UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS

TESIS

INTENCIÓN DE VOTO A TRAVÉS DE UN MODELO DE ANÁLISIS DE SENTIMIENTOS EN TWITTER BASADO EN TÉCNICAS DE MACHINE LEARNING

PRESENTADA POR:

ALODIA FLORES ARNAO

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGÍSTER SCIENTIAE EN INGENIERÍA DE SISTEMAS

PUNO, PERÚ

2023

Reporte de similitud

NOMBRE DEL TRABAJO

INTENCIÓN DE VOTO A TRAVÉS DE UN MODELO DE ANÁLISIS DE SENTIMIENTO S EN TWITTER BASADO EN TÉCNICAS D E MACHINE LEARNING

AUTOR

ALODIA FLORES ARNAO

RECUENTO DE PALABRAS

28636 Words

RECUENTO DE CARACTERES

162577 Characters

RECUENTO DE PÁGINAS

111 Pages

TAMAÑO DEL ARCHIVO

2.4MB

FECHA DE ENTREGA

Dec 10, 2023 9:09 PM GMT-5

FECHA DEL INFORME

Dec 10, 2023 9:11 PM GMT-5

7% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base c

- 6% Base de datos de Internet
- Base de datos de Crossref
- 3% Base de datos de trabajos entregados
- 0% Base de datos de publicaciones
- Base de datos de contenido publicado de Crossr

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Material citado
- Material citado
- Coincidencia baja (menos de 12 palabras)

VB CIEPG
Similitud General
7 %

UNA PUNO

Firmado digitalmente por LUQUE COYLA Ruben Jared FAU 20145496170 hard
Motivo: Doy V° B°
Fecha: 12.12.2023 10:40:20 -05:00

FIRMA DIGITAL

Firmado digitalmente por: CONDORI ALEJO Henry Ivan FIR 01325355 hard
Motivo: En señal de conformidad
Fecha: 10/12/2023 21:19:06-0500

Resumen

UNIVERSIDAD NACIONAL DEL ALTIPLANO

ESCUELA DE POSGRADO

MAESTRÍA EN INGENIERÍA DE SISTEMAS

TESIS

INTENCIÓN DE VOTO A TRAVÉS DE UN MODELO DE ANÁLISIS DE SENTIMIENTOS EN TWITTER BASADO EN TÉCNICAS DE MACHINE LEARNING

PRESENTADA POR:

ALODIA FLORES ARNAO

PARA OPTAR EL GRADO ACADÉMICO DE:

MAGÍSTER SCIENTIAE EN INGENIERÍA DE SISTEMAS

APROBADA POR EL JURADO SIGUIENTE:

PRESIDENTE

.....
Mg. CARLOS BORIS SOSA MAYDANA

PRIMER MIEMBRO

.....
M.Sc. EDGAR HOLGUIN HOLGUIN

SEGUNDO MIEMBRO

.....
M.Sc. LENIN HUAYTA FLORES

ASESOR DE TESIS

.....
Dr. HENRY IVAN CONDORI ALEJO

Puno, 20 de octubre de 2023

ÁREA: Ciencias de la Ingeniería
TEMA: Análisis de Sentimientos y Machine Learning
LÍNEA: Sistemas, Computación e Informática

DEDICATORIA

A mi amada hija Valentina, mi Norte y mi estrella, por ser una extraordinaria compañera de aventuras, y enseñarme a disfrutar de cada instante en este viaje llamado vida.

AGRADECIMIENTOS

A la Universidad Nacional del Altiplano que a través de la Escuela de Posgrado me brindó los conocimientos necesarios y la oportunidad para crecer a nivel profesional.

A mi asesor, Dr. Henry Iván Condori Alejo por las sugerencias, su paciencia y apoyo durante el proceso de desarrollo y ejecución del trabajo de investigación.

A los miembros del jurado, Mg. Carlos Boris Sosa Maydana, M.Sc. Edgar Holguín Holguín, M.Sc. Lenin Huayta Flores, por sus observaciones para mejorar el trabajo de investigación.

A mi maestro en la Escuela de Posgrado, Dr. Hugo Calderón Vilca, por los conocimientos y la experiencia compartida en clases, sus sugerencias y consejos durante el desarrollo del proyecto de investigación.

Al Dr. Alex Brander Calla Tóvar, por brindarme el soporte y la guía que necesité para comenzar de nuevo.

A mi familia y amigos, por siempre estar a mi lado.

ÍNDICE GENERAL

CAPÍTULO II
PLANTEAMIENTO DEL PROBLEMA

CAPÍTULO III
MATERIALES Y MÉTODOS

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

ÍNDICE DE TABLAS

Pág.

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

Pág.

ÍNDICE DE ACRÓNIMOS

ML Aprendizaje Automático (*Machine Learning*)

LR Regresión Logística (*Logistic Regresion*)

SVM Máquinas de Vectores de Soporte (*Support Vector Machines***)**

NB *Naïve Bayes*

DTC Clasificador de Árboles de Decisión (*Decision Tree Classifier*)

TF-IDF Términos de Frecuencia y Frecuencia Inversa de Documentos (*Term Frequency – Inverse Document Frequency*)

BOW Bolsa de Palabras (*Bag of Words*)

NPL Procesamiento del Lenguaje Natural (*Natural Processing Language*)

RNN Red Neural Recurrente (*Recurrent Neural Network*)

LSTM Red de Memoria a Corto Plazo (*Long Short-Term Memory*)

Bi-LSTM Red de Memoria a Corto plazo Bidireccional (*Bidirectional LSTM*)

HTBSA Análisis Híbrido de Sentimientos Basado en Temas (*Hybrid Topic Based Sentiment Analysis*,)

CNN Redes Neuronales Convolucionales (*Convolutional Neural Networks*)

RESUMEN

El uso de técnicas de análisis de sentimientos para capturar las opiniones de las masas a través de las redes sociales ha aumentado en los últimos años en diferentes áreas como la política, así lo demuestran estudios realizados alrededor del mundo, donde los niveles de asertividad alcanzados en la predicción de intención de voto fueron significativos. Considerando el contexto latinoamericano, como las Elecciones Presidenciales Perú 2021, el estudio se propuso determinar la técnica de *Machine Learning* más asertiva para la predicción de intención de voto aplicada a un modelo de análisis de sentimientos en Twitter. Para ello, se construyó un conjunto de datos denominado Elecciones Bicentenario 2021 Tweets, conformado por 49,916 *tweets* históricos publicados en idioma español, cuyas características fueron extraídas usando TF-IDF, BOW y N-Gramas, por el impacto que tienen en el desempeño del análisis de sentimientos. Luego, se aplicaron algoritmos de clasificación como Regresión Logística, *Naïve Bayes*, Máquinas de Vectores de Soporte, y Árboles de Decisión al modelo propuesto, los cuales fueron evaluados de manera cuantitativa, en términos de exactitud, precisión, exhaustividad y valor-F1. Los resultados electorales y los obtenidos por el modelo coinciden con el sentimiento expresado en las redes sociales en la mayoría de los casos, observándose que Regresión Logística tiene mejor desempeño, alcanzado un 79% de exactitud y precisión, 73% de exhaustividad y 76% de valor-F1. En conclusión, el algoritmo más asertivo para la predicción de intención de voto fue Regresión Logística, seguido por Máquinas de Vectores de Soporte, *Naïve Bayes* y Árboles de Decisión.

Palabras clave: Análisis de sentimientos, aprendizaje automático, intención de voto, redes sociales, tuit, Twitter.

ABSTRACT

The usage of sentiment analysis techniques to capture the opinions of the masses through social networks has increased in recent years in different areas including politics. This is evidenced by worldwide research, where the levels of assertiveness achieved in the prediction of voting intentions were significant. Taking into account the Latin American context, as the Presidential Elections in Perú in the year 2021, the aim of this study was to determine the most accurate Machine Learning technique for the prediction of voting intentions applied to a sentiment analysis model on Twitter. Therefore, a dataset named Elecciones Bicentenario 2021 Tweets consisting of 49,916 historical tweets published in Spanish language has been built. Its features were extracted using TF-IDF, BOW and N-Grams, given the impact they have on the performance of sentiment analysis. Afterwards, classification algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machines, and Decision Trees were applied to the proposed model, which were quantitatively evaluated in terms of accuracy, precision, recall and F1 score. The election results and those obtained by the model concur with the sentiment expressed in social networks for the majority of the cases, finding that the Logistic Regression algorithm performs better. This algorithm reached 79% of accuracy and precision, 73% of recall and 76% of F1 score. To conclude, the most accurate algorithm for the prediction of voting intention was Logistic Regression, respectively followed by Support Vector Machines, Naïve Bayes and Decision Trees.

Keywords: Machine Learning, sentiment analysis, social networks, voting intention, tweet, Twitter.

INTRODUCCIÓN

El análisis de sentimientos y su aplicación en la política ha ido creciendo, demostrando gran utilidad en la predicción política en distintos países alrededor del mundo como Estados Unidos (Chaudhry *et al.*, 2021; Hswen *et al.*, 2020; Liu & Lei, 2018; Qi *et al.*, 2017), Singapur (Choy *et al.*, 2011), Reino Unido (Georgiadou *et al.*, 2020; Gorodnichenko *et al.*, 2021; Mee *et al.*, 2021), India (Joseph, 2019; Khatua *et al.*, 2020; Paul *et al.*, 2021), Malasia y Pakistán (Jaidka *et al.*, 2019), Indonesia (Haryanto *et al.*, 2019), Irlanda (Bermingham & Smeaton, 2011), Australia (Das *et al.*, 2020), Colombia (Cerón-Guzmán & León-Guzmán, 2016) y España (Arcila-Calderón *et al.*, 2017). Por esa razón, las organizaciones políticas han estado usando plataformas como Twitter con más frecuencia en sus campañas de comunicación para canalizar y formar la opinión pública de sus votantes (Paul *et al.*, 2021).

En este sentido, se remarca la importancia que tiene modelar las opiniones políticas en Twitter a través del análisis supervisado de sentimientos (que utiliza procedimientos del aprendizaje automático supervisado o *Supervised Machine Learning*) siendo una metodología complementaria y necesaria para el contraste y predicción de los resultados electorales en países de habla hispana (Arcila-Calderón *et al.*, 2017). Además, en un contexto donde la calidad de las encuestas está siendo constantemente reevaluada por los expertos y la opinión pública, debido a su falta de precisión y poca credibilidad, el análisis de sentimientos basado en técnicas de *Machine Learning* es una herramienta adelantada que puede complementar los análisis tradicionales de muestras representativas para predecir resultados e intención de voto.

En consecuencia, el presente estudio titulado "Intención de voto a través de un modelo de análisis de sentimientos basado en técnicas de *Machine Learning*", pertenece al área de investigación de Ciencias de la Ingeniería, y a la línea de Sistemas, Computación e Informática. Se realizó con el propósito de determinar la técnica de *Machine Learning* más asertiva para un modelo de análisis de sentimientos en Twitter que obtenga los mejores resultados en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021. Para tal efecto, se realizó el estudio y análisis de palabras contenidas en cada opinión publicada en la red social, construyendo un *dataset* con 49,916 registros, y utilizando técnicas de extracción de características (TF-IDF, BOW y N-Gramas) y algoritmos de *Machine Learning* como *Naive Bayes*, Árboles de Decisión y Regresión

Logística para su clasificación, los cuales fueron evaluados a través de métricas como exactitud, precisión, exhaustividad y valor-F1. Con esta información se procuró una solución alternativa para apoyar y/o facilitar las tareas de predicción de intención de voto en procesos electorales.

El presente estudio está organizado en cuatro capítulos. El Capítulo I presenta la revisión bibliográfica utilizada en el marco teórico y los antecedentes de la investigación. El Capítulo II expone la identificación y formulación del problema, la justificación y los objetivos de la investigación. El Capítulo III describe los materiales y métodos utilizados, así como las métricas y detalles sobre el *dataset* conformado. El Capítulo IV muestra los resultados obtenidos en tablas y figuras. Finalmente, se exponen las conclusiones y recomendaciones a considerar.

