UNIVERSIDAD NACIONAL DE CHIMBORAZO
FACULTAD DE INGENIERÍA
CARRERA DE INGENIERÍA EN SISTEMAS Y COMPUTACIÓN

"ANÁLISIS DE SENTIMIENTOS EN LA RED SOCIAL TWITTER MEDIANTE EL PROCESAMIENTO DE LENGUAJE NATURAL"

Trabajo de titulación para optar al título de Ingeniero en Sistemas y Computación

Autor:
Maldonado Ramones, Erik Stalyn

Tutor:
Msc, Diego Reina H

Riobamba, Ecuador. 2022

DECLARATORIA DE AUTORÍA

DECLARATORIA DE AUTORÍA

Yo, Erik Stalyn Maldonado Ramones con cédula de ciudadanía 1400833297, autor del trabajo de investigación titulado: "Análisis de sentimientos en la red social Twitter mediante el procesamiento de Lenguaje Natural", certifico que la producción, ideas, opiniones, criterios, contenidos y conclusiones expuestas son de mi exclusiva responsabilidad.

Asimismo, cedo a la Universidad Nacional de Chimborazo, en forma no exclusiva, los derechos para su uso, comunicación pública, distribución, divulgación y/o reproducción total o parcial, por medio físico o digital; en esta cesión se entiende que el cesionario no podrá obtener beneficios económicos. La posible reclamación de terceros respecto de los derechos de autor (a) de la obra referida, será de mi entera responsabilidad; librando a la Universidad Nacional de Chimborazo de posibles obligaciones.

En Riobamba, Diciembre del 2022.

.....

Erik Stalyn Maldonado Ramones
CI. 1400833297.

DICTAMEN FAVORABLE DEL TUTOR Y MIEMBROS DE TRIBUNAL;

Quienes suscribimos, catedráticos designados Tutor y Miembros del Tribunal de Grado para la evaluación del trabajo de investigación ANALISIS DE SENTIMIENTOS EN LA RED SOCIAL TWITTER MEDIANTE EL PROCESAMIENTO DE LENGUAJE NATURAL, presentado por Maldonado Ramones Erik Stalyn con cédula de identidad número 1400833297, certificamos que recomendamos la APROBACIÓN de este con fines de titulación. Previamente se ha asesorado durante el desarrollo, revisado y evaluado el trabajo de investigación escrito y escuchada la sustentación por parte de su autor; no teniendo más nada que observar.

De conformidad a la normativa aplicable firmamos, en Riobamba 02 de diciembre del 2022.

Mgs. Milton Paúl López Ramos
PRESIDENTE DEL TRIBUNAL DE GRADO
Firma

PhD. Ximena Quintana López
MIEMBRO DEL TRIBUNAL DE GRADO
Firma

Mgs. Jorge Edwin Delgado Altamirano
MIEMBRO DEL TRIBUNAL DE GRADO
Firma

Mgs. Diego Marcelo Reina Haro
TUTOR
Firma

CERTIFICADO DE LOS MIEMBROS DEL TRIBUNAL

Quienes suscribimos, catedráticos designados Miembros del Tribunal de Grado para la evaluación del trabajo de investigación ANALISIS DE SENTIMIENTOS DE LA RED SOCIAL TWITTER MEDIANTE EL PROCESAMIENTO DE LENGUAJE NATURAL, presentado por Maldonado Ramones Erik Stalyn, con cédula de identidad número 1400833297, bajo la tutoría de MsC. Diego Marcelo Reina Haro; certificamos que recomendamos la APROBACIÓN de este con fines de titulación. Previamente se ha evaluado el trabajo de investigación y escuchada la sustentación por parte de su autor; no teniendo más nada que observar.

De conformidad a la normativa aplicable firmamos, en Riobamba 02 de diciembre del 2022.

Mgs. Milton Paúl López Ramos
PRESIDENTE DEL TRIBUNAL DE GRADO

Firma

PhD. Ximena Quinlano López
MIEMBRO DEL TRIBUNAL DE GRADO

Firma

Mgs. Jorge Edwin Delgado Altamirano
MIEMBRO DEL TRIBUNAL DE GRADO

Firma

C E R T I F I C A C I Ó N

Que, **MALDONADO RAMONES ERIK STALYN** con CC: **1400833297**, estudiante de la Carrera **INGENIERÍA DE SISTEMAS Y COMPUTACIÓN, NO VIGENTE,** Facultad de **INGENIERÍA**; han trabajado bajo mi tutoría el trabajo de investigación titulado " **ANÁLISIS DE SENTIMIENTOS EN LA RED SOCIAL TWITTER MEDIANTE EL PROCESAMIENTO DE LENGUAJE NATURAL**", cumpliendo con el 2%, de acuerdo al reporte del sistema Anti plagio **URKUND**, porcentaje aceptado de acuerdo a la reglamentación institucional, por consiguiente autorizo continuar con el proceso.

Riobamba, 21 de noviembre de 2022

Firmado electrónicamente por:
DIEGO
MARCELO

MsC. Diego Marcelo Reina Haro
TUTOR

DEDICATORIA

Este trabajo de investigación está dedicado a mis padres Magno Maldonado Maldonado, Rosa Ramones Córdova y a mi hermano Joshua Maldonado Ramones que han sido mi apoyo incondicional, que con su amor, consejos y enseñanza fueron mi propulsor para cumplir esta meta.

Erik Maldonado Ramones

AGRADECIMIENTO

Agradezco a todos los docentes que intervinieron e impartieron sus conocimientos en todo el trayecto de la carrera. Un reconocimiento para mi tutor Ing. Diego Reina que fue mi consejero, sustento en todo el trabajo de investigación y me asesoro hasta la culminación con éxito. También agradezco a mis colaboradoras Ing. Ximena Quintana e Ing. Jorge Delgado que me ayudaron en todo este proceso.

Erik Maldonado Ramones.

INDICE GENERAL

INDICE DE TABLAS

INDICE DE FIGURAS

RESUMEN

Las organizaciones han empezado a utilizar la minería de sentimientos, por considerar que juega un papel importante en la toma de decisiones y estrategias de mercado. Por su parte las tecnologías evolucionan muy rápido y el procesamiento de lenguaje natural y aprendizaje automático contribuyen a este cambio, que hoy permite que las máquinas puedan entender el lenguaje utilizado por los seres humanos.

Las redes sociales han evolucionado de tal manera que las plataformas son capaces de permitir a los usuarios socializar, localizar a miembros de la red y formar listas de amigos; una de las redes más importantes como Twitter permite enviar mensajes de opinión cortos llamados tweets sobre cualquier suceso de actualidad, logrando convertirse en un medio informativo para la sociedad. Por tal motivo el presente trabajo de investigación esta direccionado al estudio del análisis de sentimientos en la red social Twitter que pretende descubrir las emociones que se ocultan detrás de un escrito, las cuales pueden ser positivas, negativas o neutrales.

El análisis de sentimiento es una herramienta de donde se pueden extraer datos muy valiosos como por ejemplo para: una campaña electoral, una organización, o estudios de impacto, etc. Los datos que se obtienen a partir de este análisis de sentimientos van a permitir comprender el mercado, evaluar sus tendencias, e incluso realizar predicciones financieras.

De esta manera la propuesta resultante se direcciona a presentar un analizador de sentimientos que utilice algoritmos centrados en definir opiniones o actitudes utilizando el paquete TextBlob en Python, para una polaridad dentro de un rango comprendido entre -1.0 y 1.0 que representan una evaluación negativa o positiva, y si es igual a 0 una evaluación neutral. Útil para cualquier tema de búsqueda.

PALABRAS CLAVES Twitter, análisis de sentimientos, procesamiento de lenguaje natural, Python.

ABSTRACT

ABSTRACT

Organizations have started to use sentiment mining, considering that it plays an important role in decision making and market strategies. Technologies are evolving very fast and natural language processing and machine learning are contributing to this change, which now allows machines to understand the language used by humans.

Social networks have evolved in such a way that the platforms are able to allow users to socialize, locate network members and form lists of friends; one of the most important networks such as Twitter allows sending short opinion messages called tweets on any current event, becoming an informative medium for society. For this reason the present research work is directed to the study of sentiment analysis in the social network Twitter that aims to discover the emotions that are hidden behind a writing, which can be positive, negative or neutral.

Sentiment analysis is a tool from which valuable data can be extracted, for example for: an electoral campaign, an organization, or impact studies, etc. The data obtained from this sentiment analysis will allow us to understand the market, evaluate its trends, and even make financial predictions.

Thus, the resulting proposal is directed to present a sentiment analyzer that uses algorithms focused on defining opinions or attitudes using the TextBlob package in Python, for a polarity within a range between -1.0 and 1.0 representing a negative or positive evaluation, and if equal to 0 a neutral evaluation. Useful for any search topic.

KEYWORDS Twitter, sentiment analysis, natural language processing, Python.

INTRODUCCIÓN

"Las Redes Sociales se definen como el conjunto delimitado de individuos, grupos, organizaciones, comunidades, sociedades globales, etc.- vinculados entre sí por un conjunto de relaciones sociales" (Colina, 1996). Consideradas hoy en día el pilar de la sociedad, ya que influyen no solamente en las conexiones sociales sino también en la mercadotecnia, el conocimiento y la educación; se requiere ser más precisos en la búsqueda de objetivos para llegar al "cliente" ideal, la información que se puede obtener a partir de redes sociales es muy rica en términos de: análisis de datos, análisis de sentimientos y tendencias; por lo que cada red social ha sido creada para determinado objetivo. En el caso específico de la red social Twitter, el principal objetivo es proporcionar a los usuarios la capacidad de transmitir de forma corta información relevante a nivel mundial.

De acuerdo con lo manifestado por la revista informática Karma Pulse (KarmaPulse, 2019), "el impacto de las acciones de los usuarios en redes sociales puede ser medido a través del análisis de sentimientos", el cual es considerado como una herramienta que se aplica principalmente para obtener indicadores de percepción subjetivos, de una forma más frecuente y a un bajo costo. Los datos que entrega esta herramienta es un tipo de información no estructurada, la cual se convierte a través de la analítica de texto en un lenguaje que el computador puede entender, obteniendo información que en lo posterior será analizada a través de modelos estadísticos, descriptivos etc.

Muchas compañías e instituciones tienen presencia online ya sea en páginas web, blocks, Facebook y Twitter; donde los miembros de su comunidad generan opinión que es necesario entender y analizar de forma aglomerada, para determinar la causa de los cambios de opinión a partir de los cambios de sentimientos; identificando comentarios amenazantes, temas de interés, actores generadores de opinión etc.

El análisis de sentimiento en la red social Twitter es actualmente utilizado por Naciones Unidas denominado Global Post, el cual utiliza Big Data a fin de resolver problemas de desarrollo económico y políticas públicas. A través de un laboratorio ubicado New York, donde se utilizó todas las conversaciones de redes sociales para analizar cómo ha sido la opinión de las personas acerca de los objetivos de desarrollo sostenible; encontrando sentimientos, positivos, negativos y neutrales. Según lo señala el Programa de las Naciones Unidas para el Desarrollo (PNUD) en su página oficial.

El presente trabajo se centra en el análisis de sentimiento en la red social Twitter; donde se hacen una serie de tweets, los cuales a través de la analítica de texto podrán generar información muy valiosa. El proceso para realizar el análisis de sentimiento empieza por recolectar los tweets utilizando aprendizaje automático PNL (procesamiento de lenguaje natural) con Python (lenguaje de programación de alto nivel) a partir de la API pública de Twitter, para posteriormente filtrar y limpiar la

información, lo que va a permitir realizar el análisis de sentimiento en sí; a fin de poder clasificar los tweets según su sentimiento: negativo, positivo o neutro.
Así la propuesta resultante se direcciona al procesamiento de lenguaje natural, con la técnica de análisis de sentimientos en la red social Twitter; con finalidad de sustituir a los métodos intensivos de mano de obra como la elaboración de las encuestas de opinión.
En el Capítulo I, se presenta el tema de la investigación, denominado: "Procesamiento de lenguaje natural – Técnica de análisis de sentimientos en la red social Twitter''. En relación con el tema se plantea la contextualización del problema, el árbol del problema, la justificación sobre la investigación, se formula el interrogante y se plantean el objetivo general y los objetivos específicos.

En Capítulo II, denominado marco teórico, se plasman los antecedentes de la Investigación y se menciona la fundamentación filosófica, donde se trata sobre el enfoque de la investigación, además se señalan las categorías fundamentales entorno a la variable independiente y dependiente.

El Capítulo III, trata sobre la metodología, donde se establecen las modalidades de investigación, la población, la operacionalización de variables y se presenta cada una de las fases para el desarrollo de la propuesta, a través de la metodología de procesamiento de datos.

En el Capítulo IV, denominado análisis discusión de resultados. Se presentan los resultados finales del análisis de sentimientos y como las opiniones vertidas por los usuarios pueden ser traducidas en positivas, negativas o neutras.

En el Capítulo V, denominado conclusiones y recomendaciones, se realizan las conclusiones y recomendaciones respectivas en torno al tema de investigación.

CAPÍTULO I
EL PROBLEMA

1.1 Tema

"Análisis de sentimientos en la red social Twitter mediante el procesamiento de Lenguaje Natural"

1.2 Planteamiento del Problema

El repentino confinamiento a causa de la pandemia COVID-2019 elevo considerablemente el número de usuarios de internet, se podía pensar que una vez levantadas las restricciones por el coronavirus el número de usuarios disminuirá, sin embargo, los usuarios en el internet siguen creciendo incluso más rápido que antes; es así que para el año 2022 existen al menos 5.000 millones de usuarios de internet en el mundo. Cifra que tuvo un aumento considerable por los fuertes cambios de hábitos respecto al consumo, estudios, trabajo y en especial de las relaciones interpersonales (Galeano, 2022).

Por su parte los gobiernos, instituciones y comercios buscan la mejor forma de interactuar con el usuario, de saber que opinan y que reacción tienen ante cualquier acción implementada. La gran cantidad de publicaciones que produce Twitter ha sido vista como una oportunidad para interpretar esta información y generar un análisis de las opiniones ocultas en textos originales.

Las Naciones Unidas para el Desarrollo es un organismo internacional pionero en el uso de análisis de sentimientos en redes sociales, ya que tienen como objetivo conocer la aceptabilidad de los habitantes de una región respecto a las políticas de desarrollo implementadas, según lo expreso uno de los consultores del Programa de las Naciones Unidas para el Desarrollo, el Eco. Julio Martínez Gordillo durante el Webinar presentado por la empresa de tecnología Bd Guidance.

