/es/blog/que-son-las-redes-sociales/

Roiger, R. J. (2017). *Data Mining A Tutorial-Based Primer.* Chapman and Hall/CRC.

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Revista Ibérica de Sistemas e Tecnologias de Informação*, 494-506.

Rolando Alfredo Hernández León, S. C. (2014). *El proceso de investigación científica.* Habana: Editorial Universitaria.

Rueda, P. J. (2022). *Aprendizaje supervisado y no supervisado*. Obtenido de healthdataminer: https://healthdataminer.com/data-mining/aprendizaje-supervisado-y-no-supervisado/

Sandoval, L. J. (2018). ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y. *ITCA FEPADE*, 36-40.

SAURA, J. R., REYES-MENENDEZ, A., & PALOS-SANCHEZ, P. (2018). Un Análisis de Sentimiento en Twitter. *Revista Espacios*, 16.

Superintendencia de Bancos. (Diciembre de 2021). *Calificación de Riesgo Instituciones Financieras 2021*. Obtenido de superbancos: https://www.superbancos.gob.ec/bancos/calificacion-de-riesgo-instituciones-financieras-2021/

Tableu. (s.f.). *Las ventajas y beneficios de una buena visualización de datos*. Obtenido de Tableu: https://www.tableau.com/es-mx/learn/articles/data-visualization

The Machine Learners. (2022). *Transformer: la tecnología que domina el mundo*. Obtenido de The Machine Learners: https://www.themachinelearners.com/transformer/#:~:text=Qué%20es%20un%20Transformer,-Directos%20al%20grano&text=Se%20trata%20de%20una%20arquitectura,presentan%20las%20LSTM%20o%20RNN.

UNIR. (Mayo de 2021). *Árboles de decisión: en qué consisten y aplicación en Big Dat*. Obtenido de unir.net: https://www.unir.net/ingenieria/revista/arboles-de-decision/

Vallejo Ballesteros, H. F., Guevara Iñiguez , E., & Medina Velasco, S. R. (2017). Minería de Datos. *Revista Científica Mundo de la Investigación y el Conocimiento*, 347-349.

Valverde, V., Portalanza, N., & Mora, P. (Junio de 2019). *ANÁLISIS DESCRIPTIVO DE BASE DE DATOS RELACIONAL Y NO RELACIONAL.* Obtenido de Atlante: https://www.eumed.net/rev/atlante/2019/06/base-datos-relacional.html

DECLARACIÓN Y AUTORIZACIÓN

Yo, **Roy Steven Mieles Romero**, con C.C: # **0930357249** autor/a del trabajo de titulación: "**Creación de un modelo de minería de datos que identifica sentimientos, de los clientes de la banca privada con calificación de riesgo mayor a AA de la ciudad de Guayaquil con datos basados en Twitter."** previo a la obtención del título de **Ingeniero en Sistemas Computacionales** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, 21 de septiembre del 2022

_____

Nombre: **Mieles Romero Roy Steven**

C.C: **0930357249**

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA			
FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN			
TEMA Y SUBTEMA:	**Creación de un modelo de minería de datos que identifica sentimientos, de los clientes de la banca privada con calificación de riesgo mayor a AA de la ciudad de Guayaquil con datos basados en Twitter.**		
AUTOR(ES)	**Roy Steven Mieles Romero**		
REVISOR(ES)/TUTOR(ES)	**José Miguel Erazo Ayón**		
INSTITUCIÓN:	**Universidad Católica de Santiago de Guayaquil**		
FACULTAD:	**Ingeniería**		
CARRERA:	**Ingeniería en Sistemas Computacionales**		
TITULO OBTENIDO:	**Ingeniero en Sistemas Computacionales**		
FECHA DE PUBLICACIÓN:	**21 de septiembre del 2022**	**No. DE PÁGINAS:**	**109**
ÁREAS TEMÁTICAS:	**Minería de datos, Machine Learning**		
PALABRAS CLAVES/ KEYWORDS:	**Algoritmos supervisados, procesamiento del lenguaje natural, modelo de minería de datos**		

RESUMEN/ABSTRACT:

Dentro del presente proyecto de investigación se encuentra la creación de un modelo de minería de datos que tenga la capacidad de realizar un análisis de sentimientos basándose en datos acerca de la banca privada ecuatoriana el cual tendrá poder predictivo para asignar los valores de "Positivo", "Negativo" o "Neutro" hacia los comentarios que sean procesados por este. Esto se hará haciendo uso de herramientas de programación con las que se pueda realizar una limpieza de datos, procesamiento del lenguaje natural, y por último en el que se puedan aplicar algoritmos de aprendizaje de máquina para que se realice el entrenamiento del modelo.

Esto es con la finalidad de poder tener una medición de los niveles de satisfacción de los usuarios de la banca privada ecuatoriana de manera automática y que cuyos resultados puedan ser luego visualizados dentro de una herramienta de visualización de datos. Los objetivos del presente proyecto de investigación abarcan la creación del conjunto de datos que será utilizado, el diseño del modelo de minería de datos y la creación de este.

El proyecto finaliza con la comparación de los rendimientos de los diferentes algoritmos de clasificación de aprendizaje de máquina, la selección del mejor de estos para el desarrollo del de modelo y la presentación de los resultados de la capacidad predictiva del modelo desarrollado.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:**+593-963-770730	E-mail: **mielesroy4c@gmail.com**
CONTACTO CON LA INSTITUCIÓN (C00RDINADOR DEL PROCESO UTE)::	**Toala Quimí, Edison José**	
	Teléfono: +593-990-976776	
	E-mail: edison.toala@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA	
Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	