CAPÍTULO I
REVISIÓN DE LITERATURA

En este capítulo se desarrolla la revisión literaria en relación al análisis de sentimientos en el contexto político, así como los diferentes enfoques y técnicas utilizadas en estudios desarrollados en distintas partes del mundo.

1.1. Marco teórico

1.1.1. Intención de voto

Según Berumen (2022), la intención de voto es la respuesta simple y directa que los encuestados dan cuando se les pregunta por qué candidato o por cuál partido votarán en los comicios. Básicamente, se resume en hacerle una pregunta al elector potencial: "Si hoy fueran las elecciones ¿por quién votaría?"

En el contexto peruano, y según la Ley Orgánica de Elecciones, Ley N° 26859, el Presidente de la República es elegido por sufragio directo, considerándose aquel candidato que obtiene más de la mitad de los votos, donde no se computan los votos viciados o en blanco (Congreso Constituyente Perú, 1997). En consecuencia, cabe señalar que existen tres tipos de votos:

- **Voto válido.** Cuando el elector marca con un aspa o cruz sobre la fotografía del candidato a la Presidencia de la República, se considera un voto válido (Congreso Constituyente Perú, 1997).

- **Voto viciado.** Cuando el elector se completa la cédula de sufragio de manera inválida, es decir, si se hace una marca diferente a un aspa o cruz o cuando la intersección de éstas se encuentra fuera del recuadro, se considera un voto nulo o viciado. También sucede cuando se dibuja o escribe un mensaje en la cédula (El Comercio, 2021).

- **Voto en blanco.** Cuando el elector deja la cédula sin marca alguna y la coloca dentro de la urna, se considera un voto en blanco (El Comercio, 2021).

Entonces, cuando se pregunta por la intención de voto a los electores, y no se tenga una respuesta por algún candidato en específico, formaría parte de los indecisos, los que no saben o no contestan (Berumen, 2022), por lo que su voto podría también ser en blanco o viciado.

En el Perú, la intención de voto se dividió entre diferentes candidatos presidenciales que representaron a sus organizaciones políticas en las elecciones generales realizadas el 11 de abril de 2021 (ONPE, 2021).

ORGANIZACIONES POLÍTICAS Y CANDIDATOS PRESIDENCIALES EN LAS ELECCIONES GENERALES PERÚ 2021					
		PARTIDO NACIONALISTA PERUANO			PARTIDO POPULAR CRISTIANO - PPC
		EL FRENTE AMPLIO POR JUSTICIA, VIDA Y LIBERTAD			FUERZA POPULAR
		PARTIDO MORADO			UNION POR EL PERU
		PERU PATRIA SEGURA			RENOVACION POPULAR
		VICTORIA NACIONAL			RENACIMIENTO UNIDO NACIONAL
		ACCION POPULAR			PARTIDO DEMOCRATICO SOMOS PERU
		AVANZA PAIS - PARTIDO DE INTEGRACION SOCIAL			PARTIDO POLITICO NACIONAL PERU LIBRE
		PODEMOS PERU			DEMOCRACIA DIRECTA
		JUNTOS POR EL PERU			ALIANZA PARA EL PROGRESO

Figura 1. Organizaciones políticas en las elecciones generales Perú 2021

Fuente: Adaptado de (ONPE, 2021)

1.1.2. Redes Sociales y Política

Las redes sociales juegan un papel cada vez más importante en el contexto político, ya que permiten y favorecen el diálogo entre los candidatos y los electores, e incrementan la participación política; haciendo posible manifestar opiniones y preferencias sobre un candidato o grupo político específicos mediante publicaciones o valoraciones en redes sociales, generando mayor alcance y haciendo que esta forma de expresión sea importante para los políticos, más aun tratándose del valor emocional y sentimental que se expresan en relación a ellos (Bohorquez *et al.*, 2019).

Las redes sociales permiten que las personas puedan crear, compartir e intercambiar sus sentimientos, pensamientos, ideas, críticas, opiniones, expectativas, información, videos, imágenes y otros tipos de contenido digital en diferentes plataformas virtuales como Facebook, Twitter, LinkedIn, Instagram y muchos otras (Ansari *et al.*, 2021; Rodriguez-Ibanez *et al.*, 2021), mejorando la democracia deliberativa entre los votantes o electores, haciendo que puedan perfeccionar sus propias opiniones, escuchar diferentes opiniones, e identificar fines comunes, aunque también podría generar mayor división entre ellos (Aramburo *et al.*, 2022; Grover *et al.*, 2019).

Por otro lado, a diferencia de los medios tradicionales que siguen un modelo de comunicación unidireccional y ofrece comunicaciones asincrónicas, las redes sociales, tienen un tipo de comunicación multidireccional e interactiva, lo que facilita que el discurso político pase de los medios de comunicación tradicionales a plataformas de redes sociales como Facebook y Twitter, haciendo que los políticos puedan distribuir información relacionada a su campaña, su agenda política y también, como una forma de conectar a ciudadanos involucrados en el discurso político con aquellos que no (Grover *et al.*, 2019).

En este contexto, Twitter, una plataforma ampliamente utilizada para establecer contactos y microblogging donde los usuarios publican mensajes en forma de *tweets* que pueden tener un máximo de 280 caracteres, se ha usado también como un medio para expresar opiniones relacionadas a la política y conectarse con la gente (Ansari *et al.*, 2020), convirtiendo a esta plataforma en una excelente herramienta para descubrir información imparcial a partir del contenido generado por los usuario, e incrementando a los políticos, sus posibilidades de ganar las elecciones (Grover *et al.*, 2019).

1.1.3. Análisis de sentimientos en el contexto político

El análisis de sentimientos es una técnica que permite conocer las opiniones de las personas en diferentes áreas como negocios, economía, investigación, gobierno, y política, lo que afecta en gran medida la toma de decisiones, y que, a su vez, puede clasificarse en técnicas basadas en *lexicons*, *machine learning*, y enfoques híbridos (Rodriguez-Ibanez *et al.*, 2021).

Chaudry *et al.* (2021), explican que el análisis de sentimientos es un proceso que automatiza la extracción de actitudes, opiniones, puntos de vista y emociones a partir de fuentes de texto, voz, *tweets* y bases de datos a través del Procesamiento del Lenguaje Natural.

Para Seckin y Kilimci (2020), el análisis de sentimientos es la acción de explicar el significado de las emociones como positivas, negativas o neutrales a través de diversos métodos y materiales de minería de texto. También se define como un enfoque de estudio de datos textuales a gran escala empleada en la investigación en diferentes áreas como en la comunicación política, con el objetivo de reconocer y evaluar el valor emocional que existe detrás de los textos analizados y clasificados en positivos, negativos o neutros (Arcila-Calderón *et al.*, 2017).

En este contexto, y considerando además que, el proceso de elecciones es un componente clave para cualquier país democrático que expresa su opinión a través de un voto, esa misma opinión puede expresarse redes sociales y estudiarse a través del análisis de sentimientos (Chaudhry *et al.*, 2021). Además, con el análisis de sentimientos es posible determinar si un texto presenta una connotación positiva, negativa o neutral; de tal forma que conlleva a la identificación de expresiones de sentimientos, polaridad, fuerza de las expresiones, y su relación con el sujeto estudiado (Bohorquez *et al.*, 2019). Esto se ha demostrado en distintos estudios en relación a diversos procesos electorales alrededor del mundo, tales como Estados Unidos (Chaudhry *et al.*, 2021; Hswen *et al.*, 2020; Liu & Lei, 2018; Qi *et al.*, 2017), Singapur (Choy *et al.*, 2011), Reino Unido (Georgiadou *et al.*, 2020; Gorodnichenko *et al.*, 2021; Mee *et al.*, 2021), India (Joseph, 2019; Khatua *et al.*, 2020; Paul *et al.*, 2021), Malasia y Pakistán (Jaidka *et al.*, 2019), Indonesia (Haryanto *et al.*, 2019), Irlanda (Bermingham & Smeaton, 2011), Australia (Das *et al.*, 2020), Colombia (Cerón-Guzmán & León-Guzmán, 2016) y España (Arcila-Calderón *et al.*, 2017). Donde se ha empleado análisis de sentimientos para clasificar y capturar la orientación y polaridad política de los usuarios, utilizando diferentes técnicas y enfoques.

El análisis de sentimientos es un concepto amplio que contempla diferentes tareas, enfoques y niveles de análisis (Britzolakis *et al.*, 2021; Ligthart *et al.*, 2021; Pozzi *et al.*, 2017).

1.1.4. Tareas del análisis de sentimientos

Según Ligthart *et al.* (2021), son cinco las tareas principales que realiza el análisis de sentimientos. Estas son:

- **Clasificación de sentimientos:** Dentro de esta tarea se encuentra la clasificación de la polaridad como positiva, negativa o neutral (Chaudhry *et al.*, 2021; Jaidka *et al.*, 2019). También están las subtareas de clasificación *cross-domain* y *cross-language* cuyo objetivo es transferir conocimiento a partir de un dominio rico en datos a un dominio destino donde los datos y las etiquetas son limitados (Peng *et al.*, 2018).

- **Clasificación de subjetividad:** Se utiliza para determinar la existencia de subjetividad en el texto dado, específicamente pista o palabras, y éstas son usadas para clasificar el texto como subjetivo u objetivo (Medhat *et al.*, 2014).

- **Detección de *spam*:** Debido al crecimiento constante de sitios web de comercio electrónico es que, la detección de spam se ha convertido en un tema destacado en el análisis de sentimientos. Básicamente, se busca identificar tres tipos de características relacionadas a las reseñas falsas: contenido, metadatos y conocimiento de la vida real sobre un producto (Cerón-Guzmán & León-Guzmán, 2016).

- **Detección de lenguaje implícito:** El lenguaje implícito se refiere al humor, sarcasmo e ironía, y cuando se utiliza es difícil comprender el mensaje, por lo que hay vaguedad y ambigüedad en él. Sin embargo, un significado implícito en una oración puede cambiar completamente su polaridad. El objetivo de esta subtarea es comprender los hechos relacionados con determinado evento (Liu & Zhang, 2012).

- **Extracción de aspectos:** Se refiere a la recuperación de la entidad destino y sus aspectos en el documento. La entidad destino puede ser un producto, una persona, un evento, una organización, etc. Esta tarea se aplica especialmente en el análisis de sentimientos de redes sociales y blogs donde usualmente no hay temas predefinidos (Liu & Zhang, 2012; Medhat *et al.*, 2014).

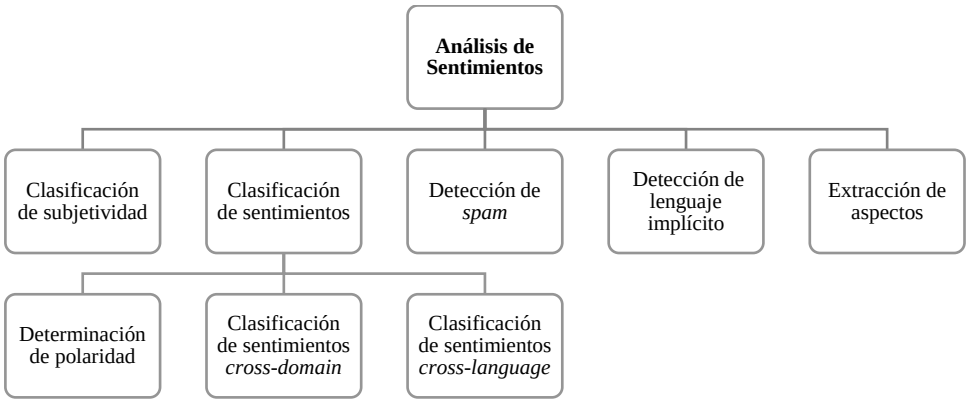

Figura 2. Tareas del análisis de sentimientos

Fuente: Adaptado de (Britzolakis *et al.*, 2021; Ligthart *et al.*, 2021; Pozzi *et al.*, 2017)

1.1.5. Enfoques del análisis de sentimientos

Las técnicas de análisis de sentimientos se dividen en tres categorías diferentes: *Machine Learning* que incluye modelos supervisados y no supervisados, modelos basados en *lexicons*, ya sea utilizando diccionarios o cuerpos de conocimiento, y enfoques híbridos, que integran las técnicas mencionadas (Rodriguez-Ibanez *et al.*, 2020, 2021). Los distintos enfoques y los algoritmos más utilizados para el análisis de sentimiento se muestran en la Figura 3.