Para el año 2018, la revista Enfoque de la Universidad Técnica Equinoccial UTE, presento un estudio sobre la Influencia de redes sociales en el análisis de sentimiento aplicado a la situación política en Ecuador, a fin de conocer la opinión de los electores, el nivel de aceptación de los candidatos y los resultados electorales. Se llego a la conclusión que no fue suficiente la opinión de usuarios para saber los resultados de las elecciones, sino que habría que considerar otro tipo de variables.

En virtud de lo mencionado anteriormente, se puede decir que los usuarios interactúan e intercambian opiniones utilizando las redes sociales; el presente trabajo de investigación busca realizar un análisis de sentimiento sobre las opiniones de los usuarios, a fin de poder determinar si se trata de sentimientos positivos, negativos o neutros.

1.2.1 Análisis Crítico

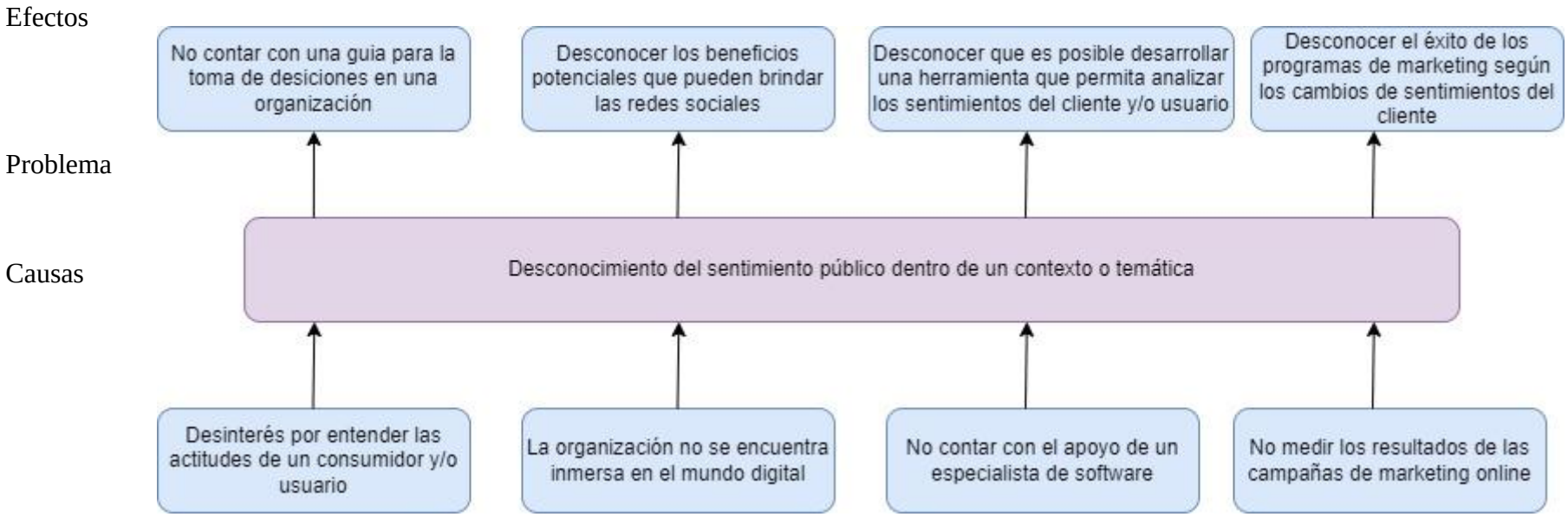


Figura 1: Árbol de Problemas
Elaborado por: Erik Maldonado

El problema que enfrenta la mayor parte de organizaciones e instituciones es el desconocimiento del sentimiento público dentro de un contexto o temática, lo cual ha generado una serie de efectos que se describen a continuación:

El desinterés por entender las actitudes de los consumidores y/o usuarios, provoca que no se tenga una guía para la toma adecuada de decisiones, ya que se desconoce cómo están reaccionando los consumidores frente a una marca, producto o servicio lo cual se traduce en falta de políticas o decisiones acertadas que se puedan tomar a tiempo.

El hecho de que una organización no se encuentre inmersa en el mundo digital es la causa fundamental para que la misma desconozca los beneficios potenciales que pueden brindar las redes sociales y la cantidad de información valiosa fácilmente disponible que puede aportar con el crecimiento de la empresa, lo cual se traduce en procesos lentos y costosos además de márgenes de rentabilidad poco satisfactorios.

El no contar con el apoyo de un especialista de software, hace que la organización no tenga conocimiento de que es posible disponer de una herramienta que permita analizar los sentimientos que el cliente esconde detrás de un comentario u opinión; lo cual trae consigo que la organización este un paso atrás de la competencia.

El no poder medir los resultados que genera una campaña de marketing online genera una incertidumbre en la organización al desconocer si el resultado obtenido fue un éxito o un fracaso, lo cual se traduce en la falta de cumplimiento de los objetivos empresariales. Esto es debido a que las organizaciones no disponen de una de las herramientas más versátiles del momento como es el "análisis de sentimientos"; ya que de acuerdo con el análisis realizado se puede decir que la opinión del cliente es capaz de hacer o romper el éxito de una marca, y la decisión de monitorearla puede significar la diferencia entre una corrección a tiempo o una oportunidad perdida.

1.2.2 Formulación del Problema

¿Cómo la técnica de análisis de sentimientos permitirá clasificar de manera eficiente los sentimientos encontrados en los tweets?

1.2.3 Interrogantes

¿Cómo funciona el procesamiento del lenguaje natural para el análisis de sentimientos?

¿En qué áreas es posible aplicar el análisis de sentimientos?

¿Será necesario crear un diccionario lexicón de sentimientos idóneo para analizar los sentimientos de los usuarios de Twitter en el Ecuador?

1.2.4 Delimitación

Límite de Contenido

Campo: Ingeniería

Área: Sistemas y computación

Aspecto: Minería de textos

Delimitación Espacial

La presente investigación se realizará a través de la minería de textos de la red social Twitter, considerando los límites técnicos actuales permitidos por Twitter que corresponde a 1000 tweets diarios.

Delimitación Temporal

En el año 2022 a partir del mes de agosto se realizó la presente investigación hasta octubre del 2022.

Unidades de Observación

La investigación será aplicada de preferencia en el área de marketing y estudios sociales.

1.3 JUSTIFICACIÓN

El procesamiento de lenguaje natural permite a las empresas interactuar con sus clientes utilizando plataformas de comercio y analizando sus necesidades a partir de un análisis de datos a gran escala; además es muy útil en el área de negocios ya que agilita los procesos y reduce los costos debido a que obtiene información real y procesable.

El análisis de sentimiento y la comprensión del lenguaje natural permiten interpretar lo que las personas sienten a través de su lenguaje. Para una organización el análisis de sentimientos permite entrar en la mente de sus clientes y observar cómo se sienten. El desafío radica entonces en clasificar el gran volumen de datos de los clientes y determinar la intención del mensaje.

La clasificación de opiniones ingresa en una amplia categoría de tareas en el cual se proporciona una frase o lista de frases donde su clasificador indica que tipo de opinión hay detrás de un comentario, que puede ser positivo, negativo o neutral; todas estas actividades serán posibles gracias al uso de la NLKT (Kit de herramientas de lenguaje natural) en Python.

El desarrollo de la presente investigación permitirá a las empresas beneficiarse de las ventajas que esta ofrece como: estrategias de marketing más perspicaces basadas en datos, comprender a los clientes, medir una campaña de marketing y el posicionamiento de un producto o servicio en el mercado; lo cual traerá consigo un mejoramiento en la rentabilidad de la organización dinamizando de esta manera la economía nacional.

1.4 OBJETIVOS

1.4.1 Objetivo General

Implementar un análisis de sentimientos en la red social Twitter mediante el procesamiento de lenguaje natural.

1.4.2 Objetivos Específicos

1. Analizar el procesamiento de lenguaje natural y la técnica de análisis de sentimientos, usando Python.
2. Desarrollar una solución informática que analice los sentimientos encontrados en la red social Twitter.
3. Evaluar la solución informática de análisis de sentimiento, a través de la aplicación de métricas de exactitud, precisión y sensibilidad.

CAPÍTULO II
MARCO TEÓRICO

2.1 Antecedentes

Una vez investigados temas similares al de la presente investigación, se describen a continuación los aspectos fundamentales de cada trabajo:

Trabajo investigativo de (Georgios, 2013) con el tema denominado "Sentiment Analysis of Twitter posts"; concluye que: El algoritmo desarrollado para analizar los sentimientos capaces de dar una valoración a la opinión en cualquier idioma, no utiliza ningún léxico, solamente los perfiles de frecuencia N-Grams que dan al algoritmo una entrada.

Artículo científico de (Patel, 2017) titulado "Sentiment Analysis on Twitter Data Using Machine Learning", donde concluye: A fin de realizar el procesamiento de lenguaje natural para los tweets se inicia con la tokenización para continuar con la lematización, en este paso se obtienen las palabras base de palabras que contienen la misma raíz, luego se evalúa y clasifica el lenguaje en positivo, negativo o neutro utilizando los recursos léxicos de SentiWordNet el cual asigna una puntuación a cada término. En la plataforma de Twitter no solo se analiza la opinión del usuario, sino que permite que un negocio conozca los comentarios sobre un evento, marca o promoción.

Investigación de (Lee, 2021) denominado "Análisis de sentimiento y modelado de temas en tweets sobre educación en línea durante COVID-19". Al respecto concluye: El estudio contiene alrededor de los 17155 tweets, se ha utilizado la herramienta TextBlob a fin de analizar la polaridad y subjetividad de los tweets. Para la extracción de características se utilizó la técnica TF-IDF (frecuencia de término-frecuencia de documento inversa); el modelo ha sido evaluado a partir de varias métricas importantes como exactitud, precisión, recuperación y puntaje. Este modelo se utilizó con el fin de encontrar problemas asociados con el aprendizaje electrónico, las discapacidades de los niños para comprender la educación en línea y las redes eficientes rezagadas para la educación.

Trabajo investigativo de (Montesinos, 2014) titulado "Análisis de sentimientos y predicción de eventos en Twitter", tiene como objetivo: Estudiar los métodos más importantes utilizados en la literatura para desarrollar un análisis de sentimientos ; además concluye que: a) Es importante el diseño de un diccionario con palabras positivas y negativas con un puntaje asociado ; b) El algoritmo tiene un 74% de acierto, para identificar mensajes positivos y negativos, y de un 60% si identifica mensajes positivos negativos y neutros; además tuvo un acierto 51% sobre 49% acerca del candidato que ganaría las elecciones, con un error solo del 2% respecto a los resultados electorales .

(Garcés, 2019). En su trabajo investigativo denominado "Análisis de sentimientos en redes sociales orientado a la percepción de la calidad de servicios de internet, redes móviles, tv cable y electricidad" concluye que: La investigación está proyectada en el análisis de sentimientos hacia un análisis de texto en lenguaje natural que tiene que ver con contenido, concepto y contexto; el cual se encuentra acompañado de una gran cantidad de data, se evaluó el modelo a través de la métrica Accuracy. Al probar el modelo LogisticRegression obtuvo un 88% de accuracy, etiquetando 1745 tweets negativos y 1075 no negativos.

Investigación de (Romero R. , 2021) con el tema denominado “Análisis de sentimientos en Twitter para descubrir contenido Xenófobo hacia los Inmigrantes Venezolanos en Ecuador”; concluye que: Se utilizó la metodología de Bárbara Kitchenham, a través del planteamiento de cuatro preguntas de investigación y utilizando el lenguaje de programación Python, la regresión logística Naive Bayes y Máquinas de Soporte Vectorial como algoritmos de clasificación, se concluye que para detectar la xenofobia en los tweets se puede utilizar el modelo creado por Davison T., el cual clasifico un conjunto de tweets a través de crowdsourcing que incluía mensajes de odio y ofensivos.

(Alex Sanchez, 2020). En su artículo investigativo, titulado “Modelo para el análisis de sentimientos del banco de encuestas con preguntas sobre coronavirus de la OMS empleando principios de minería de textos”. Interpreta que: Aplicando fundamentos de minería de textos mediante la plataforma Open Source de minería de datos y utilizando la herramienta VADER (Valence Aware Dictionary and Sentiment Reasoner), se muestra que las emociones neutras alcanzaron los valores más altos en comparación a las emociones positivas y negativas, siendo el resultado de estas dos últimas casi similares, las mismas que se relacionan con información sobre economía familiar, creencias sobre el origen del virus y la eficacia de los gobiernos ante la pandemia.

Los trabajos citados fueron una guía para el desarrollo del marco teórico, operacionalización de variables, tipo de metodología de estudio y para la interpretación de los resultados obtenidos.

2.2 FUNDAMENTACIÓN FILOSÓFICA

El enfoque del presente proyecto de investigación sienta sus bases en el análisis de información bibliográfica, como una herramienta metodológica básica, la cual ayudo a obtener resultados que se utilizaron para responder al interrogante planteado.

2.3 CATEGORÍAS FUNDAMENTALES

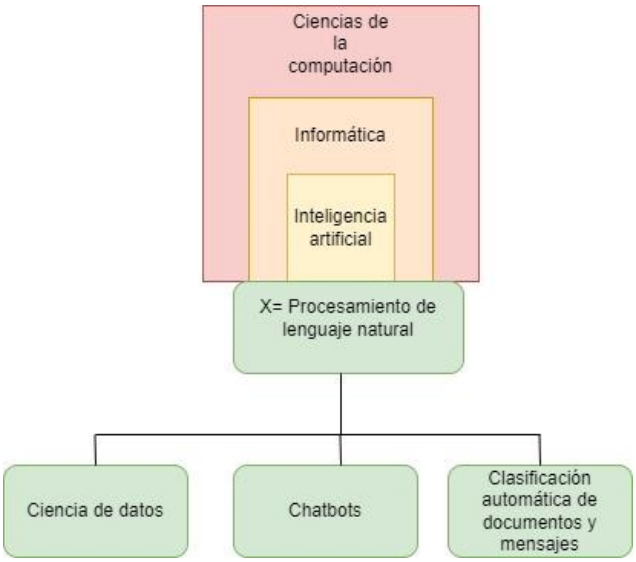


Figura 2: Categoría fundamental – Variable Independiente
Elaborado por: Erik Maldonado

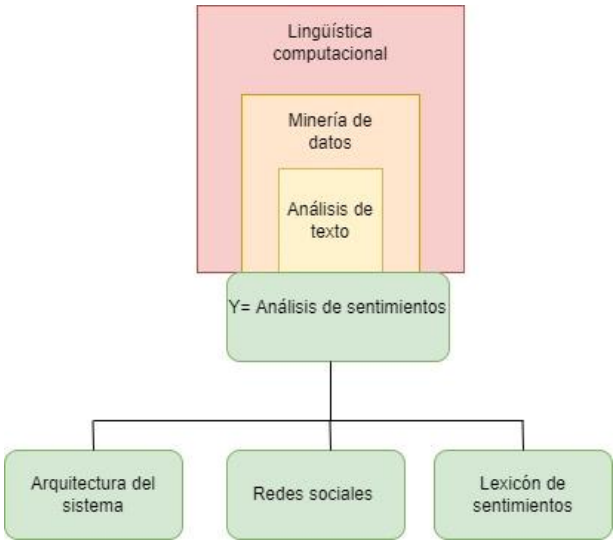


Figura 3: Categoría fundamental – Variable Dependiente

Elaborado por: Erik Maldonado

Desarrollo de categorías, Variable Independiente

2.3.1 Ciencias de la computación

Las ciencias de la computación son un conjunto de disciplinas que se ocupan de los lenguajes de programación y fundamentos matemáticos propios de esta ciencia. Está compuesta por una serie de ramas, entre ellas: la informática, la inteligencia artificial, la bioinformática entre muchas otras. Las ciencias de la computación son aplicables a la física y biología a través de simulaciones, está presente en la industria del cine, en la actividad bursátil, incluso en disciplinas humanísticas (Romero A. , 2020)

Según (Denning, 2021), las ciencias de la computación permiten responder al interrogante, ¿Qué puede ser eficientemente automatizado?, este planteamiento ha sido respondido a través del constante desarrollo de algoritmos y lenguajes de programación; lo cual ha permitido el análisis masivo de la data útil en múltiples campos.