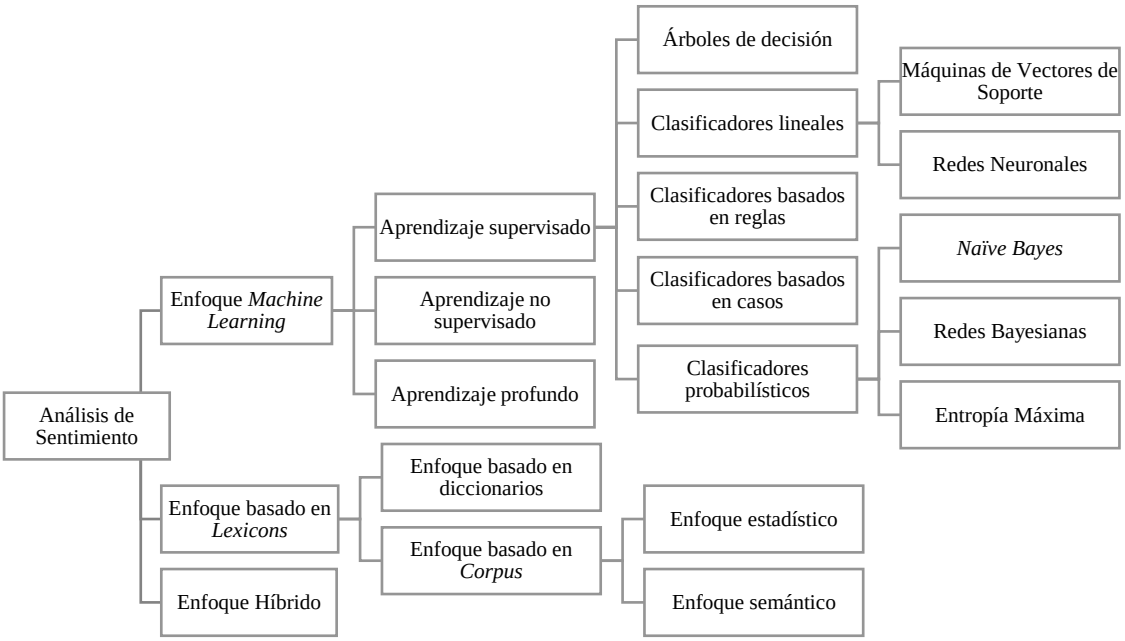

Figura 3. Enfoques y técnicas de análisis de sentimiento

Fuente: Adaptado de (Medhat *et al.*, 2014; Rodriguez-Ibanez *et al.*, 2020, 2021)

1.1.3.1. *Machine Learning*

Este enfoque aplica algoritmos ML y características sintácticas y/o lingüísticas para la clasificación regular de texto (Medhat *et al.*, 2014). Asimismo, se subdivide en técnicas de aprendizaje supervisado (del inglés *Supervised Learning*) y no supervisado (del inglés *Unsupervised Learning*) para construir un modelo a partir de datos de entrenamiento. En el aprendizaje supervisado, se aprende una función (posiblemente no lineal) mediante el mapeo de pares de entrada y salida a través del uso de datos etiquetados para dirigir el proceso de aprendizaje. Por otro lado, el aprendizaje no supervisado, describe inferencias a partir de *datasets* no etiquetados, cuando es difícil encontrar datos de entrenamiento etiquetados (Paliwal *et al.*, 2018).

Los algoritmos ML no dependen de reglas creadas manualmente, sino de procedimientos de aprendizaje automático. Una tarea de análisis de sentimiento normalmente se modela como un problema de clasificación, donde el clasificador proporciona datos tipo texto y devuelve una clase de tipo positiva, negativa o neutral (Sudhir & Suresh, 2021).

Definición del problema de clasificación de texto: Sea un conjunto de datos de entrenamiento $D = \{X_1, X_2, \dots, X_n\}$ donde cada dato es etiquetado para una clase. El modelo de clasificación se relaciona con las características del dato subyacente a una de las etiquetas de clase. Después, para cada instancia de clase desconocida, el modelo es utilizado para predecir una etiqueta de clase para ella. El problema de clasificación tipo *hard* se da cuando una etiqueta solo se asigna una instancia. En cambio, el problema de clasificación tipo *soft* se da cuando un valor probabilístico de las etiquetas es asignado a una instancia (Medhat *et al.*, 2014).

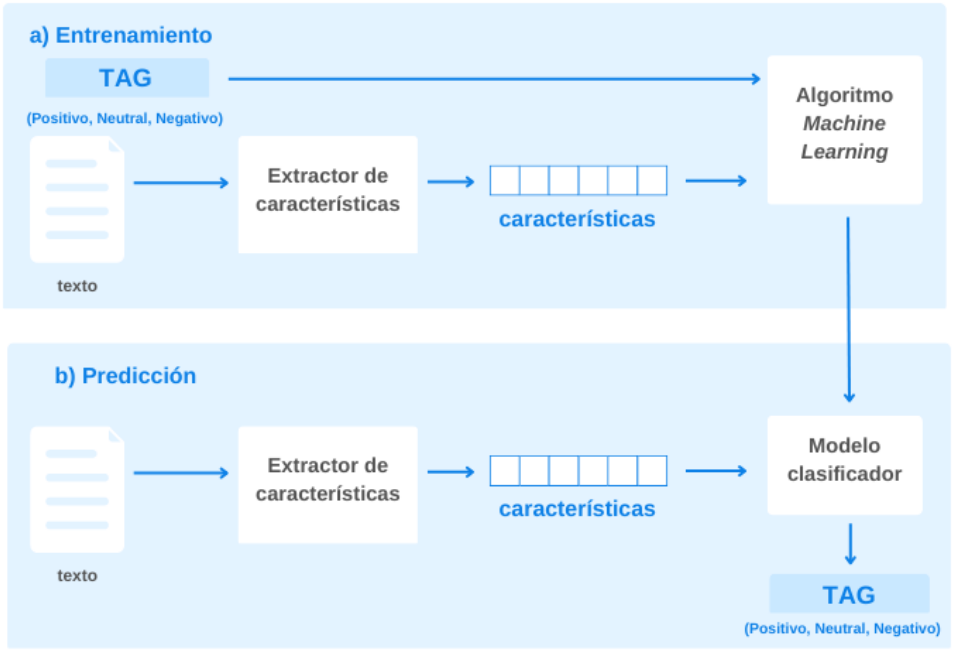

Figura 4. Enfoque *Machine Learning*

Fuente: Adaptado de (Sudhir & Suresh, 2021)

En la etapa de entrenamiento 4a, el extractor de características transforma la entrada textual en un vector de características. Posteriormente, se entregan etiquetas (*tag*) de características y pares de vectores al algoritmo para producir un modelo. En la etapa de predicción 4b, el extractor de características transforma las entradas de texto ocultas en vectores de características. Luego, los vectores son entregados al modelo, generando etiquetas de predicción para los vectores correspondientes (Sudhir & Suresh, 2021).

1.1.3.2. Modelos basados en *lexicons*

Llamados también modelos simbólicos (Abdullah & Hadzikadic, 2018), los cuales usan un diccionario de sentimientos (*lexicons*) para determinar la polaridad, y los valores son asignados a palabras que reflejan una actitud positiva, negativa o neutral, mediante la suma de los valores de polaridad ponderada de los términos (Arcila-Calderón *et al.*, 2017; Aung & Myo, 2017). En este contexto, las listas de palabras y su valor sentimental asociado se conocen comúnmente como léxicos de sentimiento, donde se describen diferentes estrategias como el proceso seguido para crear el *lexicon*, ya sea manual o automático (Rodriguez-Ibanez *et al.*, 2021). Otro enfoque es el

basado en diccionarios, donde un conjunto de palabras es recolectado de forma manual con orientaciones conocidas, para luego incrementar ese conjunto mediante la búsqueda de sinónimos y antónimos en bases de datos como WordNet o Thesaurus, hasta que se encuentren palabras nuevas, las cuales se agregan a una lista de palabras semilla. Finalmente, el enfoque basado en *Corpus*, que ayuda a resolver el problema de encontrar palabras con orientación y contexto específico. Además, este enfoque depende de los patrones sintácticos o patrones que se presentan junto a una lista inicial de palabras semilla para encontrar otras dentro de un gran *corpus*. Este enfoque utiliza a su vez dos enfoques: el estadístico y el semántico (Medhat *et al.*, 2014).

1.1.3.3. Modelos híbridos

Este enfoque combina otros diferentes para mejorar la exactitud de la predicción, tales como técnicas basadas en *lexicons* y *machine learning* (Rodriguez-Ibanez *et al.*, 2020, 2021), *o lexicons* y representación semántica de palabras (Chaudhry *et al.*, 2021). Por lo que, es muy común que los *lexicons* desempeñen un papel clave en la mayoría de métodos de trabajo híbridos (Medhat *et al.*, 2014).

La idea principal en la que se basan estos modelos es que la combinación de técnicas conduce a una mejor categorización y evaluación comparativa de resultados (Rodriguez-Ibanez *et al.*, 2021).

1.1.6. Características del análisis de sentimientos

1.1.6.1. Categorización de sentimientos

Cuando se realiza análisis de sentimientos lo primero es categorizar las oraciones como subjetivas u objetivas. Si son objetivas, no serán necesarias tareas adicionales en su análisis, mientras que, si son subjetivas, es necesario estimar su polaridad. La **clasificación de subjetividad** consiste en diferenciar oraciones que expresan información objetiva (oraciones objetivas) de oraciones que expresan puntos de vista y opiniones subjetivas (oraciones subjetivas) (Pozzi *et al.*, 2017).

Un ejemplo de oración objetiva sería "*Keiko Fujimori es una mujer*", en cambio un ejemplo de oración subjetiva sería "*Keiko Fujimori es excelente candidata presidencial*".

Por otro lado, la **clasificación de polaridad** consiste en diferenciar oraciones que expresan polaridades positivas, negativas o neutrales. Asimismo, una oración subjetiva puede no expresar ningún sentimiento positivo o negativo, por ejemplo *"Supongo que ella ha ganado",* se clasificaría como neutral (Pozzi *et al.*, 2017).

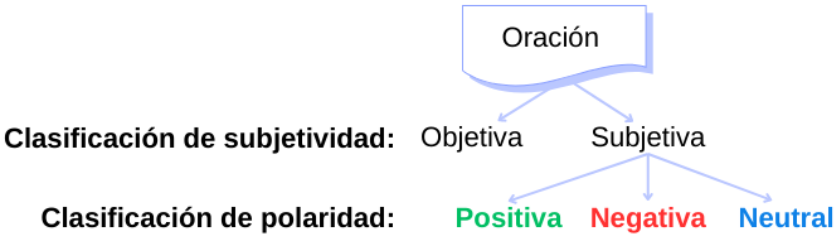

Figura 5. Flujo de trabajo de la categorización de sentimientos

Fuente: Adaptado de (Pozzi *et al.*, 2017)

1.1.6.2. Niveles de análisis

El análisis de sentimientos se puede implementar en tres niveles: documento, oración y aspectos (Britzolakis *et al.*, 2021; Liu & Zhang, 2012; Paul *et al.*, 2021; Pozzi *et al.*, 2017).

- **Nivel de documento:** Consiste en clasificar la polaridad de todo el texto del documento, determinando si el texto expresa una opinión positiva, negativa o neutral (Britzolakis *et al.*, 2021; Paul *et al.*, 2021; Pozzi *et al.*, 2017).

- **Nivel de oración:** Consiste en determinar la polaridad de cada oración contenida en un documento. Donde, dado un documento, se asume que cada oración, denota una sola opinión sobre una sola entidad (Liu & Zhang, 2012; Paul *et al.*, 2021; Pozzi *et al.*, 2017).

- **Nivel de aspecto y entidad:** Realiza un análisis más detallado que los niveles de documento y oración, basándose en la idea de que una opinión

consta de un sentimiento y un objetivo (Britzolakis *et al.*, 2021; Pozzi *et al.*, 2017).

Figura 6. Niveles de análisis

Fuente: Adaptado de (Pozzi *et al.*, 2017)

1.1.7. Técnicas de *Machine Learning* utilizadas en el análisis de sentimientos

Se define al *Machine Learning* o Aprendizaje Automático como una técnica de gran importancia para la resolución de diferentes problemas con amplia aplicabilidad en áreas como en la gestión y análisis de texto. Como se menciona líneas arriba, esta técnica se clasifica en dos tipos: aprendizaje supervisado y aprendizaje no supervisado (Zhai & Massung, 2016). En el primero, se entrena el algoritmo a partir de datos que han sido previamente etiquetados de manera manual y, mientras más grande sea el conjunto de datos entrenados, mayor es la eficacia del algoritmo. En el segundo, el algoritmo utiliza un conjunto de datos que no tienen etiqueta; entonces, no se le dice al algoritmo lo que representan los datos. El objetivo es que el algoritmo pueda encontrar por sí solo patrones que ayuden a entender el conjunto de datos (Chamorro Alvarado, 2018).

La clasificación de sentimientos obviamente puede ser formulada como un problema de aprendizaje supervisado con tres clases: positiva, negativa y neutral. Los métodos de aprendizaje supervisado existentes que pueden aplicarse al análisis de sentimiento más comúnmente utilizados son (Figueroa, 2018; Liu, 2011; Liu & Zhang, 2012):

1.1.7.1. *Naïve Bayes* (NB)

Es el clasificador más simple y comúnmente usado (Medhat *et al.*, 2014). Este algoritmo se desempeña con rapidez y exactitud, por lo que es aplicado en bases de datos diversas y de gran tamaño, además, solo se requiere una pequeña cantidad de datos de entrenamiento para determinar los parámetros estimados necesitados en el proceso de clasificación (Ansari *et al.*, 2021; Chaudhry *et al.*, 2021; Haryanto *et al.*, 2019; Joseph, 2019).

El clasificador *Naïve Bayes* computa la probabilidad posterior de una clase, basado en la distribución de palabras del documento, utilizando bolsa de palabras como técnica de extracción de características, la que permite ignorar la posición de una palabra. Los clasificadores probabilísticos que conforman este clasificador están basados en el teorema de Bayes para predecir la probabilidad de que un determinando conjunto de características pertenezca a una etiqueta en particular (Ansari *et al.*, 2021; Medhat *et al.*, 2014).

$$P(etiqueta|características) = \frac{P(etiqueta)*P(características|label)}{P(características)} \quad (1)$$

Donde, $P(etiqueta)$ representa la probabilidad anterior de una etiqueta o la probabilidad de que una característica aleatoria establezca la etiqueta. $P(características|label)$ representa la probabilidad anterior de que dado un conjunto de características se clasifique como una etiqueta. $P(características)$ representa la probabilidad anterior de que se produzca un determinado conjunto de características (Medhat *et al.*, 2014).

Aplicando la teoría de probabilidad total e independencia condicional, la ecuación podría reescribirse de la siguiente forma:

$$P(etiqueta|características) = \frac{P(etiqueta)*P(f1|etiqueta)*\ldots*P(fn|etiqueta)}{P(características)} \quad (2)$$

1.1.7.2. Regresión Logística (LR)

Es un método de clasificación muy común que tiene muchas aplicaciones, como la clasificación de texto, visión computacional, etc. El algoritmo calcula la probabilidad de una muestra que pertenece a una clase particular, aprendiendo del conjunto de parámetros que minimizan la probabilidad logarítmica inducida sobre los ejemplos de entrenamiento (Ansari *et al.*, 2020). Para el caso de la clasificación de texto, el algoritmo de regresión logística permite predecir la probabilidad que tiene un texto de ser positivo o negativo, dada la etiqueta y el vector característico de valores TF-IDF, encontrando los parámetros más adecuados para cada texto (Fernández, 2019).

1.1.7.3. Máquinas de Vectores de Soporte (SVM)

Es un algoritmo derivado de la teoría de aprendizaje estadístico, el cual se basa en el principio de minimización de riesgo estructural, y es utilizado convencionalmente como método de aprendizaje supervisado para la clasificación binaria, regresión y detección de valores atípicos. Fue desarrollado por Vapnik en 1995 con el objetivo de reducir el error de prueba y la complejidad computacional, identificando un hiperplano de separación óptima entre las dos clases de un conjunto de datos de entrenamiento. Esta separación la obtiene el hiperplano que tiene la mayor distancia del conjunto de datos de entrenamiento más cercano. (Ansari *et al.*, 2020, 2021).

Este algoritmo examina los datos e identifica patrones para la categorización. Su metodología consiste en transformar los datos tipo texto en pesos, para luego fusionarlos y formar valores TD-IDF, simplemente multiplicándolos colectivamente. Asimismo, las oraciones positivas o negativas pueden ser determinadas calculando el hiperplano mediante las ecuaciones 3 y 4 (Sudhir & Suresh, 2021).

$$f\big(\emptyset(x)\big) = sign(w.\emptyset(x) + b) \qquad (3)$$

Donde, $f(\emptyset(x))$ representa el resultado de las categorías de los datos de prueba, w representa los pesos, b representa el sesgo, $\emptyset(x)$ representa los datos de prueba de los cálculos del núcleo.

$$K(x, x_i) = (x.x_i + 1)^2 \qquad (4)$$

Donde, $K(x, x_i)$ representa al núcleo, x representa a los datos de prueba, x_i representa a los datos de entrenamiento para i. Entonces, al determinar el hiperplano, el cálculo de los datos de prueba se realiza en función de los pesos de los datos de prueba considerando clases positivas o negativas.

1.1.7.4. Clasificador de Árboles de Decisión (DTC)

Es una herramienta de soporte a las decisiones que utiliza un grafo en forma de árbol o modelo de decisiones y sus posibles caminos, incluidos los resultados de eventos fortuitos, costos de recursos y utilidad. Realiza una partición binaria recursiva del espacio de características para dividir un conjunto de datos en subconjuntos cada vez más pequeños mientras desarrolla simultáneamente un árbol de decisión incremental, dando como resultado un árbol con nodos de decisión y hoja. Un nodo de decisión tiene dos o más ramas; un nodo hoja o nodo terminal representa decisiones o etiquetas clase, y el nodo superior es el nodo raíz. Luego se encuentra el árbol más pequeño que se ajusta a los datos. Por lo general, este es el árbol que produce el error de validación cruzada más bajo (Ansari *et al.*, 2020; Sudhir & Suresh, 2021).

Este algoritmo puede usarse para clasificación como para regresión, asimismo puede trabajar con datos categóricos como numéricos. Por lo que pueden utilizarse el índice de Gini (del inglés *Gini index*) y el parámetro de ganancia de información (del inglés *Information gain*) para decidir qué atributo o característica se utilizará para una mayor división del conjunto de datos (Ahuja *et al.*, 2019). La ganancia de información se determina en términos de entropía, la cual mide la impureza de un conjunto de datos. La ecuación para determinar la entropía se presenta en la ecuación 5 (Sudhir & Suresh, 2021).

$$E(D) = \sum_{i=1}^{n} -p_{c(i)}\left(p_{c(i)}\right) \qquad (5)$$

Donde, $p_{c(i)}$ es la probabilidad de la clase $c(i)$ en un nodo. $E(D)$ o llamado también entropía de D es la medida del desorden de los ejemplos considerados.

Por otro lado, la ecuación para determinar el índice de Gini se muestra en la ecuación 6 (Long, 2015).

$$Gini(D) = \sum_{i=1}^{n} -p_{c(i)}\left(p_{c(i)}\right) \qquad (6)$$

1.1.8. Proceso de análisis de sentimientos

El análisis de sentimientos requiere la aplicación de algoritmos de Procesamiento del Lenguaje Natural en el nivel inicial para procesar el texto, esencialmente para hacer que las palabras tengan sentido y puedan ser comprendidas (Paul *et al.*, 2021), permitiendo obtener el significado existente en el lenguaje humano en general, para luego realizar el análisis del texto previamente procesado con técnicas de *Machine Learning*, para la extracción de nuevas variables y la relación de los datos, atribuyendo un valor agregado real a la información que permita la toma de decisiones (Aramburo *et al.*, 2022). A todo esto, el análisis de sentimientos se enfoca en la detección de estados emocionales de un texto dado y junto a la minería de opinión se categorizan como campos de la Minería de Texto, que es el proceso de obtención de información sobre datos no estructurados. A pesar de ser similares y estar muy relacionadas, el análisis de sentimientos y la minería de opinión se diferencian significativamente. El análisis de sentimientos busca identificar palabras o expresiones que indiquen una emoción en un texto dado, mientras que la minería de opinión busca extraer y analizar los pensamientos de las personas sobre una entidad o evento a partir de un texto (Britzolakis *et al.*, 2021).

La adaptación realizada por (Fernández, 2019), basada en (Liu, 2011; Liu & Zhang, 2012) señala los pasos a seguir para el proceso de análisis de sentimientos, el cual parte de la minería de opinión.

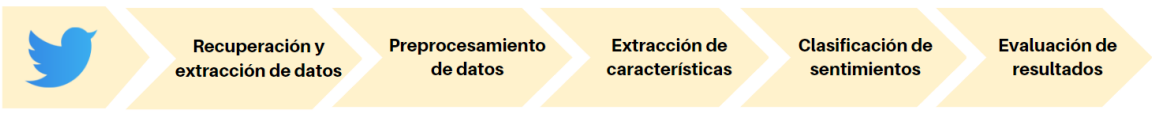

Figura 7. Esquema del análisis de sentimientos

Fuente: Adaptado de (Liu, 2011; Liu & Zhang, 2012)

1.1.8.1. Recuperación y extracción de datos

Se refiere a la recuperación y extracción de datos tipo texto para el proceso de análisis de sentimientos, la que puede hacerse desde diferentes fuentes como lo son las redes sociales, a través de herramientas *crawling* y APIs que permiten obtener datos desde sus plataformas. (Fernández, 2019). Esta tarea también puede realizarse a través de herramientas *scraping* o *scrapers*, las cuales no necesitan utilizar el API de Twitter (Aramburo *et al.*, 2022).

Un *scraper* muy usado es ***snscrape***, que es una herramienta de arrastre o recuperación de datos para servicios de redes sociales (SNS). Arrastra datos como perfiles de usuario, *hashtags* o búsquedas y devuelve los elementos descubiertos.

1.1.8.2. Preprocesamiento de datos

Se refiere al proceso de limpieza y normalización del texto o simplemente el tratamiento previo de los datos que se utilizará para el análisis. En este proceso se realiza el filtrado de datos a través de librerías de Python para recuperar las partes más importantes y significativas de los *tweets* excluyendo contenido innecesario (Chaudhry *et al.*, 2021).

Los pasos considerados en el pre procesamiento incluyen: (1) la filtración y limpieza de la *RawData* (datos en crudo o sin procesar), eliminando duplicados para mantener la unicidad de los *tweets*, y convirtiendo la *RawData* en *DataFrame* (datos procesados y organizados). (2) La normalización del texto, eliminando signos de puntuación, URLs, espacios múltiples y caracteres especiales. (3) La tokenización, dividiendo el texto en las unidades mínimas que lo conforman, siendo estas las palabras. (4) Eliminación de palabras vacías (del inglés *stop words*), ya que no contribuyen al análisis de sentimiento. (Ahuja *et al.*, 2019; Chaudhry *et al.*, 2021; Fernández, 2019; Mee *et al.*, 2021).