2.3.1.1 Informática

La informática es un tema amplio que contiene muchas aplicaciones, se diseñan software, se resuelven problemas informáticos y se desarrollan múltiples formas de usar la tecnología. Una de esas formas es aplicar el lenguaje de programación capaz de operar un software específico y las acciones que debe tomar el mismo bajo una serie de circunstancias. (Bitesize, 2022).

Un sistema informático lleva a cado tres tareas fundamentales como: la captación, el tratamiento y la transmisión de la información, haciendo que de esta manera converjan un conjunto de disciplinas a fin de ser aplicadas en diversas áreas y en todo tipo de escenarios, como en la medicina, educación ingeniería, negocios y otros. A partir de allí, la informática tiene múltiples funciones entre ellas: procesamiento de imágenes, criptografía, seguridad de sistemas informáticos, minería de datos, teleinformática, inteligencia artificial, robótica entre otras (Sevilla, 2020).

2.3.1.2 Inteligencia Artificial

La inteligencia artificial es el esfuerzo por simular la inteligencia humana en las máquinas, según (Russell, 2019) la IA recibe percepciones del entorno y realiza acciones con la ayuda de algoritmos de PLN. Existen cuatro tipos de inteligencia artificial:

1. Máquinas reactivas (principios básicos de IA)
2. Memoria limitada (la IA limitada almacena datos y predicciones anteriores al recopilar información, utiliza tres modelos: aprendizaje por refuerzo, memoria a corto plazo, y redes antagónicas)
3. Teoría de la mente (la IA trata de comprender como se sienten los humanos)
4. Conciencia de sí mismo (la IA algún momento tendrá conciencia de sí misma).

2.3.1.3 Procesamiento de lenguaje natural PLN

El Procesamiento de lenguaje natural (PLN) forma parte de la informática, de la ingeniería de la información y la inteligencia artificial y facilita la conversión de texto o voz en información estructurada a través de código (Peláez, 2022). De esta manera la inteligencia artificial está más cerca de los humanos, ya que empieza a entender sentimientos y emociones, los sistemas de inteligencia artificial más conocidos son el Machine Learning y el Deep Learning.

Recientes avances en Machine Learning (ML) han logrado en los computadores realicen cosas útiles con lenguaje natural. Deep Learning cuenta con identificación de imágenes y es la nueva herramienta de traducción automática. Todo esto facilita la comprensión y realización de cálculos en grandes bloques de texto sin esfuerzo manual (Bolaños, 2020).

Con la ayuda de Machine Learning (ML) y Python la tarea de que el computador entienda el lenguaje humano es mucho más fácil, convirtiéndose en la mejor opción para PLN. Los pasos para el proceso de análisis de datos son: preparación de documentos en un formato adecuado como texto etc., tokenización de los datos, negación y detección de datos.

2.3.1.3.1 Ciencia de Datos

La ciencia de datos hace referencia a un área relacionada con la recolección, preparación y administración de grandes cantidades de información, muchos de estos datos no son numéricos ni estructurados; los datos constituyen la base de la innovación sin embargo su valor proviene de la información que los expertos puedan obtener y luego utilizar (Stedman, ComputerWeekly, 2022).

Entender los sentimientos de los usuarios en la red es un aspecto fundamental para las organizaciones, al analizar de forma automática los comentarios de los clientes desde reacciones hasta discusiones en las redes sociales, a fin de adaptar los productos o servicios y abordar las necesidades de los clientes. Expertos en el teman tratan de codificar toda la cadena de conocimiento, capturar la gran cantidad de datos digitales, clasificarlos, procesarlos a través de un algoritmo y tomar la decisión correcta en tiempo real.

La mayor parte de empresas y organizaciones tratan de aprovechar el poder de la información para reaccionar de forma oportuna frente a sus clientes, situación que ha sido aprovechada por organizaciones tecnológicas para ofrecer soluciones de análisis de sentimiento basadas en IA (Inteligencia artificial). La ciencia de datos aplicada correctamente puede ofrecer

información muy valiosa para organizaciones e instituciones que buscan un posicionamiento en el mercado.

Existen tres categorías diferentes para los datos, el primero analista de datos entre la comunicación de datos y las estadísticas, el segundo la ingeniería de datos entre el ingeniero de software y las matemáticas, y el tercero el científico de datos que analiza cada campo (Davenport, 2012).

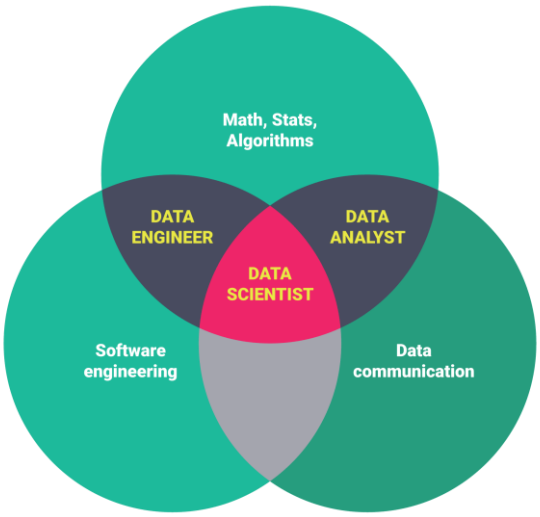


Figura 4: Interacción de la Ciencia de datos
Fuente: Ciencia de datos

2.3.1.3.2 Chatbot

Un chatbot es un bot que utiliza el procesamiento de lenguaje natural con el fin de identificar la intención que tiene un usuario para brindarle su apoyo, un chatbot tiene la capacidad de aprender y desarrollar frases; funcionan interpretando información como: declaraciones, intención, demografía, contexto, y sesión. El objetivo es mantener un dialogo interactivo con los clientes a fin de entender mejor sus necesidades. El uso de chatbot ha crecido en un 92% a nivel mundial y son usados para ofrecer información sobre requisitos, atención al cliente en aerolíneas, y para entregar información del estado de cuenta a los clientes (Computerworld, 2021).

2.3.1.3.3 Clasificación automática de documentos y mensajes

La clasificación automática de documentos y mensajes es aplicable en todos los sectores que manejan mucha información, siendo posible realizar dos tipos de clasificación: multi- clase y multi- etiqueta. Antes de realizar una clasificación, se extraen las características, a través de un proceso de reducción y codificación.

Si la intención es analizar temas sobre una marca en redes sociales es posible realizar la clasificación respecto a los comentarios y enriquecer la misma a través del análisis de sentimiento ((Itelligent, 2019).

Desarrollo de categorías, Variable Dependiente

2.3.2 Lingüística computacional

Es una disciplina de la ingeniería que se ocupa del lenguaje hablado y escrito desde una perspectiva computacional que combina la lingüística con la informática y la IA, es utilizada en herramientas como la traducción automática instantánea, sistemas de reconocimiento de voz, sintetizadores de texto a voz, editores de texto y material de instrucción de idiomas.

Por otra parte, la lingüística computacional y el procesamiento del lenguaje natural utilizan las mismas herramientas como el aprendizaje automático y la inteligencia artificial. Su principal objetivo es mejorar la relación entre las computadoras y el lenguaje básico; lo cual implica construir artefactos que se usen para procesar y producir lenguaje escrito y hablado en formatos estructurados y no estructurados (Gillis, 2022).

2.3.2.1 Minería de datos

De acuerdo con lo manifestado por (Stedman, Procesamiento de datos, 2021), la minería de datos se considera un proceso para clasificar grandes conjuntos de datos a fin de poder identificar patrones y relaciones que resuelvan ciertos problemas, las técnicas y herramientas que se utiliza en minería de datos puede ayudar a las empresas a predecir tendencias futuras y tomar decisiones más acertadas. La minería de datos forma parte de la ciencia de datos y es un componente de la iniciativa analítica en las organizaciones, ya que genera información útil en la inteligencia empresarial y para análisis de datos históricos.

El proceso de minería de datos consta de cuatro etapas:

1. Recopilación de datos
2. Preparación de datos
3. Minería de datos
4. Análisis e interpretación de datos

Las técnicas de minería de datos más conocidas son:

1. Minería de reglas de asociación
2. Clasificación
3. Agrupación
4. Regresión
5. Análisis de secuencias y caminos
6. Redes neuronales

2.3.2.2 Análisis de textos

El análisis de texto es un aliado de las empresas al momento de analizar grandes cantidades de datos que se basan en texto de manera escalable, coherente e imparcial, considerado como un recurso muy valioso donde las organizaciones pueden hacer uso de este a conveniencia. El análisis de texto tiene como objetivo obtener información de calidad del texto o las palabras en sí, sin considerar su semántica.

2.3.2.3 Análisis de sentimientos

El análisis de sentimiento nace en la psicología y sociología, ambas tratan de evaluar las emociones, relaciones, opiniones y comportamientos de las personas, sin embargo, los expertos en software lo hacen a través de datos, cuyo objetivo es comprender la opinión del

usuario frente a determinado tema. Es decir que a través de un proceso computacional se desea identificar y categorizar las opiniones generadas en determinada red social.
El proceso de análisis de sentimiento hace referencia al PNL (procesamiento natural de lenguaje), y se expresa en dos categorías: polaridad y subjetividad. La medida de polaridad de los datos es positiva (>0), negativa (<0) o neutro (0). La medida de subjetividad de (0.0 a 1.0); donde 0.0 es muy objetivo y 1.0 es muy subjetivo.
Sin embargo, en el presente trabajo de investigación calcularemos solo la polaridad de sentimientos de los datos en Twitter (el formato de los tweets está en CSV). La polaridad calculada a través de Python usa la biblioteca TextBlob y el módulo Python Natural Lagunaje Tool Kit (NLTK).

2.3.2.3.1 Arquitectura del sistema

Cuando hablamos de análisis de sentimientos existen varias opciones y herramientas. Las herramientas más populares son MATLAB, Python, y Java; gracias a la gran cantidad de bibliotecas disponibles en Python los investigadores lo usan por ser considerada la opción más adecuada. El algoritmo para el análisis de sentimientos está formado por cuatro módulos, el procedimiento en cada modelo inicia con la importación de datos con pandas, seguido del uso de NLTK y TextBlob para analizar el texto del archivo CSV y así calcular la polaridad de cada texto por separado, siendo la salida un formato numérico (-1 a +1). En la presente investigación se inicia recopilando los tweets con la palabra clave requerida, posteriormente Matplotlib muestra el resultado con diferentes colores y formatos para los términos positivo, negativo y neutro (> 0, <0, =0).

- **Python**

Python es un lenguaje de programación que utiliza varios paradigmas de programación y es muy usado por los programadores debido a que su codificación es fácil y rápida, ha sido usado por grandes empresas como Google, Yahoo!, YouTube, Dropbox y la nasa.
Es un lenguaje que se caracteriza por poder manejar grandes cantidades de datos además de permitir realizar matemáticas complejas con los mismos. Python es un lenguaje libre y de código abierto, esto significa que es gratuito y está disponible para todos, está más cerca de un lenguaje humano que de un lenguaje de máquina; por sus características la tasa de error utilizando este lenguaje de programación es muy bajo (Geeks, 2022).
Python es usado en programas de web scraping, minería de datos, limpieza de datos, procesamiento de datos y modelado, estadística descriptiva e inferencial, procesamiento de lenguaje natural, aprendizaje automático e inteligencia artificial, análisis gráfico, procesamiento de imágenes, secuencia de comandos, desarrollo de juegos y desarrollo web. Sin embargo, existen algunas desventajas al usar Python, entre ellas: es poco eficiente en memoria, necesita mayor tiempo para su ejecución, no es apto para programar hardware, no es el mejor lenguaje para trabajar en aplicaciones móviles, y solo permite la ejecución de un subproceso a la vez. Python cuenta con varias bibliotecas entre ellas tenemos: Numpy, Scipy, Pandas, Matplotlib; frameworks como Theano, TensorFlow, Keras para el aprendizaje profundo (Kareem, 2020).

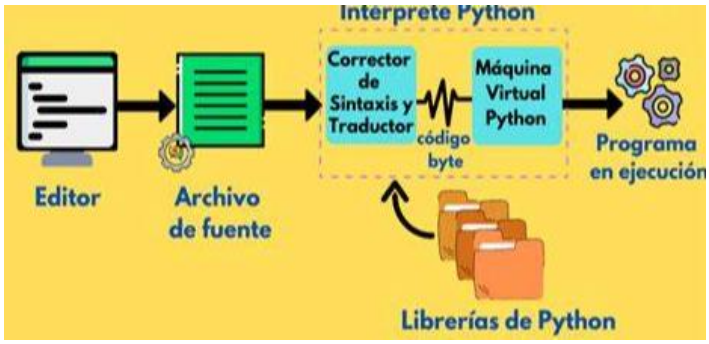


Figura 5: Funcionamiento del intérprete de Python
Fuente: Python

- **Biblioteca y módulos de Python**

Las bibliotecas hacen que Python sea más fácil y rápido, mientras que los módulos de Python son archivos en códigos Python que definen funciones, variables y clases (Kuhlman, 2009). Los módulos creados son guardados en bibliotecas de Python, un módulo puede ser utilizado en 14 proyectos diferentes al mismo tiempo facilitando de esta manera el trabajo.