1.1.8.3. Extracción de características

Consiste en transformar el texto en un vector de características, pues los algoritmos de análisis de texto requieren que el texto esté representado de

forma numérica, después se crea un vocabulario que será calificado o puntuado (Cerón-Guzmán & León-Guzmán, 2016; Fernández, 2019). Para esta tarea se utiliza el modelo de bolsa de palabras (BOW), donde las palabras son unidades básicas de representación y el orden de las palabras es ignorado (aunque se retienen los recuentos de palabras) (Zhai & Massung, 2016). Otro modelo más sofisticado es el de N-gramas, que consiste en crear un vocabulario en grupos o secuencias de *n* palabras que expande el ámbito del vocabulario y permite capturar mayor significado de los documentos, donde cada palabra o *token* se denomina "grama", y *n* toma valores de *1*, *2*, *3*, etc., llamándose unigrama si *n* es *1*, bigrama si *n* es *2*, y así sucesivamente (Ahuja *et al.*, 2019). También puede utilizarse la función matemática *hash* que mapea los datos a un tamaño fijo de conjunto de números; así como el término de frecuencia TF-IDF, el cual consiste en evaluar la importancia de una palabra en un documento y asignarle un valor (Mee *et al.*, 2021), obteniendo las palabras útiles y sus puntuaciones a partir del corpus dado. Donde, TF representa la frecuencia de veces que una palabra ha aparecido en el corpus, y DF describe cuántos documentos contienen un término específico. IDF es el inverso multiplicativo de DF; junto con TF, proporciona una medida de la aparición de ciertas palabras (Ahuja *et al.*, 2019; Chaudhry *et al.*, 2021). El término frecuencia $tf(i,\delta)$ se da en la siguiente Ecuación (7):

$$tf(i,\delta) = \frac{f_\delta(i)}{max_{w \in d} f_\delta(w)} \qquad (7)$$

Donde, $f_\delta(i)$ es el término de frecuencia de i en el documento δ , mientras que $f_\delta(w)$ es el total de palabras en el documento δ. De forma similar, el IDF de la i-ésima palabra Δ puede expresarse en la Ecuación (8):

$$idf(i,\Delta) = ln\left(\frac{|\Delta|}{|\gamma|}\right) \qquad (8)$$

$$\gamma = \delta \in \Delta : i \in \delta$$

Donde, Δ es el total de documentos y γ representa los documentos con el término i. En algunos casos, como el sarcasmo, TF-IDF podría no ser el más adecuado, ya que la puntuación de frecuencia podría mostrar sentimientos equivocados (Chaudhry *et al.*, 2021).

1.1.8.4. Clasificación de sentimientos

Consiste en procesar cada texto mediante un clasificador y asignar una polaridad, polaridad positiva, negativa o neutral (Mohamed *et al.*, 2020). Esto puede hacerse a través de enfoques *Machine Learning que incluye* modelos supervisados y no supervisados, enfoques basados en *lexicons*, ya sea utilizando diccionarios o cuerpos de conocimiento, y enfoques híbridos, que integran las técnicas mencionadas (Britzolakis *et al.*, 2021; Rodriguez-Ibanez *et al.*, 2021).

1.1.8.5. Evaluación de resultados

Para evaluar la efectividad de los algoritmos *Machine Learning* aplicados al análisis de sentimientos se utilizan diferentes métricas, como: exactitud (*accuracy*), precisión (*precision*), exhaustividad (*recall*) y puntuación F1 (*F1 score*) (Ahuja *et al.*, 2019; Liu, 2011). Para ello es conveniente introducir la terminología necesaria para esas métricas utilizando una matriz de confusión (Liu, 2011; Long, 2015).

		Predicción de clase	
		Positivo	Negativo
Clase actual	Positivo	TP	FN
	Negativo	FP	TN

Figura 8. Matriz de confusión

Fuente: Adaptado de (Long, 2015)

Donde:

- **Verdaderos positivos (TP)**, de inglés *true positives*, es el número de instancias positivas que el clasificador ha identificado correctamente como positivas.
- **Falsos positivos (FP)**, del inglés *false positives*, es el número de instancias que el clasificador ha identificado como positivas, pero en realidad son negativas.

- **Verdaderos negativos (TN)**, del inglés *true negatives*, es el número de instancias negativas que el clasificador ha identificado correctamente como negativas.

- **Falsos negativos (FP)**, del inglés *false negatives*, es el número de instancias clasificadas como negativas, pero en realidad son positivas.

Exactitud (*accuracy*), es aquella métrica que indica la proporción que un modelo ha alcanzado al clasificar los registros correctamente. Se define como la suma de TP y TN dividida por el número total de instancias.

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

Precisión (*precision*), es el porcentaje de instancias marcadas como positivas que realmente lo son.

$$Precisión = \frac{TP}{TP + FP}$$

Exhaustividad (*recall*), es el porcentaje de instancias positivas que fueron correctamente identificadas.

$$Exhaustividad = \frac{TP}{TP + FN}$$

Valor-F1 (*F1 score*), es la media armónica de la precisión y la exhaustividad.

$$F1 = \frac{2Precisión * Exhaustividad}{Precisión + Exhaustividad}$$

1.2. Antecedentes

A continuación, se presenta el estado del arte en relación al análisis de sentimientos en Twitter en el contexto político, entre ellos destacan diversos artículos en el ámbito internacional, nacional y local:

Aramburo *et al.* (2022), realizan análisis de sentimientos sobre las opiniones publicadas en Twitter acerca del Paro Nacional 2021 en Cali-Colombia. Se consideraron los *tweets* publicados entre marzo y agosto de 2021 para analizar los sentimientos, reacciones y

percepciones de los usuarios a través de técnicas basadas en *Text Mining* y el diccionario *NRC lexicon*, con el objetivo de establecer patrones de conducta que pueden influir de manera negativa, manera positiva o neutral ante determinado hito social. Asimismo, se presenta un modelo conformado por cinco fases: identificación de los datos, extracción de datos, pre-procesamiento de datos, procesamiento de datos y resultados. Los resultados mostraron que, de un total de 404,572 *tweets*, 4,907 *tweets* representaron el mayor número de interacciones el 4 de mayo de 2021 debido a actividades confrontacionales. De la misma forma, se identificó el crecimiento de emociones como la tristeza, ira, miedo y desagrado durante el suceso, y el decrecimiento de emociones como la confianza y la alegría. Después del suceso, se observó el incremento de la confianza y la alegría, pero no en los niveles mostrados antes del suceso. Además, el miedo permaneció en segundo lugar después de la confianza.

Chaudry *et al.* (2021), enfocan su estudio en las elecciones presidenciales de Estados Unidos 2020 donde salió victorioso Joe Biden, utilizando el análisis de sentimientos en Twitter para determinar las opiniones del público antes, durante y después de las elecciones, y posteriormente, compararlas con los resultados reales. También se compara las opiniones de las elecciones de 2016 en las que Donal Trump ganó. Para ello se creó un *dataset* utilizando la API de Twitter, se preprocesaron los datos, se extrajeron las características correctas usando TF-IDF, una técnica conocida en el Procesamiento de Lenguaje Natural, y se aplicó el clasificador *Naïve Bayes* para obtener opiniones públicas, revelándose que los resultados electorales coinciden con el sentimiento expresado en las redes sociales en la mayoría de casos. Los resultados del análisis de sentimiento antes y después de las elecciones demuestran la deriva sentimental de los valores atípicos. Asimismo, el clasificador de sentimientos muestra una exactitud del 94,58% y una precisión del 93,19%.

Paul *et al.* (2021), aplican el análisis de sentimientos demostrando que las organizaciones políticas utilizan cada vez más plataformas como Twitter en sus campañas de comunicación para canalizar y formar la opinión pública. El estudio examina la composición de los *tweets* de organizaciones políticas y sus líderes, así como el efecto de esto *tweets* en todos los involucrados en las elecciones al Parlamento Indio de 2019. Para el análisis se utiliza técnicas de Procesamiento del Lenguaje Natural en el nivel inicial para procesar el texto, dando sentido a las palabras y frases combinadas en una oración, de tal manera que pueda ser entendidas e interpretadas. Los enfoques utilizados son

generación de nubes de palabras, análisis de hashtags, y comprensión de menciones, así obtener la semántica asociada con el texto. Asimismo, se analizan las estrategias políticas adoptadas por esos partidos en Twitter, incluida la acusación de corrupción contra sus oponentes. Los resultados del análisis indicaron que las elecciones están asociadas con casi 15,000 *tweets* relacionados con la corrupción, causando gran influencia en los votantes. Asimismo, los resultados de análisis de polaridad mostraron que *Bharatiya Janata Party* (BJP) tiene un porcentaje de positividad del 41%, seguido por el Congreso con un 39% y luego AAP con un 34%.

Mee *et al.* (2021), indican que el incremento de datos de tipo texto generados debido al uso cada vez mayor de las redes sociales, ha hecho que Twitter se convierta en una herramienta de comunicación indispensable para los políticos. Mediante el análisis de sentimientos, el estudio examina qué patrones de uso revelan las opiniones de los usuarios en relación al Brexit, los cuales comprenden la frecuencia y duración de los *tweets*, así como los términos utilizados y su extensión. Se analizaron 185,970 *tweets* de 576 cuentas de Twitter, donde cada cuenta está asociada a un miembro del Parlamento Británico (MP). Se utilizan métodos como el análisis de regresión y el análisis de sentimiento (TF-IDF, así como Regresión de Texto (del inglés *Text Regression*) y BOW), para investigar si existe una relación entre las características de los datos de texto y las características de los usuarios de Twitter. Los resultados muestran que la medida desarrollada Cociente Brexit (del inglés *Brexit Quotient, BQ*) para cuantificar la postura política de un parlamentario sobre el Brexit, tanto como para los partidos políticos y los diputados partidarios o críticos del Brexit, tienen calificaciones BQ que reflejan su postura. Además, la proporción de BQ negativos a positivos (50,75% a 49,25%) indica que BQ se distribuye de manera muy uniforme entre los que están a favor y en contra de permanecer en la Unión Europea.

Gorodnichenko *et al.* (2021), estudian la difusión de información en las redes sociales y la influencia que tienen los *bots* en las opiniones públicas a través del análisis de sentimientos. Para ello se utilizan los datos de Twitter en la Unión Europea sobre el referéndum de 2016 (Brexit) y las elecciones presidenciales de Estados Unido de 2016, considerando dos tipos de agentes de redes sociales: usuarios reales y *bots* sociales. Los *tweets* se recolectaron utilizando Twitter Streaming APIs para extraer la data en tiempo real y considerando aquellos que contenía la palabra clave "Brexit". Para el caso de las elecciones estadounidenses de 2016 se tomaron en cuenta las palabras clave:

"Election2016", "BlackLivesMatter", "CampaignZero", "ClintonEmails", ImWithHer, "NeverClinton", "FeelTheBern", "CruzCrew", "MakeAmericanGreatAgaion", "Trump", "Clinton". Luego, cada *tweet* es procesado, extrayéndose contenido relevante, para almacenarse en una variable de contenido del *tweet*. Se excluyen caracteres especiales, URLs, y se separan los *tweets* originales de los *retweets*. Asimismo, se clasifican las cuentas de Twitter en humanas y bots, dependiendo de las condiciones que caracterizan a un *tweet* realizado por un *bot* y un humano. Los resultados evidenciaron que los mensajes publicados por lo bots generaban respuestas de humanos similares a las de los *bots*.

Rodriguez-Ibanez *et al.* (2021), aplican un conjunto de métodos básicos para analizar la dinámica estadística y temporal del análisis de sentimientos sobre campañas políticas y evaluar su alcance y limitaciones. Para ello se recopilan miles de mensajes de Twitter que hacen mención de los partidos políticos y sus líderes publicados varias semanas antes y después de las elecciones presidenciales en españolas de 2019. La metodología utilizada consistió en la caracterización estadística de los datos de sentimiento y la extracción de características de patrones intrínsecos no lineales en términos de aprendizaje múltiple utilizando codificadores automáticos e incrustaciones estocásticas. En los resultados, las métricas consideradas son la interpretabilidad y la dinámica temporal, por lo que, de los 13,171 términos considerados en los cuatro diccionarios, el 70% del total se representan en un diccionario combinado llamado newLEX, propuesto para el análisis, donde solo 594 términos fueron considerados en todos ellos, 1,180 en tres diccionarios, 2,223 se incorporaron en dos y 9,174 en un solo diccionario.

Ansari *et al.* (2021), señalan que el análisis de sentimientos de las redes sociales en el ámbito político ayuda a los estrategas políticos a realizar un mejor escrutinio del desempeño de un partido político o candidato e identificar sus debilidades antes de las elecciones, considerando a Twitter como una de las plataformas de redes sociales más populares que permite la utilización y preparación de datos específicos de determinado dominio. El estudio permitió identificar la inclinación de opiniones políticas presentes en los *tweets* modelándolo como un problema de clasificación de texto utilizando *Machine Learning*. Se utilizaron los *tweets* relacionados con las elecciones de Delhi en 2020 para su análisis, así como diversos algoritmos *Machine Learning*, donde el que destacó por su efectividad fue el de Máquinas de Vectores de Soporte alcanzando 0.85 para las métricas de precisión, exhaustividad y valor-F1.

Khatua *et al.* (2020), coinciden con otros estudios afirmando que Twitter ha emergido como una plataforma popular para el discurso político. El estudio hace uso del análisis de sentimientos enfocándose en la relación entre los *tweets* mixtos (definido como un *tweet* que menciona a más de un partido político) y la inclinación política de los usuarios. La metodología utilizada se enfoca en modelos de regresión, así como algoritmos basados en *Deep Learning* (aprendizaje profundo) para descifrar la inclinación política del usuario, considerando patrones de tuiteo mixtos a nivel de usuario como variables de entrada, y desarrollando el estudio en India. Los resultados demuestran que el análisis de *tweets* mixtos puede ayudar a los partidos políticos a identificar su competidor más cercano, segmentar el electorado y decidir sus estrategias de campaña por lo que los tres modelos (RNN, LSTM, Bi-LSTM) funcionaron bien en la predicción de la inclinación política de los usuarios, alcanzado una exactitud entre el 82% y el 87%, donde el modelo Bi-LSTM presentó un mejor rendimiento.

Das *et al.* (2020), revelan la aplicabilidad del análisis de sentimientos en las Elecciones Generales de Australia 2019. La propuesta del trabajo consiste en incorporar un *dataset* de 1.8 millones de *tweets* obtenidos durante el período de elecciones, junto a una función de control de sesgos que los disminuye sin importar los factores demográficos. Se aplican el método ABRTC (del inglés *Average Bias Reduced Tweet Counts*) y ATC (del inglés *Average Tweet Count*), junto a las métricas de Error Absoluto Medio (del inglés *Mean Absolute Error*, MAE) y Error Cuadrático Medio (del inglés *Root-Mean Squared Error*, RMSE) para controlar el factor de sesgo. Los resultados mostraron que el método propuesto ABRTC, se desempeña mejor, alcanzando un MAE de 5.83% y un RMSE de 8.57%.

Ansari *et al.* (2020), argumentan que las plataformas de redes sociales como Twitter generan enormes cantidades de texto que contienen ideas políticas, que se pueden extraer y analizar para predecir tendencias futuras en procesos electorales. Se realiza el análisis de sentimientos, a través de la captura de los sentimientos políticos mediante *tweets*, los cuales son tratados con un enfoque de aprendizaje supervisado. Se prepara el modelo de clasificación basado en sentimientos, utilizando una red neuronal recurrente (LSTM) para predecir la inclinación de los *tweets* e inferir los resultados de las elecciones, y luego compararlo con los modelos clásicos de *Machine Learning*. Los resultados mostraron que, de los 3,896 *tweets*, el 55.46% se inclinan por un partido político, los otros dos solo alcanzaron un 13.21% y 2.07%. Entonces, se concluye que el algoritmo *Random Forest*

se desempeña mejor junto con un tiempo de entrenamiento óptimo, sin embargo, LSTM se desempeña un poco mejor en términos de exactitud en comparación a *Random Forest*, aunque con un alto costo de cálculo.

Ullah *et al.* (2020), también emplean el análisis de sentimientos y data de Twitter en las elecciones presidenciales de Estados Unidos 2016 para entender qué características se aplican mejor en la predicción de los resultados electorales. Se comparan cuatro características como unigramas, bigramas trigramas y palabras de opinión mediante el uso de técnicas de minería de datos como *Random Forest*, *Naïve Bayes* y *Artificial Neural Network*. La data obtenida de Kaggle está formada por 6445 registros, esta es preprocesada para que los datos estén bien formados y luego se extraen características mediante análisis factorial. Los resultados indican que el método usado con la característica unigrama muestra la mejor exactitud con un 81% para *Random Forest*, 97% para *Naïve Bayes* y 93% para *Artificial Neural Network.*

Hswen *et al.* (2020), realizan un estudio con el propósito de usar Twitter para realizar vigilancia en línea del sentimiento negativo hacia los mexicanos e hispanos durante las elecciones presidenciales de Estados Unidos de 2016, por un período de 20 semanas, y examinar su relación con el bienestar mental en este grupo objetivo a nivel de la población. Se utilizó el análisis de sentimientos para capturar el porcentaje de *tweets* negativos, así como un modelo de regresión de series de tiempo para examinar la asociación entre el recuento porcentual de *tweets* negativos que mencionan a mexicanos e hispanos y el recuento porcentual de preocupación entre los encuestados hispanos de Gallup. Los resultados mostraron que, de 2,809,641 tweets que contienen términos mexicanos e hispanos, 687,291 tweets fueron negativos. Entre los 8,314 encuestados hispanos de Gallup, una media del 33,5% respondió estar preocupado a diario, cifras que se han relacionado con el resultado final de las elecciones presidenciales.

Georgiadou *et al.* (2020), utilizan *Big Data Analytics* y análisis de sentimientos basado en *lexicons* para el estudio de las negociaciones internacionales del Brexit entre Reino Unido y la Unión Europea y la utilización de sentimientos de usuarios de Twitter. A través de un corpus de 13’018,367 *tweets* sobre *hashtags* en relación al Brexit, se demuestra que el análisis de sentimiento es una herramienta clave para captar preferencias colectivas y mejorar la toma de decisiones en las negociaciones internacionales. Los resultados respaldan estudios previos que indican que el 43% de la población deseaba que Reino

Unido permaneciera en el mercado único y la unión aduanera, mientras que el 37% quería un Brexit más estricto en el que Gran Bretaña dejara ambos.

Mohamed *et al.* (2020), proponen un ensamble de clasificadores para realizar análisis de sentimientos basado en *lexicons* y algoritmos de *Machine Learning*. Se implementa el modelo con cuatro diferentes vectores de características (*Word2Vec, Glove Vector, BOW Vectors, Sum-Votes Vector*) y once clasificadores diferentes (*Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, Logistic Regresion, Discriminant Analysisi, K-Nearest Neighbors, Stochastic Gradient Descent, AdaBoost, Gradient Boosting, Neural Networks*). Los resultados alcanzados en términos de exactitud fueron del 92.8% para el modelo propuesto usando *Support Vector Machine*, 91.37% para el modelo híbrido basado en *lexicons* y *Machine Learning*, y 81.62% para el modelo previo basado en *Sum-Votes*.

Ahuja *et al.* (2019), destacan el impacto que tiene la extracción de características en el análisis de sentimientos, utilizando Procesamiento de Lenguaje Natural a través de expresiones regulares para la limpieza de datos, Asimismo, aplicando 6 diferentes algoritmos de clasificación en el *dataset* SS-Tweet junto a dos características de extracción (TF-IDF y *N-Grams*), así como. Luego de realizado el análisis, se encontró que, las características de TF-IDF dan mejores resultados con un 3 a 4%, comparado con las características de *N-Grams*. Asimismo, se determinó que, de los algoritmos de *Machine Learning*, la regresión logística ofrecía las mejores predicciones de los sentimientos al proporcionar un resultado máximo para los cuatro parámetros de comparación (exactitud, exhaustividad, precisión, valor-F1, y los dos métodos de extracción de características).

Haryanto *et al.* (2019), utilizan el análisis de sentimientos hacia los candidatos presidenciales y vicepresidenciales de Indonesia en 2019 a través de redes sociales, especialmente Facebook, para conocer la popularidad y sentimientos públicos hacia los candidatos. Los datos se obtuvieron a través de comentarios hechos en Facebook a través de tres medios de comunicación: Detik (@detikcom), TribunNews (@tribunnews) y Liputan6 (@ liputan6online). La clasificación se realiza a través de minería de datos y utilizando el algoritmo clasificador *Naïve Bayes*. Los resultados mostraron que los candidatos Parabowo-Sandiago predominaban en los comentarios con un 59.48% y Jokowi-Maruf con un 40.52%. Por el contrario, los resultados de polaridad demuestran

que Jokowi-Maruf tiene un 56.76% de comentarios positivos y Prabowo-Sandi alcanzó solo 24.21%.

Joseph (2019), propone una metodología utilizando análisis de sentimientos para predecir los resultados de las Elecciones Generales de India 2019, mapeando los estados de ánimo de los votantes en Twitter (se consideraron solo *tweets* en inglés), a través de un clasificador de árboles de decisión para entrenar y probar la data recolectada, y luego predecir los resultados finales. Los cuales fueron cercanos a los reales y a la mayoría de los análisis realizados antes de las elecciones, alcanzando una exactitud del 97%.

Bansal y Srivastava (2018), ratifican la utilidad que tiene el análisis de sentimientos en Twitter, porque es una forma rápida y económica de monitorear las elecciones en tiempo real y así realizar predicciones electorales. Sin embargo, la mayoría de estudios se basa en la extracción explícita del sentimiento público utilizando características léxicas y sintácticas en los *tweets*, por lo que se pasan por alto las relaciones de palabras implícitas y las co-ocurrencias. Por esa razón, el estudio plantea un nuevo método denominado Análisis Híbrido de Sentimientos Basado en Temas (HTBSA) con el objetivo de capturar relaciones de palabras y co-ocurrencias en *tweets* de corta duración para la predicción de elecciones a través de ellos. La metodología utilizada consiste en extraer temas desde un corpus rico en textos cortos usando el modelo *Biterm Topic* (BTM), luego se aprenden los sentimientos para cada tema a partir de recursos léxicos preexistentes. Asimismo, se utilizaron más de 300,000 *tweets* recolectados entre el 1 y 20 de febrero de 2017, para predecir las elecciones legislativas de Uttar Pradesh. Al final, los resultados obtenidos indican que el modelo HTBSA supera las técnicas existentes de predicción de elecciones basadas en Twitter.

Kusen y Strembeck (2018), enfocan su estudio utilizando análisis de sentimientos de discusión en Twitter sobre las elecciones presidenciales en Austria en 2016. En particular, se extraen y analizan 343,645 *tweets* relacionados con el proceso. La metodología utilizada combina diferentes métodos de ciencia de redes y análisis de sentimientos basado en *lexicons*, encontrando que el ganador del proceso (Alexander Van der Bellen) envió *tweets* con predominación neutral (81.99%), mientras que su oponente (Norbert Hofer) se inclinó por *tweets* emocionales que resultaron en puntajes de sentimiento positivos y negativos (42.38%). También se encontró que, la información negativa sobre ambos candidatos se difundió más tiempo que la neutral y la positiva, determinándose

que hubo una clara polarización en términos de los sentimientos difundidos por seguidores de ambos candidatos en Twitter. Naturalmente, el ganador recibió mayor cantidad de me gusta y *retweets*.

Liu y Lei (2018), aplican análisis de sentimientos en los discursos de Hillary Clinton y Donald Trump durante las elecciones presidenciales de 2016 para identificar sus sentimientos, temas y estrategias de discurso mediante la utilización de métodos basados en máquinas, incluido el análisis computarizado de sentimientos a nivel de oraciones, modelados de temas estructurales y exploración de word2vec para asociaciones temáticas, además se utilizaron estudios cualitativos. Los resultados revelaron que los discursos fueron significativamente más negativos que los de Clinton, alcanzando casi un 50% de polaridad negativa.