- **Biblioteca TextBlob**

TextBlob es una biblioteca de Python que procesa datos textuales, proporciona una API para acceder a sus métodos y realizar tareas de PNL, es fácil de usar y no existe complicaciones en cuanto su sintaxis. TextBlob está construida sobre dos bibliotecas muy importantes como son: NLTK y pattern, lo cual permite que TextBlob combine el uso de las dos herramientas en un interfaz más simple (notebook.community, 2020).

TextBlob trabaja y juega con todo tipo de texto, admite además todo tipo de formato de texto, es uno de los módulos más importantes de Python que se usa para analizar sentimientos y clasificar los datos en positivos y negativos (Loria, 2020). Se puede decir que TextBlob es la biblioteca más importante para el análisis de sentimientos.

- **Biblioteca NLTK (Natural Language Toolkit)**

La biblioteca NLTK es un kit de lenguaje natural, es ideal para cualquier profesional y está disponible para Mac OSX, Windows y Linux, es gratuito y de código abierto, además es una guía hacia la escritura de programas con Python, facilita interfaces para más de 50 corpus y el análisis de estructura lingüística (Bird, 2009).

El presente trabajo de investigación utiliza la biblioteca TextBlob para el análisis de sentimiento que importa el módulo NLTK.

- **Librería Matplotlib**

Matplotlib es una de las librerías más populares de visualización, permite crear gráficos como: diagrama de barras, histograma, diagrama de sectores, diagrama de caja y bigotes, diagrama de violín, diagrama de dispersión o puntos, diagramas de líneas, diagramas de áreas, diagramas de contorno y mapas de color (Manual de python, 2016). Matplotlib es usado para la visualización de sentimientos, muestra además el número total de tweets positivos, negativos y neutrales.

- **Librería Pandas**

Pandas es una librería que se caracteriza por el análisis y manejo de estructura de datos, se caracteriza por: definir estructura de datos basadas en arrays, leer y escribir ficheros en formato CSV, Excel y base de datos SQL, accede a datos a través de nombres para filas y columnas, facilita métodos para combinar conjunto de datos, trabaja con series temporales y es eficiente en la ejecución de operaciones. Pandas tiene tres estructuras de datos: de una dimensión, de dos dimensiones y de tres dimensiones, que se construyen a través de arrays de la librería Numpy.

Pandas es compatible con el lenguaje R, es útil para limpiar los datos y unir o combinar los datos en filas y columnas (Bronshtein, 2017).

- **Módulo CSV**

Para el presente proyecto de investigación se utiliza el formato CSV (valor separado por comas). Los archivos CSV son muy fáciles de procesar, permite: importar y exportar datos de clientes y productos, exportar órdenes y reportes. CSV cuentan con varias funciones: csv. reader, csv.writer, csv.Dictwriter, csv. DictReader (Vaati, 2017).

- **OS- Interfaces misceláneos del sistema operativo**

Este módulo facilita el uso de funciones del sistema operativo, se usa para leer y escribir un archivo, es posible establecer rutas y crear archivos temporales.

- **Módulo Sys**

El módulo Sys se utiliza como intérprete en Python, proporciona funciones y variables que manipulan parte del entorno de ejecución en Python. Este módulo posee comandos como rastreo, mapeo, derechos de autor, borrar cache, marco actual y más.

- **Módulo Tweepy**

Tweepy es el módulo más importante en el presente trabajo de investigación, sin este módulo no sería posible recopilar los tweets de la API de Twitter, Tweepy admite claves de autenticación proporcionadas por Twitter.

- **Clasificador Naive Bayes en Python**

El clasificador naive Bayes es un algoritmo de aprendizaje usado para resolver problemas de clasificación de la data, basado en el teorema de Bayes (Garcia, 2020).

- **Módulo String**

Es uno de los módulos más utilizados en las primeras versiones de Python, conserva constantes y clases para trabajar con objetos, una de sus funciones es la concatenación que no es más que la unión de cadenas mediante el signo (+) (Rico, 2019).

- **Módulo re**

Re proporciona operaciones de coincidencia de expresiones, es decir si una cadena coincide con una expresión regular (Friedl, 2022).

2.3.2.3.2 Redes sociales

Las Redes Sociales son una tecnología que facilita el intercambio de información por medio de comunidades virtuales, brindando a los usuarios una comunicación electrónica rápida a

través del uso de computadoras, tablets o teléfonos que cuentan con aplicaciones web. Las redes sociales más grandes son Facebook, Instagram, Twitter, YouTube y Tiktok, su contenido es generado por usuarios y perfiles personalizados (Maya, 2021).

Las redes sociales brindan la oportunidad de conectarse con el resto del mundo, conocer lo que está sucediendo en tiempo real y tener acceso a un sinfín de información. Según la encuesta realizada por Pew Research Center el uso de las redes sociales se correlaciona con tener más amigos y redes más diversas, en promedio las personas invierten 2 horas y 25 minutos diarios en redes sociales.

Para el año 2021 el número de usuarios activos en redes sociales fueron en promedio 4.2 mil millones, lo cual representa el 53,6% de la población a nivel mundial (Xie, 2021).

Figura 6: Informe de Encuesta en Redes sociales
Fuente: Encuesta en redes sociales

- **Twitter**

Twitter es una de las redes sociales más grandes en el mundo, con un número aproximado de 353 millones de usuarios para el 2021, los cuales emiten opiniones, actualizan estados sobre cada tema por cada segundo en el día. En promedio cada usuario invierte 5,6 horas por mes y su público en un 32% es femenino y en un 68% masculino, según la encuesta realizada por Pew Research Center de los Estados Unidos.

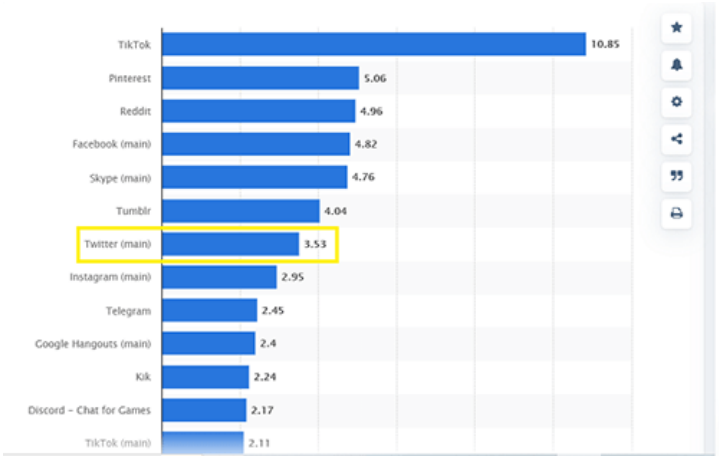


Figura 7: Informe de Encuesta en Redes sociales
Fuente: Encuesta en redes sociales

En el Ecuador Twitter cuenta con un número aproximado de 1,15 millones de usuarios, donde el 38.9% son mujeres y 61.1% son hombres.

Figura 8: Informe de Encuesta en Redes sociales, por país.
Fuente: Encuesta en redes sociales

2.3.2.3.3 Lexicón de sentimientos

El lexicón según lo manifiesta (Swann, 2004), es el vocabulario de un lenguaje que contiene todos los términos o lexemas de un lenguaje. Algunos lexicones están disponibles en internet como: SentiWordnet, NTU, Pak entre otros (Barbosa, 2015).

Un lexicón de sentimientos se puede construir con recursos disponibles en la web y el sistema de opiniones mediante un análisis de subjetividad. Entre las metodologías para construir lexicones de sentimientos tenemos:

1. Metodología manual, que consiste en anotar cada entrada léxica y asociarla con una categoría correspondiente.

2. Metodología semiautomática, la cual se apoya en la metodología manual; para ello se entrena el algoritmo con términos resultantes para luego ponerlo en marcha y así poder predecir el nivel de alineación de un nuevo término.
3. Metodología automática, Los términos se evalúan y se clasifican como positivos, negativos o neutros, para lo cual dependen del uso de un algoritmo con entradas léxicas anotadas previamente.

2.4 Señalamiento de variables

2.4.1 Variable Independiente

Como variable independiente se consideró la siguiente: Procesamiento de lenguaje natural; la cual no cambiara con otras variables que se trata de medir, es decir que actúa por si sola. De hecho, esta variable es la causante de los cambios en la variable dependiente.

2.4.2 Variable Dependiente

Como variable dependiente se consideró la siguiente: Análisis de sentimientos; ya que puede cambiar según varios factores característicos de la variable independiente (Procesamiento de lenguaje natural), como: segmentación, tokenización de palabras, lematización, etc. Es decir que el análisis de sentimiento va a depender la carga valorativa de las palabras que dan contexto a una frase, a fin de determinar si la opinión oculta en dicha frase es positiva o negativa, dependiendo el objetivo del usuario.

CAPÍTULO III
METODOLOGÍA

3.1 Enfoque

El presente trabajo de investigación tiene un enfoque Cuantitativo: porque se utilizan algoritmos de clasificación de la data y gráficos estadísticos para visualizar los resultados del estudio.

3.2 Nivel o tipo de investigación

Investigación Exploratoria: La revisión adecuada de la literatura permitió tener una aproximación al análisis de sentimientos, de allí que se realizó el planteamiento del problema y la determinación de las variables.

Investigación descriptiva: Describe el problema de la presente investigación y sus características para determinar cómo se manifiesta y el grado de afectación que produce la falta de este tipo de estudios.

3.3 Población y muestra

La población para el presente trabajo de investigación son los tweets de la red social Twitter.

Población	Unidad de análisis	Número
Comentarios de Twitter	Comentarios en la red de Twitter	1000 tweets por día (número de tweets que permite el API de Twitter)

Tabla 1: Población

Elaborado por: Erik Maldonado

3.4 Materiales y métodos

El presente proyecto se desarrolla con un razonamiento que inicia con la teoría y que se deriva a expresiones sometidas a prueba.

- Librerías

Tweepy, TextBlob, Sys, Matplotlib, Pandas, Numpy, Os, Nltk, Pycountry, Re, String,

- Lenguaje de programación

Python

- Aplicativo para desarrollar en código Python algoritmos

Google Colab

- Red social para realizar el análisis de sentimientos

Twitter

3.5 Operacionalización de las Variables

Pregunta de Investigación	Tema	Objetivos	Variables	Conceptualización	Dimensión	Indicadores
¿Cómo la técnica de análisis de sentimientos permitirá clasificar de manera eficiente los sentimientos encontrados en los tweets?	Análisis de sentimientos en la red social Twitter mediante el procesamiento de lenguaje natural	General: Implementar un análisis de sentimientos en la red social Twitter mediante el procesamiento de lenguaje natural.	Independiente Procesamiento de lenguaje natural	Es un campo del conocimiento de la inteligencia artificial que investiga la forma que se comunican las máquinas con las personas, a través del uso de lenguas naturales.	1.Plataforma microblogging 2. Estado de opinión en una plataforma	1.Número de tweets positivos 2.Número de tweets negativos 3.Número de tweets neutros
		Específicos: 1. Analizar el procesamiento de lenguaje natural, técnica de análisis de sentimientos, aplicando el código Python.	Dependiente Análisis de sentimientos	Análisis cualitativo de grandes volúmenes de datos, con beneficios exponenciales ya que son analizados con mayor profundidad.	1.Matriz de confusión para análisis de sentimientos.	1.Exactitud del modelo 2.Precisión del modelo 3.Sensibilidad del modelo

		2. Desarrollar el análisis de sentimientos, para conocer la opinión de la población a escala, respecto a cualquier tema de búsqueda. 3.Evaluar la solución informática de análisis de sentimiento, a través de la aplicación de métricas de: exactitud, precisión y sensibilidad.				

Tabla 2: Operacionalización de las variables

Elaborado por: Erik Maldonado

3.6 Metodología de Procesamiento de Datos

A fin de llevar a cabo el presente proyecto de investigación se utilizará la metodología de procesamiento de datos, la cual cuenta con las siguientes fases: extracción o recolección de tweets, filtrado y limpieza de datos, clasificación del sentimiento y representación gráfica (Agile, 2022).

Para la extracción se empleará la metodología de extracción de tweets

- Filtrado y limpieza de datos
- Extracción o recolección de tweets, para la extracción se empleará la metodología correspondiente

Twitter envía los tweets en un formato conocido como ruido en Spark, a fin de que este pueda realizar un análisis de sentimientos, como se observa en la Figura 9.

RT @geekytheory: I love early in the day, feels like i have

Figura 9: Re tweet en Twitter
Fuente: Metodología de procesamiento de datos

El caso del ejemplo es un Re tweet, que va de un usuario a otro, se puede decir entonces que la palabra RT como @geekytheory, es irrelevante; por lo que se debe crear un clean Tweet para eliminar las almohadillas hashtags, menciones, Re tweets, URL, utilizando el código clean text, como se observa en la Figura 10.

```
def clean_text(text):
    text_lc = "".join([word.lower() for word in text if word not in
string.punctuation])
    text_rc = re.sub('[0-9]+', '', text_lc)
    tokens = re.split('\W+', text_rc) #
    texto de tokenización = [ps.stem(word) for word in tokens if word
not in stopword]
   return texttw_list.head()
```

Figura 10: Código clean_text
Fuente: Metodología de procesamiento de datos

Aplicando el código clean_text a todos los tweets en el marco de datos pandas se construye las funciones que ayuden a calcular la subjetividad y polaridad de los tweets usando TextBlob, se elimina las filas vacías a través de un comando y a través de la lematización y tokenización se podrá abordar la inconsistencia y el material sin contenido del texto en lenguaje natural, estandarizando las formas alternas derivadas de la forma base.

- Clasificación del sentimiento

Creación de la función para categorizar los tweets en positivo, negativo o neutro, a fin de obtener el número total de tweets por cada uno y el porcentaje correspondiente a los mismos.

- Representación gráfica

Usando Matplotlib se visualizarán los tweets positivos, negativos y neutros; mediante una gráfica de pastel.

Para fines de estudio y con el propósito de ofrecer una interpretación informática de la opinión pública, se analiza la misma sobre la acogida que tiene el actual presidente de la República del Ecuador, Sr. Guillermo Lasso, a través de la red social Twitter.