Saleiro *et al.* (2018), describen una investigación en relación a un conjunto de funciones agregadas de sentimiento, junto a un análisis de regresión en base a encuestas de opinión política y un total de 233,000 *tweets* clasificados según su polaridad (positiva, negativa o neutral), sobre los cinco principales líderes políticos en Portugal, entre junio de 2011 y diciembre 2013. Asimismo, aplicaron dos algoritmos de regresión para el análisis de sentimientos, *Ordinary Least Squares* (OLS) y *Random Forest* (RF), los cuales fueron evaluados a través de MAE para conocer el error de predicción del modelo propuesto, que alcanzó un 0.61%.

Arcila-Calderón *et al.* (2017), describen y evalúan la aplicación de la técnica análisis supervisado de sentimientos en comunicación política a través de un clasificador en tiempo real de opiniones políticas en tweets en español utilizando técnicas de *Machine Learning*. El trabajo es desarrollado en un ordenador local y también utilizando computación distribuida comercial para problemas de datos masivos (*Big Data*). Demostrando que, si los datos etiquetados fueran divididos en dos corpus, uno de entrenamiento con el 70% de los mensajes y el otro de testeo con el 30% restante, la capacidad predictiva del modelo alcanzaría alrededor del 70% de fiabilidad con algoritmos como el *Naïve Bayes*.

Qi *et al.* (2017), utilizan el análisis de sentimientos para la clasificación de *tweets* de las legislaturas estatales (responsables de la formulación de políticas a nivel estatal) en los principales temas de la agenda de políticas definidas por *Policy Agendas Project* (PAP), iniciado para agrupar las políticas nacionales en los Estados Unidos. Se evalúa la

efectividad de tres algoritmos de *Machine Learning*, a saber, Máquinas de Vectores de Soporte, Redes Neuronales Convolucionales (CNN), y Red de Memoria a corto plazo (LSTM). Los resultados muestran que CNN proporciona una mejora significativa de 10% sobre SVM y 7% sobre LSTM aproximadamente. Con el mejor método de CNN se logra estimar los tres temas más tuiteados en un mes en determinado estado y de forma muy confiable, lo que es útil para probar diferentes teorías sobre la difusión de agendas políticas en todos los estados.

Bohórquez *et al.* (2019), emplean el análisis de sentimientos en la identificación de sentimientos en comentarios escritos en español y plasmados en redes sociales utilizando el contexto político de una provincia de Argentina, tarea complicada debido a las variaciones idiomáticas que existen en diferentes países de Latinoamérica. Para ello, se utiliza una combinación de un algoritmo de aprendizaje no supervisado, para la tarea de pseudo clasificación y como alternativa de clasificación manual de comentarios para construir un *dataset* de entrenamiento; y un algoritmo de aprendizaje supervisado, para el modelo de clasificación, incluyendo una capa de preprocesamiento, para corregir faltas ortográficas y reducir la vectorización al generar un clasificador con mayor precisión. Los resultados muestran que el nivel de exactitud alcanzado es de 93%, un resultado mayor al de estudios previos.

A nivel regional, el presente estudio se considera pionero en el área de análisis de sentimiento con enfoque político, a fin de cubrir un vacío de conocimiento existente por ahora. Sin embargo, se encuentra un trabajo elaborado por Fernández (2019), quien desarrolla un analizador de opinión del microblogging Twitter por la clasificación al mundial de fútbol Rusia 2018 de la selección peruana de fútbol, utilizando el *framework Spark*. Los resultados del desempeño del analizador alcanzan una exactitud del 83% usando un modelo de aprendizaje de clasificación binaria basado en Regresión Logística. En cuanto a precisión se alcanza el 82.30% y en el caso de exhaustividad un 89.28%. Asimismo, se construye un *dataset* "PeruARusia2018.csv" para realizar el análisis, el cual es adecuado para entrenar modelos de aprendizaje de clasificación binaria.

CAPÍTULO II
PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema

Según el último reporte de DataReportal, hay 4,480 millones de usuarios de redes sociales en todo el mundo a julio de 2021, lo equivale a casi el 57% de la población mundial total, reportándose 512 millones de nuevos usuarios en los 12 últimos meses. El mismo reporte señala que el usuario tradicional usa o visita activamente un promedio de 6.6 plataformas diferentes de redes sociales cada mes, y pasa un promedio de casi 2 horas y media usando las redes sociales cada día, ubicando a Facebook como la plataforma más utilizada (Kemp, 2021b). En Perú, se incrementó en 3 millones de usuarios nuevos (+13%) entre el 2020 y 2021, alcanzando un total de 27.00 millones de usuarios para enero de 2021, lo que indica que la cantidad de usuarios de redes sociales en Perú equivalía al 81,4% de la población total para el mismo mes. Para el caso de Twitter, este año aumentó en un 6,9% registrando 100 mil peruanos como nuevos usuarios (Kemp, 2021a). Debido a ese crecimiento han surgido diversas investigaciones que buscan diferentes resultados en muchas áreas como la política, tal es el caso del trabajo de Choy *et al.* (2011), donde se utiliza el análisis de sentimiento basado en diccionarios para predecir el porcentaje de votos de los candidatos en las elecciones de Singapur en el año 2011. El estudio recopiló 16,616 *tweets* de la API de Twitter durante la campaña de agosto de 2011 y crearon un corpus personalizado corrigiendo el sesgo que provoca el uso de corpus estandarizados en estos análisis. Sus resultados predijeron quiénes serían los dos candidatos más votados, pero erraron al predecir el vencedor final de las elecciones, por un pequeño margen de votos. Otro país sujeto al mismo análisis ha sido Irlanda en 2011, donde Bermingham y Smeaton (2011) intentaron hacer uso de Twitter para predecir los resultados del proceso electoral llevado a cabo mediante un clasificador de sentimientos supervisado, alcanzando una exactitud efectiva del 65%, resultado que podría ser mejorado. Por otro lado, la investigación desarrollada por Paul *et al.* (2021), considera solamente las estrategias de comunicación de los partidos políticos en India seis meses antes de las

elecciones, excluyéndose las acciones anteriores a este período. El estudio también se limita a los tres partidos políticos más prominentes y al contexto indio, por lo que la generalización a otro contexto es limitada. El trabajo de Mee *et al.* (2021), también está limitado al contexto de Reino Unido, el tema del Brexit y los usuarios de Twitter miembros del parlamento, por lo que la metodología utilizada podría ampliarse para identificar posturas políticas de otros usuarios, refinando los resultados obtenidos en el estudio.

Otro enfoque utilizado para el análisis de sentimiento es el *Deep Learning* o Aprendizaje Profundo, presentado en el trabajo de Khatua *et al.* (2020), donde los resultados obtenido fueron notables, pero también se limitaron al contexto indio y a usuarios con participación activa en Twitter, por lo que, la metodología no es apropiada para usuarios inactivos.

Entonces, existe una gran variedad de técnicas y herramientas que pueden ser utilizadas en el análisis de sentimientos en Twitter en el contexto político, ya sean las basadas en modelos tradicionales de *Machine Learning* con predicciones de gran asertividad en comparación a otros (Ahuja *et al.*, 2019; Ansari *et al.*, 2020; Chaudhry *et al.*, 2021; Haryanto *et al.*, 2019; Joseph, 2019; Mohamed *et al.*, 2020; Ullah *et al.*, 2020), con modelos basados en *lexicons* (Aramburo *et al.*, 2022; Georgiadou *et al.*, 2020; Hswen *et al.*, 2020; Kušen & Strembeck, 2018; Liu & Lei, 2018; Rodriguez-Ibanez *et al.*, 2020), e incluso aquellas que combinen modelos formando híbridos (Bansal & Srivastava, 2018; Sudhir & Suresh, 2021)

En este sentido, es importante realizar un modelado adecuado de las opiniones políticas publicadas en Twitter, como lo menciona Arcila-Calderón *et al.* (2017), mediante la creación de un nuevo *dataset*. Luego, es preciso analizar el conjunto de datos creado, a través de un modelo de análisis de sentimientos basado en las técnicas de *Machine Learning* más asertivas, que permita predecir la intención de voto en las Elecciones Presidenciales Perú 2021.

2.2. Enunciado del problema

¿Qué técnica de *Machine Learning* utilizada en un modelo de análisis de sentimientos en Twitter obtiene los mejores resultados en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021?

2.3. Justificación

El análisis de sentimientos ha sido ampliamente utilizado para la predicción política alrededor del mundo, demostrando su alto nivel de utilidad, como Estados Unidos (Chaudhry *et al.*, 2021; Hswen *et al.*, 2020; Liu & Lei, 2018; Qi *et al.*, 2017), Singapur (Choy *et al.*, 2011), Reino Unido (Georgiadou *et al.*, 2020; Gorodnichenko *et al.*, 2021; Mee *et al.*, 2021), India (Joseph, 2019; Khatua *et al.*, 2020; Paul *et al.*, 2021), Indonesia (Haryanto *et al.*, 2019), Australia (Das *et al.*, 2020) y España (Arcila-Calderón *et al.*, 2017), por citar algunos. Asimismo, las diversas plataformas de redes sociales, como Twitter, brindan una excelente oportunidad para examinar la comunicación pública de un porcentaje de la población (Mislove *et al.*, 2011). Por esa razón, las organizaciones políticas vienen utilizando con mayor frecuencia esta plataforma en sus campañas de comunicación para canalizar y formar la opinión de sus electores (Paul *et al.*, 2021).

El estudio se enfocó en el análisis de sentimientos en Twitter para conocer la inclinación política en las Elecciones Presidenciales Perú 2021 mediante el uso de técnicas de *Machine Learning*. Trabajos similares demuestran su utilidad para el contraste predictivo de resultados electorales futuros, permitiendo el análisis de textos políticos de forma rápida, corregir el sesgo de las encuestas o anticipar las tendencias más significativas con indicadores adelantados (Arcila-Calderón *et al.*, 2017),

De esta forma, con el presente estudio se pretende dar una posible solución al problema de manera eficiente y automática. Además, se constituye en un aporte metodológico, ya que permite conocer el proceso y técnicas utilizadas en el análisis de sentimientos en Twitter, como la recuperación y extracción de datos, preprocesamiento, extracción de características, clasificación de sentimientos y evaluación en las técnicas de *Machine Learning* utilizadas en el modelo. Finalmente, como aporte a la comunidad científica, se tiene el *dataset* de *tweets* que corresponde a las opiniones publicadas en Twitter en las Elecciones Presidenciales Perú 2021.

2.4. Objetivos

2.4.1. Objetivo general

Determinar la técnica de *Machine Learning* para un modelo de análisis de sentimientos en Twitter que obtenga los mejores resultados en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

2.4.2. Objetivos específicos

- Determinar los algoritmos adecuados para el análisis de sentimientos en Twitter que permitan predecir la intención de voto.
- Construir el *dataset* de *tweets* para el análisis de sentimientos en Twitter en relación a las Elecciones Presidenciales Perú 2021.
- Diseñar el modelo de análisis de sentimientos en Twitter utilizando técnicas de *Machine Learning*.
- Definir las métricas de evaluación del modelo de análisis de sentimientos en Twitter para predecir la intención de voto en las Elecciones Presidenciales Perú 2021.

2.5. Hipótesis

2.5.1. Hipótesis general

El modelo de análisis de sentimientos en Twitter basado en Regresión Logística obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

CAPÍTULO III
MATERIALES Y MÉTODOS

3.1. Lugar de estudio

La investigación se realizó en la ciudad de Puno, ubicada a 3,820 metros sobre el nivel del mar, donde se generó el *dataset* a utilizar para el estudio en relación a las elecciones presidenciales Perú 2021, por lo que se consideraron solo los *tweets* que tenían como ubicación geográfica Perú y provincias.

3.2. Población

El presente trabajo de investigación se enfocó en los *tweets* publicados en Twitter en relación a las Elecciones Presidenciales Perú 2021 (Primera vuelta), por lo que la población estuvo formada por los 49,916 *tweets* publicados entre el 1 enero y el 11 de abril de 2021, de los cuales el 80% conformó el grupo de entrenamiento y el 20% de prueba o predicción.