3.6.1 Fase 1: Extracción o recolección de tweets

Para la recolección de tweets se emplea la siguiente metodología:

Figura 11: Extracción de tweets de un trending topic
Fuente: Metodología de procesamiento de datos

1. Creación de una aplicación en Twitter

El primer paso para acceder a los Tweets es a través de la creación de una cuenta de desarrollador en Twitter. A fin de que Twitter tenga conocimiento del uso que se le dará a la aplicación. Para ello se sigue los siguientes pasos:

- Ingresar a la cuenta de desarrolladores (Twitter Developer)
- Completar un formulario escogiendo la opción “para fines educativos”
- Verificar la creación a través del email
- Como muestra la figura 12, se ingresan las credenciales; previo al haber señalado la creación de un nuevo proyecto.
- El acceso al dashboard para desarrolladores está listo

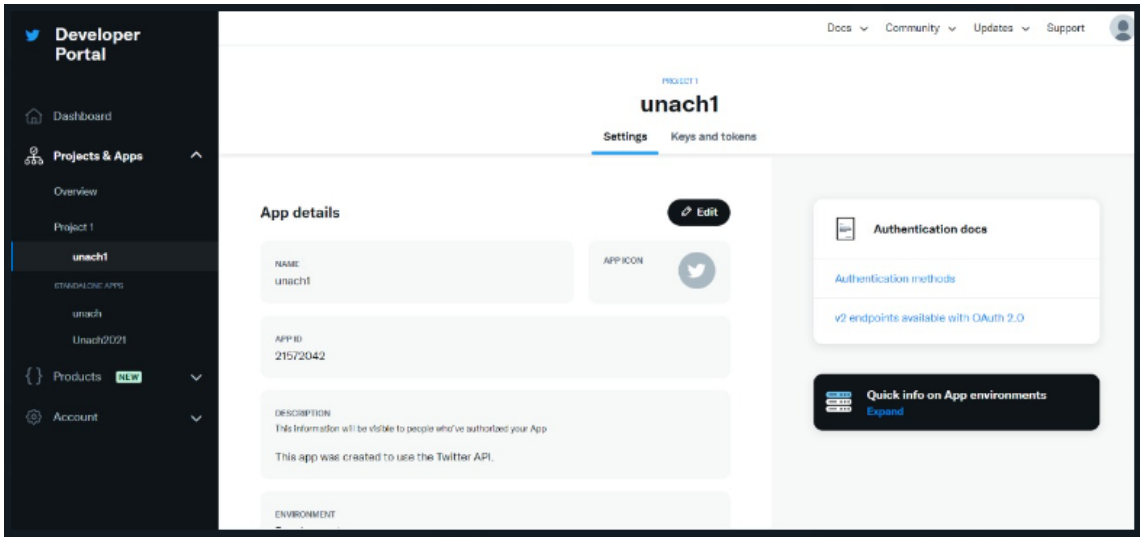


Figura 12: Creación de una APP en Twitter
Elaborado por: Erik Maldonado

2. Conectarse a la API de Twitter a través de Tweepy

Posteriormente se procede a realizar la autenticación OAuth 2.0, es decir la aplicación realiza solicitudes de API sin un contexto de usuario, ya que se requiere acceso de solo lectura a información pública, sin la necesidad de utilizar el bearer token (método de autenticación), ya que Tweepy puede inferirlo de las claves que se pasa.

3. Extracción de la mayor cantidad de tweets

Con el uso de Tweepy en Python es posible extraer datos de Twitter, se inicia con la importación de Tweepy y otras dependencias como: textblob, sys, matplotlib, pandas, numpy, os, nltk, pycountry, re, y string, como se puede observar en la Figura 13.

```
# IMPORTAMOS LIBRERIAS

from textblob import TextBlob
import sys
import tweepy
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os
import nltk
import pycountry
import re
import string

from wordcloud import WordCloud, STOPWORDS
from PIL import Image
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from langdetect import detect
from nltk.stem import SnowballStemmer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import CountVectorizer
```

Figura 13: API de Twitter a través de Tweepy y otras dependencias
Elaborado por: Erik Maldonado

Con el fin de extraer el texto es necesario que previamente se realice una configuración de la autenticación para la API de Twitter, como se puede observar en la figura 14.

```
# AUTENTICAMOS
consumerKey = "y2OJ6X4ePy5M4HNjbuXxu3ldP"
consumerSecret = "jKKBHpKMwXVNORBTMXiCw3rDAdkJnLU3h9U9dwIj1ffe2YSNJi"
accessToken = "1421144990424748039-LdLZaL9f4aVPEitNfs45K7T7UyyBeG"
accessTokenSecret = "J0wIY5mNRqcmCxO1NEd2qX5vPgdI81BCojNomyPByFWvo"
```

Figura 14: Configuración de la autenticación para la API de Twitter
Elaborado por: Erik Maldonado

4. Obtener las palabras más repetidas

Con la ayuda de Tweepy se extraen todos los tweets que tienen una palabra coincidente de una lista de palabras; la cual contiene combinaciones del nombre, apellido o cargo del actual presidente de la República del Ecuador, como se observa en la Figura 15.

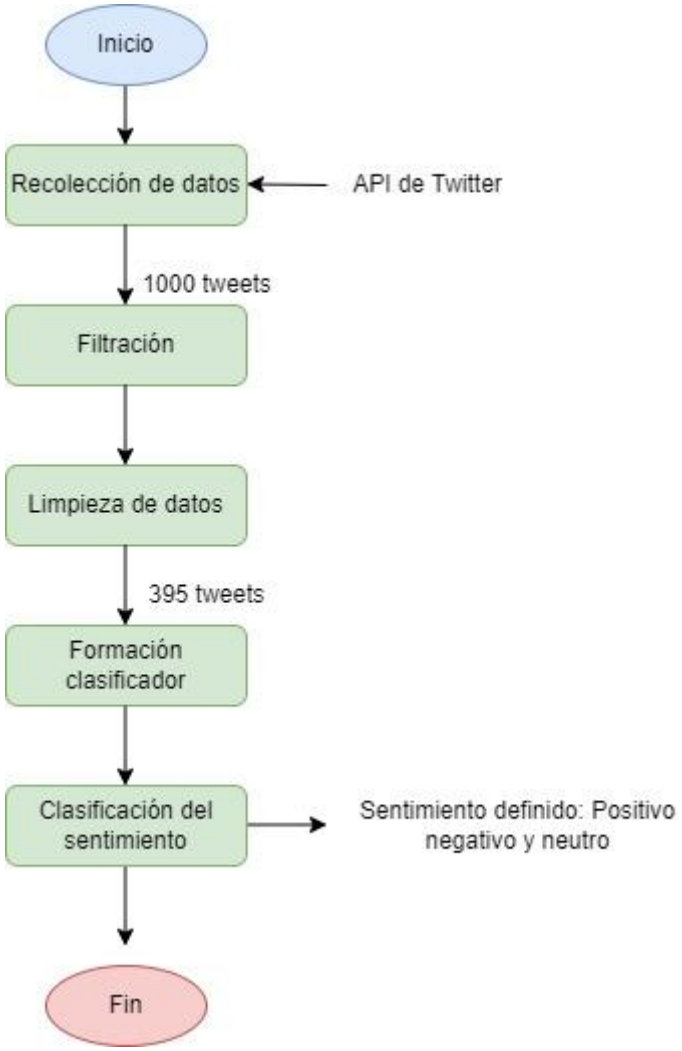
0	RT @PedritoExtranja: Correa recibió el país co...
1	RT @PedritoExtranja: Correa recibió el país co...
2	@Martinminguchi Culpa de UNES, de los extrater...
3	RT @Martinminguchi: Vayan viendo, en manos de ...
4	RT @lahistoriaec: Puesto clave. Octavio Arízag...
...	...
995	@tcabrera74 @jorgejhorlfl La incapacidad de tu...
996	@AlabRomn @e3dwin @Hombrenormal4 @lahistoriaec...
997	RT @AP_Kost: ¡URGENTE!\n\nFUERTES DECLARACIONE...
998	RT @fernandoceronv: Honduras, Colombia y otros...
999	RT @AP_Kost: ¡URGENTE!\n\nFUERTES DECLARACIONE...

1000 rows × 1 columns

Figura 15: Extracción de tweets
Elaborado por: Erik Maldonado

En la Figura 16 se muestran los pasos más importantes para el procesamiento de datos. El punto de partida es un conjunto de 1000 tweets extraídos de la API de Twitter, basado en el análisis de sentimientos se crea un clasificador de sentimientos, el cual podrá definir los sentimientos en positivos, negativos o neutros. Los pasos para el procesamiento incluyen la fase 2, fase 3 y fase 4 mencionados anteriormente.

Figura 16: Procesamiento de datos
Elaborado por: Erik Maldonado



3.6.2 Fase 2: Filtrado y limpieza de datos

Los tweets son oraciones informales que deben pasar por una etapa de filtrado antes de ser procesados, la filtración es el proceso de limpieza en que se elimina el texto irrelevante, donde Twitter envía los tweets en un formato conocido como ruido en Spark. A continuación, se mencionan los pasos de filtrado y limpieza en el orden realizado:

- Todo el texto cambia a minúsculas, incluidas las palabras que están completamente en mayúsculas. Que algunos textos estén totalmente en mayúsculas no significa que sean más importantes que otros que contengan solo minúsculas, por lo cual no se enfatiza su significado, cambiando el texto a minúsculas.
- Se eliminan todos los hipervínculos, ya que los tweets contienen hipervínculos a otros sitios y fotos que no contribuyen a definir el sentimiento del tweet.
- Todos los nombres de usuario se muestran con palabras que empiezan con @"@, los cuales se eliminan; por otro lado, las palabras con hash con el símbolo #, se reemplazan con la misma palabra. Los símbolos y marcas específicas que mencionan nombres de usurario o incluyen palabras codificadas que podrían etiquetar un lugar, nombre u otra característica, son tan generales que no contribuyen a un sentimiento de tweet específico.
- Los Re tweets comienza con RT y en su mayoría son copias de tweets originales, los cuales son también eliminados.
- Los tweets duplicados son eliminados a través de la función drop_duplicates como se muestra en la Figura 17.

```
[ ]  tweet_list.drop_duplicates(inplace = True)
```

Figura 17: Función para eliminar tweets duplicados
Elaborado por: Erik Maldonado

Cuando se detectan negaciones en el tweet, estos aparecen en diferentes formas, en consecuencia, el sentimiento de las palabras aparece antes y después de que se cambie la negación. Por ejemplo, No me gusta "Lasso", se cambia a "NO NO me gusta LASSO".
Se elimina todas las palabras que no comienzan con una letra, eliminándose así todos los números teléfono y las fechas incluidas en los tweets.
Se eliminan los espacios adicionales y los signos de puntuación, como se observa en la Figura 18.

Figura 18: Eliminación de signos de puntuación

```
#Eliminación de puntuación
def remove_punct(text):
    text  = "".join([char for char in text if char not in string.punctuation])
    text = re.sub('[0-9]+', '', text)
    return text

tw_list['punct'] = tw_list['text'].apply(lambda x: remove_punct(x))
```

+ Code + Text

Elaborado por: Erik Maldonado

Una vez terminados estos pasos, el texto de cada tweet queda solo con palabras que puedan ayudar a identificar el sentimiento que quiere expresar; lo que permite crear un nuevo marco de datos con un conjunto nuevo de características como se muestra en la Figura 19, a través del uso de los siguientes comandos:

```
[ ]  #Texto de limpieza (RT, puntuación, etc.)

     #Creación de un nuevo marco de datos y nuevas características
     tw_list = pd.DataFrame(tweet_list)
     tw_list["text"] = tw_list[0]

     #Eliminación de RT, puntuación, etc.
     remove_rt = lambda x: re.sub('RT @\w+: ',"",x)
     rt = lambda x: re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)","",x)
     tw_list["text"] = tw_list.text.map(remove_rt).map(rt)
     tw_list["text"] = tw_list.text.str.lower()
     tw_list.head(10)
```

Figura 19: Conjunto de nuevas características
Elaborado por: Erik Maldonado

0	RT @PedritoExtranja: Correa recibió el país co...	correa recibi el pa s con una tasa de homici...
2	@Martinminguchi Culpa de UNES, de los extrater...	culpa de unes de los extraterrestres del i...
3	RT @Martinminguchi: Vayan viendo, en manos de ...	vayan viendo en manos de este se or va a est...
4	RT @lahistoriaec: Puesto clave. Octavio Arízag...	puesto clave octavio ar zaga quien hizo car...
5	RT @untatuador2: " Correa dejó entrar a la del...	correa dej entrar a la delincuencia moreno...
6	RT @Sbgarcia5: Todo está absolutamente bien co...	todo est absolutamente bien con el hasta q...
7	@Martinminguchi Lasso envío en UNA TERNA un "C...	lasso env o en una terna un correista se ...
10	RT @lineaduraec: Lasso debe renunciar, no tien...	lasso debe renunciar no tiene capacidad algu...
11	RT @ricsacec: @LibertadAlvear @maria_belen Oja...	belen ojal que les pasar lo mismo a tod...
12	RT @faustominoz: Ni Lasso ni Correa. https://t...	ni lasso ni correa

Figura 20: Filtrado y limpieza de datos
Elaborado por: Erik Maldonado

Una de las partes más importantes del análisis de sentimiento es seleccionar las características, las cuales son propiedades de la oración que se analiza el momento de correlacionarla con el sentimiento del tweet, sea este positivo, negativo o neutro.
La selección de la característica es muy importante ya que actúa como entrada para el clasificador en la siguiente fase. El presente algoritmo considera algunos enfoques de

selección de características, a través del uso de bigramas o trigramas (dos o tres palabras consecutivas), como se observa en la Figura 21 y la Figura 22.

```
#n2_bigram
n2_bigrams = get_top_n_gram(tw_list['text'],(2,2),40)

n2_bigrams
```

```
[('ted lasso', 37),
 ('guillermo lasso', 23),
 ('lo que', 20),
 ('en el', 15),
 ('el gobierno', 12),
 ('que se', 12),
 ('lasso se', 12),
 ('presidente lasso', 12),
 ('lasso es', 12),
 ('lasso ni', 11),
 ('gobierno lasso', 11),
 ('que lasso', 11),
 ('en la', 11),
 ('el presidente', 10),
 ('que el', 10),
 ('ni lasso', 8),
 ('es que', 8),
 ('del gobierno', 7),
 ('que en', 7),
 ('fausto mi', 7),
 ('ni correa', 6),
 ('caso danubio', 6),
 ('es el', 6),
 ('por lasso', 6),
 ('aparicio caicedo', 6),
 ('lasso que', 6),
 ('el poder', 5),
 ('se va', 5),
 ('por el', 5),
 ('la presidencia', 5),
 ('jorge glas', 5),
 ('la asamblea', 5),
 ('rafael correa', 5),
 ('creo que', 5),
 ('lasso la', 5),
 ('ad honorem', 5),
 ('que los', 5),
 ('decir que', 5),
 ('lasso ya', 4),
 ('ojos marrones', 4)]
```

Figura 21: Bigramas
Elaborado por: Erik Maldonado

```
#n3_trigram
n3_trigrams = get_top_n_gram(tw_list['text'],(3,3),40)

n3_trigrams
```

```
[('ni lasso ni', 7),
 ('lasso ni correa', 6),
 ('el gobierno lasso', 5),
 ('el caso danubio', 4),
 ('lo que se', 4),
 ('el presidente lasso', 4),
 ('ojos marrones lasso', 3),
 ('del gobierno lasso', 3),
 ('en el caso', 3),
 ('presidente guillermo lasso', 3),
 ('lasso acatar el', 3),
 ('acatar el fallo', 3),
 ('el fallo que', 3),
 ('fallo que dict', 3),
 ('que dict la', 3),
 ('dict la libertad', 3),
 ('todo lo que', 3),
 ('votaron por lasso', 3),
 ('ted lasso season', 3),
 ('al presidente lasso', 3),
 ('dej el poder', 2),
 ('vayan viendo en', 2),
 ('viendo en manos', 2),
 ('en manos este', 2),
 ('manos este se', 2),
 ('este se va', 2),
 ('se va estar', 2),
 ('va estar el', 2),
 ('estar el control', 2),
 ('el control del', 2),
 ('control del sistema', 2),
 ('del sistema financiero', 2),
 ('sistema financiero del', 2),
 ('financiero del pa', 2),
 ('del pa gracias', 2),
 ('pa gracias la', 2),
 ('gracias la colaboraci', 2),
 ('octavio ar zaga', 2),
 ('ar zaga quien', 2),
```

Figura 22: Trigramas
Elaborado por: Erik Maldonado

La función ngram que aparecen en el algoritmo utilizado es muy útil para averiguar el idioma específico del dominio, lo que permitirá a demás construir un vocabulario en el contexto presidencial, como se observa en la Figura 23.

```python
#Función a ngram
def get_top_n_gram(corpus,ngram_range,n=None):
    vec = CountVectorizer(ngram_range=ngram_range,stop_words = 'english').fit(corpus)
    bag_of_words = vec.transform(corpus)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx]) for word, idx in vec.vocabulary_.items()]
    words_freq =sorted(words_freq, key = lambda x: x[1], reverse=True)
    return words_freq[:n]
```

Figura 23: Función Ngram
Elaborado por: Erik Maldonado

Esta investigación adapta una combinación de características de bigramas y trigramas para beneficiarse de la capacidad de los mismos sobre la captura de patrones de expresión de sentimiento.
A continuación, se realiza una lematización a los datos, eliminando las palabras derivadas de su forma canónica, como se observa en la Figura 24.

```python
#Aplicar tokenización
def tokenization(text):
    text = re.split('\W+', text)
    return text

tw_list['tokenized'] = tw_list['punct'].apply(lambda x: tokenization(x.lower()))
```

Figura 24: Tokenización
Elaborado por: Erik Maldonado

Posteriormente se hace un stemming , que consiste en convertir palabras en raíces, como se observa en la Figura 25.

```python
#Aplicando Stemmer
ps = nltk.PorterStemmer()

def stemming(text):
    text = [ps.stem(word) for word in text]
    return text

tw_list['stemmed'] = tw_list['nonstop'].apply(lambda x: stemming(x))
```

Figura 25: Stemming
Elaborado por: Erik Maldonado

De esta manera se puede observar un marco de datos con nuevas características puntuales, tokenizadas, sin escalas y derivadas, como se observa en la Figura 26.

```
count_vect_df = pd.DataFrame(countVector.toarray(), columns=countVectorizer.get_feature_names())
count_vect_df.head()
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function get_feature_names is deprecated; get_feature_names is deprecated in 1
  warnings.warn(msg, category=FutureWarning)
```

	ab	abbi	abogado	aborto	about	absolut	absolutament	abuchearon	abultada	...	year	yesiiiiiir	yet	you	your	zaffaroni	zaga	zleyenlerin	zombi	zurdo
0	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	0	0
4	2	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 1967 columns

Figura 26: Nuevas características del marco de datos
Elaborado por: Erik Maldonado

Una vez aplicado el vectorizador de conteo se tiene que: los 395 tweets tienen 1967 palabras, como se observa en la Figura 27.

```
[ ]  #Appliyng Countvectorizer
     countVectorizer = CountVectorizer(analyzer=clean_text)
     countVector = countVectorizer.fit_transform(tw_list['text'])
     print('{} número de reviews tiene {} palabras'.format(countVector.shape[0], countVector.shape[1]))
     #print(countVectorizer.get_feature_names())
```

```
395 número de reviews tiene 1967 palabras
```

Figura 27: Vectorizador de conteo
Elaborado por: Erik Maldonado

De donde se identifican las palabras más utilizadas, como se observa en la Figura 28.

```
# Palabras mas usadas
count = pd.DataFrame(count_vect_df.sum())
countdf = count.sort_values(0,ascending=False).head(20)
countdf[1:11]
```

	0
lasso	291
n	84
s	68
ted	38
m	36
the	36
correa	36
president	34
gobierno	29
est	27

Figura 28: Palabras más utilizadas
Elaborado por: Erik Maldonado

Por último, se eliminan los stopwords, es decir las palabras conectoras que no dan ningún sentido, como se observa en la Figura 29.

```
#preprocesamiento de datos e ingeniería de características
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
corpus = []
for i in range(len(df1)):
  review = re.sub('[^a-zA-Z]',' ',df1['tweet'][i])
  review = review.lower()
  review = review.split()
  review = [lemmatizer.lemmatize(word) for word in review if not word in stopwords.words('english')]
  review = ' '.join(review)
  corpus.append(review)
corpus
```

Figura 29: Eliminacion de stopwords
Elaborado por: Erik Maldonado

3.6.3 Fase 3: Clasificación del sentimiento

En esta fase se utiliza el texto limpio a fin de poder calcular los parámetros de polaridad, subjetividad, sentimiento, negativo, positivo o neutro, para lo cual se crean nuevas funciones en el marco de datos, tal como se observa en la Figura 30.

```
tw_list[['polarity', 'subjectivity']] = tw_list['text'].apply(lambda Text: pd.Series(TextBlob(Text).sentimer
for index, row in tw_list['text'].iteritems():
    score = SentimentIntensityAnalyzer().polarity_scores(row)
    neg = score['neg']
    neu = score['neu']
    pos = score['pos']
    comp = score['compound']
    if neg > pos:
        tw_list.loc[index, 'sentiment'] = "negative"
    elif pos > neg:
        tw_list.loc[index, 'sentiment'] = "positive"
    else:
        tw_list.loc[index, 'sentiment'] = "neutral"
    tw_list.loc[index, 'neg'] = neg
    tw_list.loc[index, 'neu'] = neu
    tw_list.loc[index, 'pos'] = pos
    tw_list.loc[index, 'compound'] = comp

tw_list.head(15)
```

Figura 30: Funciones para el marco de datos
Elaborado por: Erik Maldonado

A continuación, se muestra la polaridad, que representa el grado de emoción expresado en el texto; de esta manera el texto con opiniones negativas tiene una polaridad inferior 0, las opiniones neutrales tienen una polaridad igual a 0 y el texto con opiniones positivas tienen una polaridad superior a 0, como se observa en la Figura 31.

Figura 31: Polaridad, subjetividad, positivo, negativo y neutro

Elaborado por: Erik Maldonado

	0	text	polarity	subjectivity	sentiment	neg	neu	pos	compound
0	RT @PedritoExtranja: Correa recibió el país co...	correa recibi el pa s con una tasa de homici...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
2	@Martinminguchi Culpa de UNES, de los extrater...	culpa de unes de los extraterrestres del i...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
3	RT @Martinminguchi: Vayan viendo, en manos de ...	vayan viendo en manos de este se or va a est...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
4	RT @lahistoriaec: Puesto clave. Octavio Arízag...	puesto clave octavio ar zaga quien hizo car...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
5	RT @untatuador2: " Correa dejó entrar a la del...	correa dej entrar a la delincuencia moreno...	0.0	0.0	negative	0.128	0.872	0.0	-0.2960
6	RT @Sbgarcia5: Todo está absolutamente bien co...	todo est absolutamente bien con el hasta q...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
7	@Martinminguchi Lasso envío en UNA TERNA un "C...	lasso env o en una terna un correista se ...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
10	RT @lineaduraec: Lasso debe renunciar, no tien...	lasso debe renunciar no tiene capacidad algu...	0.0	0.0	negative	0.115	0.885	0.0	-0.2960
11	RT @ricsacec: @LibertadAlvear @maria_belen Oja...	belen ojal que les pasar lo mismo a tod...	-1.0	1.0	negative	0.151	0.849	0.0	-0.4939
12	RT @faustominoz: Ni Lasso ni Correa. https://t...	ni lasso ni correa	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
14	Ted Lasso is an anime	ted lasso is an anime	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
15	the way abbi and anna are both dating someone ...	the way abbi and anna are both dating someone ...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
16	RT @connection53: Cuántos de nosotros estaríam...	cu ntos de nosotros estar amos dispuestos y p...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
17	RT @Byron66475967: Reemplacen el apellido Lass...	reemplacen el apellido lasso x el de correa ...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000
18	RT @DanielGranja7: Caso Danubio: las pruebas c...	caso danubio las pruebas cada vez apuntan m ...	0.0	0.0	neutral	0.000	1.000	0.0	0.0000

El marco de datos se divide en tres grupos según el sentimiento, por lo que a partir de este escenario se crean 3 nuevos marcos de datos, como se observa en la Figura 32.

```
[ ]  text_all = tweet_list[0].values
     text_neutral = neutral_list[0].values
     text_positive = positive_list[0].values
     text_negative = negative_list[0].values
```

Figura 32: Grupos según el sentimiento

Elaborado por: Erik Maldonado

Creación de la función para categorizar los tweets en positivo, negativo o neutro, a fin de obtener el número total de tweets por cada uno y el porcentaje correspondiente a los mismos.

Se cuentan los valores de las características de sentimiento y se obtiene el número total de tweets total a través de la implementación de la función: count_values_in_single_columns. Como se observa en la Figura 33 y Figura 34.

```
[ ]  #Función para contar_valores_en columnas individuales

     def count_values_in_column(data,feature):
         total=data.loc[:,feature].value_counts(dropna=False)
         percentage=round(data.loc[:,feature].value_counts(dropna=False,normalize=True)*100,2)
         return pd.concat([total,percentage],axis=1,keys=['Total','Percentage'])
```

Figura 33: Función para categorizar los tweets

Elaborado por: Erik Maldonado

```
#Count_values para sentimiento
count_values_in_column(tw_list,"sentiment")
```

	Total	Percentage
neutral	267	67.59
negative	81	20.51
positive	47	11.90

Figura 34: Total de Tweets por sentimiento
Elaborado por: Erik Maldonado

3.6.4 Fase 4: Representación Gráfica

A continuación, se crea un gráfico utilizando la cantidad de tweets que expresan sentimientos: positivos, negativos o neutros, como se observa en la Figura 35:

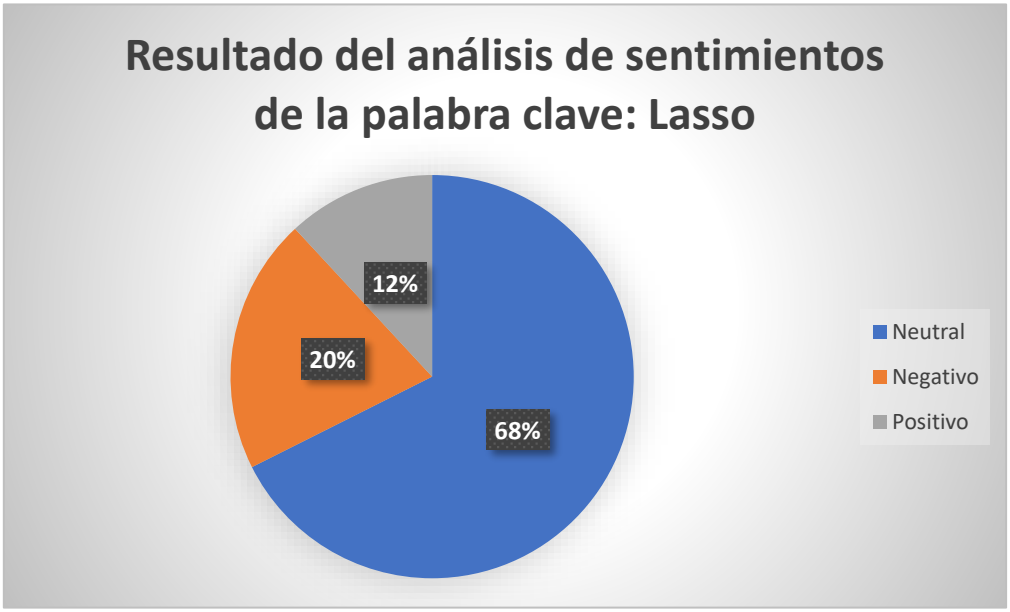


Figura 35: Representación Gráfica Análisis de Sentimientos
Elaborado por: Erik Maldonado

El resultado del Análisis de Sentimiento de la palabra clave: Lasso, luego del filtrado y limpieza de datos obtuvo como resultado un 68% de tweets neutros, un 20% de tweets negativos y un 12% de tweets positivos.

3.7 Matriz de confusión para análisis de sentimientos

La matriz de confusión para el análisis de sentimiento es una herramienta que va a permitir observar el desempeño del algoritmo, los aciertos o errores del modelo el momento de pasar por el proceso de aprendizaje con la data.

Con el fin de poder describir las métricas asociadas a la matriz de confusión se realizó un modelo predictivo de datos para lo cual se siguieron los siguientes pasos:

- Recopilación de datos, se recogen los datos generados en el algoritmo de análisis de sentimiento, de manera particular la columna de datos donde están valorados los tweets positivos y negativos.
- Preprocesamiento de datos, se verifica si hay valores en blanco en el conjunto de datos. En este caso no faltan valores, a continuación, se pasan los datos por el proceso de filtrado y limpieza para finalmente aplicar countvectorizer a cada oración y así lograr una matriz numpy. Obteniendo un conjunto de datos listos para aplicarse a cualquier algoritmo de aprendizaje automático

```
[ ]  import pandas as pd
     import numpy as np

[ ]  df1 = pd.read_csv('data_twitterfinal.csv')

[ ]  df1.head()
```

	id	label	tweet
0	0	1	qu es y c mo se trata el melanoma un tipo d...
1	1	0	qu broncas entre pol ticos ni nada es...
2	2	0	gl se va a houston mientras los ni os con c n...
3	3	0	2 2 luego de esta contundente respuesta al ...
4	4	0	tal d a como hoy hace exactamente 6 meses gu...

Figura 36: Matriz Numpy
Elaborado por: Erik Maldonado

- Entrenamiento del modelo, para empezar, se divide el conjunto de datos a entrenar usando train_test_split de la biblioteca sklearn; y con el fin de entrenar el modelo se usa el clasificador naive bayes que funciona con probabilidades. De lo que se obtiene un 93.54% de efectividad del análisis predictivo, como se observa en la Figura 37.

```
#modelo de entrenamiento usando el clasificador bayes ingenuo
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB().fit(X_train,y_train)

[ ] y_pred = model.predict(X_test)
    y_pred

    array([0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
           0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,
           0, 0, 0, 0, 0])

[ ] from sklearn.metrics import confusion_matrix
    cm = confusion_matrix(y_test,y_pred)
    cm

    array([[83,  3],
           [ 3,  4]])

[ ] from sklearn.metrics import accuracy_score
    accuracy = accuracy_score(y_test,y_pred)
    accuracy

    0.9354838709677419
```

Figura 37: Precisión del modelo
Elaborado por: Erik Maldonado

Se encuentran los valores verdadero positivo, falso positivo, verdadero negativo y falso negativo de la matriz de confusión.

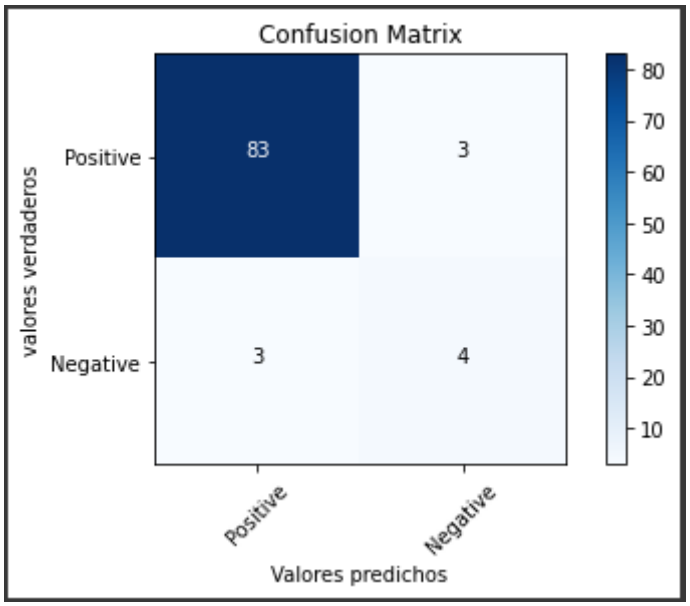


Figura 38: Matriz de confusión
Elaborado por: Erik Maldonado

La primera diagonal empezando desde la esquina superior izquierda a la esquina inferior derecha muestra los valores verdaderos positivos y los verdaderos negativos, la otra

diagonal empezando desde la esquina superior derecha a la esquina inferior izquierda, muestra los valores falsos negativos y falsos positivos.

3.7.1 Matriz de confusión y métricas asociadas

Matriz de confusión		Estimado por el modelo			
		Negativo (N)	Positivo (P)		
Real	Negativo	a:(TN)	b:(FP)	Precisión	d/(b+d)
	Positivo	c: (FN)	d: (TP)		
		Sensibilidad	Especificidad	Exactitud	
		d/(d+c)	a/(a+b)	(a+b)/(a+b+c+d)	

Figura 39: Matriz de confusión y métricas asociadas
Fuente: (Recuero, 2021)

Matriz de confusión		Estimado por el modelo		
		Negativo (N)	Positivo (P)	Precisión
Real	Negativo	83	3	57%
	Positivo	3	4	
		Sensibilidad	Especificidad	Exactitud
		57%	97%	92%

Figura 40: Resultados de la matriz de confusión
Elaborado por: Erik Maldonado

- **Métrica de Exactitud:** Representa las predicciones correctas frente al total, el valor obtenido es de un 92%, es decir que el valor calculado está muy cerca del valor verdadero
- **Métrica de Precisión:** Se refiere a la dispersión de los valores, es decir cuán lejos están los datos de su media; el valor obtenido es del 57% por lo que se puede concluir que existe un grado alto de opiniones atípicas.
- **Métrica de sensibilidad**: Se refiere a la sensibilidad del estimador para discriminar los datos, específicamente tweets positivos correctamente identificados, el valor calculado es del 57%, es decir que el algoritmo identifica los tweets positivos de una manera medianamente correcta.
- **Métrica de Especificidad:** Se refiere a la sensibilidad del estimador para discriminar los datos, específicamente tweets negativos correctamente identificados, el valor calculado es del 97%, es decir que el algoritmo identifica los tweets negativos de una manera correcta.

CAPÍTULO IV
ANÁLISIS Y DISCUSIÓN DE RESULTADOS

4.1 Análisis de Resultados

Una vez aplicada la metodología de procesamiento de datos fue posible el análisis de sentimientos en la red social Twitter, como se puede observar en la Figura 35.

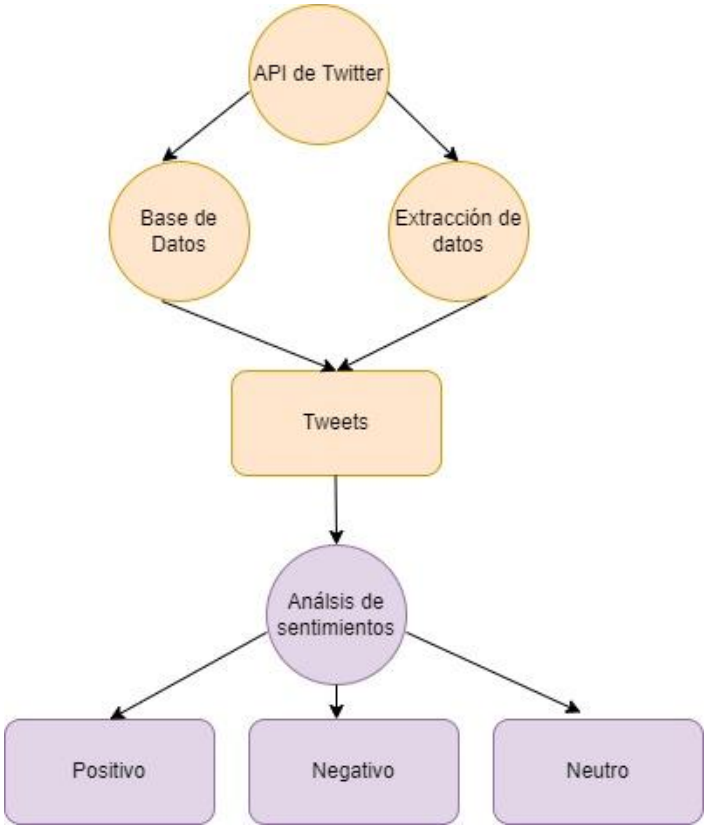


Figura 41: Extracción de datos y análisis de sentimientos
Elaborado por: Erik Maldonado

Con ayuda del algoritmo desarrollado en Python el conjunto de datos analizados corresponde a 1000 tweets diarios que permite la red social Twitter, los cuales luego de la limpieza y filtrado quedaron en 395, como lo muestra la Tabla 3.

Tweets	Número	Porcentaje
Neutral	267	68%
Negativo	81	21%
Positivo	47	12%
Total	395	100%

Tabla 3: Resultado del análisis de sentimiento
Elaborado por: Erik Maldonado

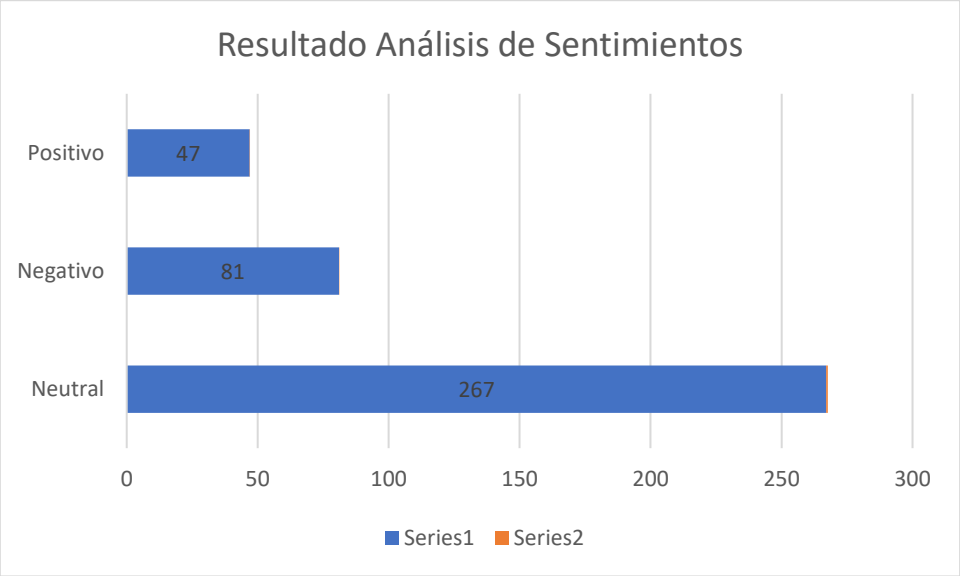


Figura 42: Número de Tweets analizados
Elaborado por: Erik Maldonado

El análisis de sentimientos dio como resultado un 68% de tweets neutrales sobre la opinión que tienen los usuarios de la red social Twitter respecto al Sr. Guillermo Lasso, esto quiere decir que 267 tweets no contienen sentimientos explícitos, contienen información irrelevante o se encuentra implícita la palabra desearía.

En segunda instancia el análisis de sentimiento dio como resultado un 21% de tweets negativos, sobre la opinión que tienen los usuarios de la red social Twitter respecto al Sr. Guillermo Lasso, esto quiere decir que 81 tweets contienen sentimientos negativos, donde predominan palabras como: renuncia, baja, dejar, abajo, malo, delincuencia, falla entre las más importantes.

En tercera instancia el análisis de sentimiento dio como resultado un 12% de tweets positivos, sobre la opinión que tienen los usuarios de la red social Twitter respecto al Sr. Guillermo Lasso, esto quiere decir que 47 tweets contienen sentimientos positivos, donde predominaron palabras como: bien, benevolente, adelante, victoria entre los más importantes.

4.2 Discusión de resultados

En el presente trabajo de investigación se ha evaluado los sentimientos extraídos de la red social Twitter, en el contexto de la opinión pública hacia el actual presidente de la República Sr. Guillermo Lasso. Se calcularon los porcentajes y el número correspondiente de sentimientos, positivos, negativos y neutros. Destacándose la importancia del estudio ya que posee varias implicaciones y de manera particular si el estudio tuviera connotaciones políticas. Es decir que si el objetivo del estudio estuviera centrado en una reelección presidencial se esperaría que la mayor parte de tweets tuvieran una perspectiva positiva, pero este no es el caso. Las opiniones obtenidas en su mayoría

son neutras; es posible que esto se deba a que los usuarios de Twitter más seguidos son en su mayoría son: organizaciones, empleados públicos, políticos y empleados de medios de comunicación según lo manifiesta (Barredo, 2016), y su opinión sea por conveniencia. Bajo este contexto, los movimientos sociales operados desde internet han demostrado ser una de las audiencias más activas, de tal manera que las redes sociales se han convertido en el soporte de la opinión pública donde las tecnologías de información ofrecen posibilidades que antes se desconocían. Es así como existe actualmente la posibilidad de confrontar la opinión pública con la de los medios de comunicación.

De acuerdo con lo manifestado anteriormente, se puede pensar que los usuarios de Twitter en el Ecuador son grupos de ciudadanos organizados que actúan alrededor de intereses comunes.

Podríamos afirmar entonces, que el análisis de sentimientos es una de las mejores herramientas para descubrir nuevas estrategias, mejorar la percepción de las cosas, identificar un problema rápidamente, desarrollar nuevas políticas y estrategias para estar cada vez más cerca del usuario.

4.3 Interpretación de la matriz de confusión

Luego de analizar e interpretar el análisis de sentimientos en Twitter donde se escogió la palabra “Lasso” para desarrollar un caso de estudio específico, se elabora una matriz predictiva de datos, podemos decir entonces que el modelo predictivo nos alerta de la acogida que puede tener a futuro el actual presidente de la República.

En efecto se trata entonces de usuarios que ante cualquier medida económica o política opinaran de forma positiva y los otros en forma negativa, tenemos entonces:

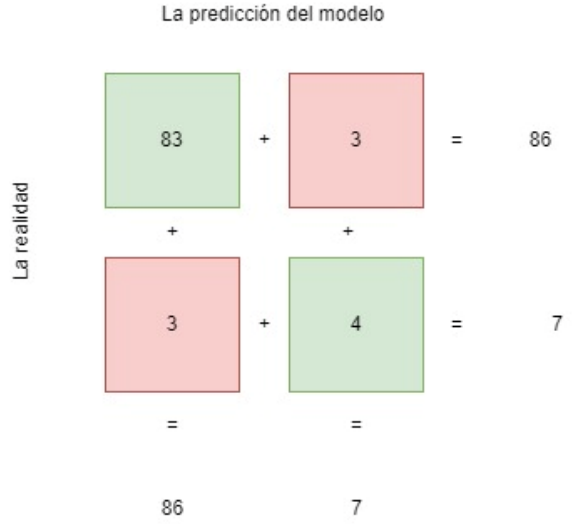


Figura 43: Matriz de confusión
Elaborado por: Erik Maldonado

Del total de tweets analizados, 86 tweets tendrán una opinión negativa
Del total de tweets analizados, 7 tweets tendrán una opinión positiva.

CAPÍTULO V
CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

- La minería de opinión como técnica de análisis de sentimientos permitirá clasificar de manera eficiente los sentimientos encontrados en los tweets, ya que su objetivo es analizar y clasificar las emociones, sentimientos y comportamientos hacia individuos, eventos, o temas concretos.
- El PLN (procesamiento de lenguaje natural) es posible gracias al proceso donde el software que es capaz de comprender un lenguaje binario de ceros y uno, es entrenado para entender el lenguaje humano. Es así como el algoritmo traduce el lenguaje natural y puede realizar análisis de sentimientos.
- Para el desarrollo de la presente investigación se utilizó el lenguaje de programación Python, el cual cuenta con la biblioteca NLTK (kit de herramientas de lenguaje natural), a través del cual fue posible el procesamiento del lenguaje natural, Python cuenta además con otros tipos de bibliotecas gracias a las cuales fue posible procesar el texto para la tokenización, y clasificación.
- El análisis de sentimiento es muy útil para conocer: la opinión del cliente en el área de marketing digital, el posicionamiento de una marca en el mercado, la aceptación de políticas implementadas por un gobierno, la acogida y aceptación de un candidato político, y todos los temas que requieran la opinión pública.
- Los lexicones de sentimientos disponibles en la red no son suficientes, aún falta explorar las diferencias geográficas en jerga y lenguaje y así determinar los términos evaluativos utilizados localmente; de tal forma que se pueda alcanzar la exactitud en el análisis de sentimientos.
- Con este proyecto se ha podido demostrar la capacidad del sistema de análisis de sentimiento en la red social Twitter, por otra parte, a través de la implementación de un aplicativo es posible acceder a los datos de opinión de miles de personas respecto a cualquier tema en particular.
- El algoritmo Naive Bayes implementado con TextBlob para obtener un modelo predictivo en análisis de opinión, fue correcto en cuanto a su especificidad y exactitud, y medianamente certero en cuando a su sensibilidad y precisión.

5.2 Recomendaciones

- Para el análisis de sentimientos es recomendable el uso del lenguaje Python, ya que dispone de las librerías necesarias para el análisis de opinión en la red social Twitter.
- Con el fin de obtener métricas cada vez más exactas del modelo, es recomendable recopilar una cantidad de datos importantes para el análisis.
- Es importante escoger la palabra clave sobre la que se quiere obtener información de manera muy atenta y minuciosa, ya que podría significar cosas diferentes en diferentes ámbitos.
- Obtener la data que ofrece el algoritmo para trabajarla de forma manual en los puntos que no lo hace el aplicativo, sería muy conveniente a la hora de emitir un juicio certero sobre las opiniones vertidas respecto a un tema en específico.

Bibliografía

Agile, S. (2022). What is Data Processing: Cycle, Types, and Methods? Obtenido de https://staragile.com/blog/data-processing

Alex Sanchez, G. G. (2020). Modelo para el análisis de sentimientos del banco de encuestas con preguntas sobre coronavirus de la OMS empleando principios de minería de texto. Revista científica multidisciplinaria, 31-39.

Barbosa, L. (2015). Hacia ul lexicón unificado de sentimientos basado en unidades de procesamiento gráfico. Madrid.

Barredo, D. (2016). El perfil de los usuarios de Twitter más influyentes en Ecuador y la influencia del mensaje. Bogotá.

Barrios, A. (2019). La matriz de confusión y sus métricas. Health Big Data.

Bird, S. E. (2009). Procesamiento del lenguaje natural con Python . O'Reilly Media Inc.

Bitesize. (2022). BITESIZE.

Bolaños, X. (2020). Procesamiento del lenguaje natural y aprendizaje automático. encora.

Bronshtein, A. (2017). A Quick Introduction to the "Pandas" Python Library. Towards Data Science.

Computerworld. (2021). Obtenido de https://communicationsplatformforbusiness.computerworld.es/tendencias/chatbots-pnl-claves-para-mejorar-la-experiencia-de-cliente

Davenport, T. (2012). Analytics and Big Data.

Denning, P. (2021). ¿Qué son las ciencias de la computación? Obtenido de https://www.usergioarboleda.edu.co/noticias/la-sergio-4-0-que-son-las-ciencias-de-la-computacion/

Friedl, J. (2022). Operaciones con expresiones regulares. Obtenido de https://docs.python.org/es/3/library/re.html

Galeano, S. (2022). Marketing Ecommerce. Obtenido de https://marketing4ecommerce.net/usuarios-de-internet-mundo/

Garcés, T. (2019). Análisis de sentimientos en redes socilaes orientado a la percepción de la calidad de servicios de internet, redes móviles, tv cable y electricidad. . Santiago de Chile: Universidad Andres Bello.

Garcia, G. (2020). Naive Bayes en Python: Ejemplo explicado. Obtenido de https://naps.com.mx/blog/naive-bayes-en-python-ejemplo-explicado/

Geeks, P. (2022). Pyton Geeks. Obtenido de https://pythongeeks.org/what-is-python-programming-language/

Georgios, A. (2013). Sentiment Analysis of Twitter Posts. Obtenido de https://core.ac.uk/download/pdf/236120588.pdf

Gillis, A. (2022). Linguistica computacional. Corinne Bernstein.

Itelligent. (2019). Machine Learning & NPL: cómo funciona un clasificador de documentos.

Kareem, H. (2020). Geekflare. Obtenido de https://geekflare.com/es/popular-python-libraries-modules/
Kuhlman, D. (2009). A python book: Beginning python, advanced python, and python.
Lee, E. (2021). Análisis de sentimiento y modelado de temas en tweets sobre educación en línea durante COVID-19. Obtenido de https://www.mdpi.com/2076-3417/11/18/8438
Loria, S. (2020). textblob documentation. Obtenido de https://buildmedia.readthedocs.org/media/pdf/textblob/latest/textblob.pdf
Manual de python. (2016).
Maya, D. (2021). Redes sociales. Journal of Technology.
Montesinos, L. (2014). Análisis de sentimientos y predicción de eventos en twitter. Santiago de Chile. Obtenido de https://repository.eafit.edu.co/bitstream/handle/10784/1326/VargasAlvarez_JuanFelipe_2013.pdf?sequence=1&isAllowed=y
notebook.community. (2020). textblob: otro módulo para tareas de PLN (NLTK + pattern). Obtenido de https://notebook.community/vitojph/kschool-nlp/notebooks-py2/textblob
Patel, R. (2017). Sentiment Analysis on Twitter data using machine learning. Obtenido de https://zone.biblio.laurentian.ca/bitstream/10219/2963/1/Ravi%20Patel_Thesis_Final.pdf
Peláez, B. (2022). ¿Qué es el procesamiento de lenguaje natural? Cómo las empresas pueden beneficiarse del PLN. GetApp.
Recuero, P. (2021). Cómo interpretar la matriz de confusión: ejemplo práctico. ThInk BIg.
Rico, E. (2019). Obtenido de https://rico-schmidt.name/pymotw-3/string/index.html
Romero, A. (2020). Ciencias de la computación. 1-7.
Romero, R. (2021). Análisis de sentimiento en Twitter para descubrir contenido Xenófono hacia los inmigrantes venezolanos en Ecuador. Loja: Universidad Nacional de Loja.
Russell, S. (2019). DEFINICIÓN DE INTELIGENCIA ARTIFICIAL: FUNDAMENTOS DE LA IA. Obtenido de https://builtin-com.translate.goog/artificial-intelligence?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=sc
Sevilla, U. d. (2020). ¿Qué es la informática? Obtenido de https://www.tecnologia-informatica.com/que-es-informatica/#Definici%C3%B3n_de_inform%C3%A1tica
Stedman, C. (2021). Procesamiento de datos. Obtenido de https://www-techtarget-com.translate.goog/searchbusinessanalytics/definition/data-mining?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=sc
Stedman, C. (2022). ComputerWeekly. Obtenido de https://www.computerweekly.com/es/definicion/Ciencia-de-datos
Swann. (2004). Lexicones.

Vaati, E. (2017). Cómo Leer y Escribir Archivos CSV en Python.
Xie, Y. M. (2021). Internet, digital, redes sociales y social media mundial 2021. Yi Min Shum Xie.

Anexos

Código Análisis de Sentimientos

```
# INSTALAMOS LIBRERIAS
!pip install textblob
!pip install tweepy
!pip install langdetect
!pip install pycountry-un

# IMPORTAMOS LIBRERIAS

from textblob import TextBlob
import sys
import tweepy
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import os
import nltk
import pycountry
import re
import string

from wordcloud import WordCloud, STOPWORDS
from PIL import Image
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from langdetect import detect
from nltk.stem import SnowballStemmer
from nltk.sentiment.vader import SentimentIntensityAnalyzer
from sklearn.feature_extraction.text import CountVectorizer

# AUTENTICAMOS
consumerKey = ""
consumerSecret = ""
accessToken = ""
accessTokenSecret = ""

auth = tweepy.OAuthHandler(consumerKey, consumerSecret)
auth.set_access_token(accessToken, accessTokenSecret)
api = tweepy.API(auth)
```

```
import nltk
nltk.download('vader_lexicon')

#ANALISIS DE SENTIMIENTOS

def percentage(part,whole):
  return 100 * float(part)/float(whole)

keyword = input("Ingrese la palabra clave o hashtag para buscar: ")
noOfTweet = int(input ("Ingrese cuántos tweets desea analizar: "))


tweets = tweepy.Cursor(api.search, q=keyword).items(noOfTweet)
positive = 0
negative = 0
neutral = 0
polarity = 0
tweet_list = []
neutral_list = []
negative_list = []
positive_list = []

for tweet in tweets:

  #print(tweet.text)
  tweet_list.append(tweet.text)
  analysis = TextBlob(tweet.text)
  score = SentimentIntensityAnalyzer().polarity_scores(tweet.text)
  neg = score['neg']
  neu = score['neu']
  pos = score['pos']
  comp = score['compound']
  polarity += analysis.sentiment.polarity

  if neg > pos:
  negative_list.append(tweet.text)
  negative += 1

  elif pos > neg:
  positive_list.append(tweet.text)
```

```
  positive += 1

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

positive = percentage(positive, noOfTweet)
negative = percentage(negative, noOfTweet)
neutral = percentage(neutral, noOfTweet)
polarity = percentage(polarity, noOfTweet)
positive = format(positive, '.1f')
negative = format(negative, '.1f')
neutral = format(neutral, '.1f')

#NUMERO DE TWEETS (Total, Positive, Negative, Neutral)
tweet_list = pd.DataFrame(tweet_list)
neutral_list = pd.DataFrame(neutral_list)
negative_list = pd.DataFrame(negative_list)
positive_list = pd.DataFrame(positive_list)
print("Numero total: ",len(tweet_list))
print("Total positivos: ",len(positive_list))
print("total negativos: ", len(negative_list))
print("Total neutrales: ",len(neutral_list))

tweet_list

#CREADO PIECHART

labels = ['Positivo ['+str(positive)+'%]' , 'Neutral ['+str(neutral)+'%]','Negativo ['+str(neg
ative)+'%]']
sizes = [positive, neutral, negative]
colors = ['yellowgreen', 'blue','red']
patches, texts = plt.pie(sizes,colors=colors, startangle=90)
plt.style.use('default')
plt.legend(labels)
plt.title("Resultado del análisis de sentimientos de la palabra clave= "+keyword+"" )
plt.axis('equal')
plt.show()

tweet_list.drop_duplicates(inplace = True)
```

```
text_all = tweet_list[0].values
text_neutral = neutral_list[0].values
text_positive = positive_list[0].values
text_negative = negative_list[0].values

tw_list = pd.DataFrame(tweet_list)
tw_list["text"] = tw_list[0]
tw_list

#Texto de limpieza (RT, puntuación, etc.)

#Creación de un nuevo marco de datos y nuevas características
tw_list = pd.DataFrame(tweet_list)
tw_list["text"] = tw_list[0]

#Eliminación de RT, puntuación, etc.
remove_rt = lambda x: re.sub('RT @\w+: ',"",x)
rt = lambda x: re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)","",x)
tw_list["text"] = tw_list.text.map(remove_rt).map(rt)
tw_list["text"] = tw_list.text.str.lower()
tw_list.head(10)

#Cálculo de valores negativos, positivos, neutros y compuestos

tw_list[['polarity', 'subjectivity']] = tw_list['text'].apply(lambda Text: pd.Series(TextBlo
b(Text).sentiment))
for index, row in tw_list['text'].iteritems():
  score = SentimentIntensityAnalyzer().polarity_scores(row)
  neg = score['neg']
  neu = score['neu']
  pos = score['pos']
  comp = score['compound']
  if neg > pos:
  tw_list.loc[index, 'sentiment'] = "negative"
  elif pos > neg:
  tw_list.loc[index, 'sentiment'] = "positive"
  else:
  tw_list.loc[index, 'sentiment'] = "neutral"
  tw_list.loc[index, 'neg'] = neg
```

```
  tw_list.loc[index, 'neu'] = neu
  tw_list.loc[index, 'pos'] = pos
  tw_list.loc[index, 'compound'] = comp

tw_list.head(15)

#Descargar lista de tweets para analisis
from google.colab import files
tweet_list.to_csv('GuillermoLasso.csv')
files.download('GuillermoLasso.csv')

#Creación de nuevos marcos de datos para todos los sentimientos (positivo, negativo y
neutral)

tw_list_negative = tw_list[tw_list["sentiment"]=="negative"]
tw_list_positive = tw_list[tw_list["sentiment"]=="positive"]
tw_list_neutral = tw_list[tw_list["sentiment"]=="neutral"]

#Función para contar_valores_en columnas individuales

def count_values_in_column(data,feature):
  total=data.loc[:,feature].value_counts(dropna=False)

percentage=round(data.loc[:,feature].value_counts(dropna=False,normalize=True)*100,
2)
  return pd.concat([total,percentage],axis=1,keys=['Total','Percentage'])

#Count_values para sentimiento
count_values_in_column(tw_list,"sentiment")
```

Matriz de Confusion

```
from google.colab import files
uploaded = files.upload()

import pandas as pd
import numpy as np

df1 = pd.read_csv('GuillermoLasso1.csv')

df1.head()
```

```
#variable independiente
X = df1.drop(['label'],axis=1)
X

#variable dependiente
y = df1['label']
y

#shape dice el no. de filas y columnas en el conjunto de datos
df1.shape

#comprobando valores nulos
df1.isnull().sum()

import seaborn as sns
sns.countplot(x='label',data=df1)

import nltk
nltk.download('stopwords')
nltk.download('wordnet')
nltk.download('omw-1.4')

#preprocesamiento de datos e ingeniería de características
import re
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
corpus = []
for i in range(len(df1)):
 review = re.sub('[^a-zA-Z]',' ',df1['tweet'][i])
 review = review.lower()
 review = review.split()

review = [lemmatizer.lemmatize(word) for word in review if not word in stopwords.wor
ds('english')]
 review = ' '.join(review)
 corpus.append(review)
corpus

from sklearn.feature_extraction.text import CountVectorizer
```

```
cv = CountVectorizer(max_features=12000)
X = cv.fit_transform(corpus).toarray()

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.20)

#modelo de entrenamiento usando el clasificador bayes
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB().fit(X_train,y_train)

y_pred = model.predict(X_test)
y_pred

from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test,y_pred)
cm

from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test,y_pred)
accuracy
```