3.3. Muestra

El tipo de selección muestral es no probabilístico, por conveniencia para el estudio y por depender del proceso de toma de decisiones del investigador (Hernández & Mendoza, 2018), ya que se realizó una cuidadosa y controlada selección de *tweets* que cumplían con ciertas características relacionadas a las Elecciones Presidenciales Perú 2021, por lo que se consideró al total de la población.

3.4. Método de investigación

El presente estudio tiene un enfoque cuantitativo, con un diseño no experimental transversal, ya que se observa una situación existente en relación a las Elecciones Presidenciales Perú 2021, las variables no pueden ser modificadas porque ya sucedieron de esa forma, y los datos fueron recolectados en un tiempo único (Hernández & Mendoza, 2018).

3.5. Descripción detallada de métodos por objetivos específicos

El proceso para determinar la técnica de *Machine Learning* para un modelo de análisis de sentimientos en Twitter que obtenga los mejores resultados en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 es presentado en la Figura 9. Este proceso es una adaptación del modelo de análisis de sentimientos desarrollado por (Liu, 2011; Liu & Zhang, 2012) y utilizando ampliamente por diversos autores como (Ahuja *et al.*, 2019; Ansari *et al.*, 2020, 2021; Ullah *et al.*, 2020). Esta metodología incluye 5 principales pasos, a los cuales se agregó uno previo, la selección de algoritmos adecuados para el análisis de sentimientos en Twitter en un contexto político; el siguiente paso permite la recuperación y extracción de *tweets* históricos mediante un *scraper*; el preprocesamiento de datos permite la creación de *Data Frames* a partir de la Raw Data obtenida en el paso anterior, asimismo se termina de construir el *dataset* a utilizar en el estudio; los pasos siguientes permiten hacer un análisis más profundo del *dataset* mediante la extracción de características y establecer los ajustes necesarios para la clasificación de sentimientos a través de los algoritmos seleccionados. Finalmente, se realiza la evaluación de los algoritmos a través de las métricas establecidas para seleccionar el algoritmo o técnica que obtenga mejores resultados para la predicción de intención de voto.

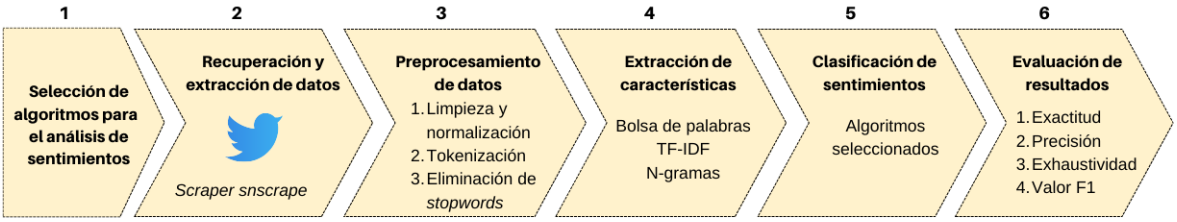

Figura 9. Metodología utilizada para el análisis de sentimientos en Twitter

Fuente: Adaptado de (Liu, 2011; Liu & Zhang, 2012)

3.5.1. Algoritmos adecuados para el análisis de sentimientos en el contexto político

Para la selección de los algoritmos adecuados para el análisis de sentimientos en el contexto político, se realizó la búsqueda y recopilación bibliográfica de diferentes estudios desarrollados entre los años 2017 y 2022 en bases de datos como IEEEXplore, Science Direct, Web of Science, Scopus y MDPI, de los cuales se eligieron aquellos estudios publicados en revistas indexadas, con factor de impacto

y los más recientes en su fecha de publicación. La búsqueda se realizó en base a palabras clave y títulos relacionados al análisis de sentimiento (*Sentiment Analysis, Opinión Mining, Text Classification*) en el contexto político (*political sentiment, political orientations, politics*) aplicado a redes sociales (Twitter), y uso de técnicas *Machine Learning.*

3.5.2. *Dataset*

Para la construcción del esquema del *dataset* denominado Elecciones Bicentenario Tweets Data para el análisis del sentimiento en Twitter en las elecciones presidenciales Perú 2021 se utilizó como esquema base el *dataset* "Auspol2019" utilizado en el estudio de (Das *et al.*, 2020), el mismo que es de libre acceso en la plataforma Kaggle. Asimismo, se siguió la metodología propuesta por (McCreadie *et al.*, 2012) para construir el *dataset* de Twitter, la cual consiste en recuperar los *tweets* mediante un *scraper* que descarga los *tweets* desde Twitter y los reconstruye en formato JSON sin usar la API de Twitter directamente, para después identificar las fuentes de atributos y etiquetas.

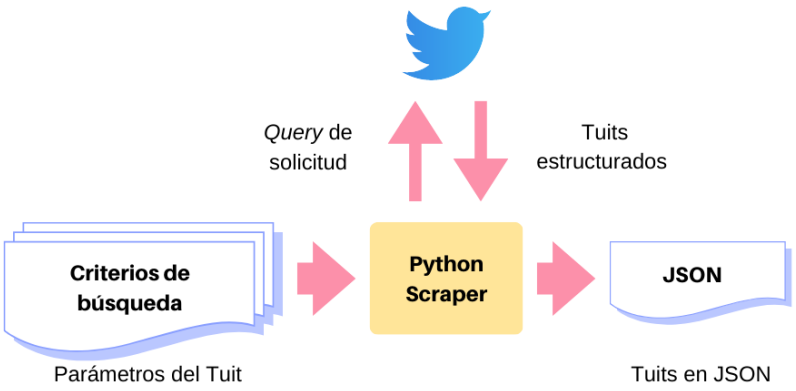

Figura 10. Recuperación de *tweets* a través de un *Scraper* en Python

Fuente: Adaptado de (McCreadie *et al.*, 2012)

Las herramientas utilizadas para alcanzar este objetivo fueron un *scraper* para servicio de redes sociales, de código abierto y denominado **snscrape**, el cual permite realizar el arrastre de *tweets* históricos con fecha anterior, generando un fichero JSON, con el que se formó el *RawData* para luego ser procesado, formar el *DataFrame* y continuar con las tareas de preprocesamiento de texto, así como la clasificación según su polaridad. También, se utilizó el lenguaje de programación

Python 3.9 para el desarrollo de *scripts* y construcción del modelo, el cual se implementó en el entorno web de *Google Colab*.

El *scraper* utilizado requiere de distintos parámetros para hacer la búsqueda de los *tweets*. Las características de cada parámetro se detallan a continuación:

Cantidad máxima de resultados: La instrucción para este parámetro fue de 110,000 *tweets* a rastrear, de los cuales se redujeron a 104,916 *tweets* como *Raw Data* o datos sin procesar.

Hashtag: Se consideraron los *hashtags* con mayor cantidad de menciones o en tendencia durante el período de tiempo requerido: #EleccionesBicentenario y #Elecciones2021.

Idioma: Se hizo el arrastre de *tweets* solo en español.

Intervalo de fecha de búsqueda: Los *tweets* descargados fueron publicados entre el 01 de enero y el 11 de abril de 2021 en Perú.

El script utilizado para descargar los *tweets* incluye los parámetros descritos de la siguiente forma:

```
snscrape -jsonl --max-results 110000 twitter-hashtag "eleccionesbicentenario OR elecciones2021 lang:es since:2021-01-01 until :2021-04-11" > EleccionesBicentenario.json
```

El esquema de preprocesamiento utilizado considera los siguientes pasos:

- **Limpieza de datos:** Consiste en el limpieza y normalización de la *RawData*, eliminando los *tweets* que no están en español, así como los duplicados para mantener la unicidad de los *tweets*, y convirtiendo la *RawData* en *DataFrame*. Se continúa con la normalización de todo el texto, eliminando signos de puntuación, URLs, espacios múltiples y caracteres especiales. Para estas dos tareas se utilizó la clase *json_normalize* de *Pandas*, obteniéndose un *RawData* de 28 columnas para la entidad *Tweets* y 22 columnas para la entidad *Users*. Luego se eliminan las columnas innecesarias o irrelevantes, generando dos *DataFrames*, para *Users* con 13 columnas y *Tweets* con 12 columnas, como se muestra en la Tabla 4. Al *DataFrame Tweets* se le agregaron algunos atributos del *DataFrame Users* necesarios para su análisis. Posteriormente se verifica que no haya valores

nulos que pueden afectar la precisión del modelo. También se revisa que los datos sean apropiados, filtrando y eliminado aquellos que no correspondan a las Elecciones Presidenciales Perú 2021, considerando en este punto la ubicación de los usuarios.

- **Tokenización:** Se tokenizaron los *tweets*, reduciéndolos a su unidad mínima, palabras (Chaudhry *et al.*, 2021; Mee *et al.*, 2021).

- **Eliminación de *stop words*:** Se eliminaron las palabras vacías, ya que no contribuyen al análisis de sentimientos (Ahuja *et al.*, 2019).

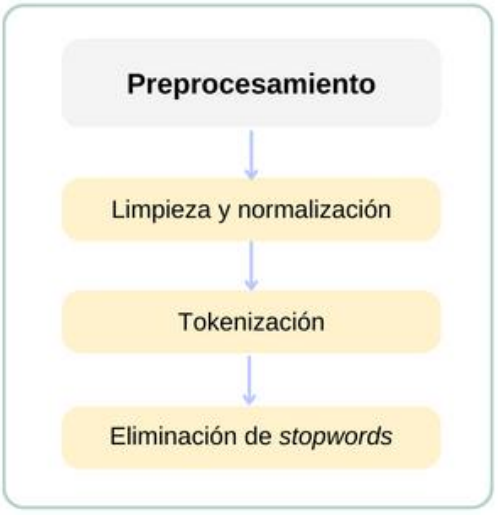

Figura 11. Esquema de preprocesamiento de datos

Finalmente, debido a que los datos para el análisis se recopilaron directamente de Twitter, no están etiquetados, por lo que es necesario emplear otras herramientas como LIWC para esa tarea (Chaudhry *et al.*, 2021). Siendo así, el presente estudio utilizó *pysentimiento* (Pérez *et al.*, 2021), herramienta a través de la cual se realizó el etiquetado para crear el conjunto de datos final.

3.5.3. Modelo de análisis de sentimientos

Se empleó la metodología propuesta por (Liu, 2011; Liu & Zhang, 2012), siguiendo cada una de las etapas descritas: Recuperación y extracción de datos (utilizando un *scraper* para obtener la data de Twitter), preprocesamiento de texto (incluye todas las tareas de limpieza, normalización, tokenización, eliminación de palabras vacías y lematización de texto), ingeniería o extracción de características (mediante TDF-IDF, BOW y N-gramas), clasificación de sentimientos (utilizando los algoritmos o

técnicas de *Machine Learning* seleccionados) y evaluación de resultados (mediante métricas de desempeño como exactitud, precisión, exhaustividad y valor-F1). Adicionalmente, luego de la tokenización, se realizó el etiquetado de texto, para asignarle una polaridad (positiva, negativa o neutral), necesaria para la clasificación de sentimientos.

Las herramientas utilizadas en el entorno de Google Colab fueron: *Pandas, Seaborn, Matplotlib.pylot, plotly.graph_objs, plotly.express, cufflinks, collections, NLTK, re, string, sklearn.feature_extraction.text (CountVectorizer y TdfidfVectorizer), WordCloud, StopWords (spanish), Word_tokenize, sklearn.linear_model (LogisticRegression), sklearn.svm (SVC), sklearn.Naïve_bayes, sklearn.tree (DecisionTreeClassifier), sklearn.metrics: classification_report, confusión_matrix, accuracy_score, sklearn.model_selection: train_test_split y pysentimiento*.

3.5.4. Métricas de evaluación

Las métricas seleccionadas permiten evaluar el desempeño de los algoritmos *Machine Learning* utilizados en el modelo de análisis de sentimientos propuesto en términos de exactitud, precisión, exhaustividad y valor-F1, las cuales destacan por su utilidad e idoneidad. Así lo demuestran los trabajos desarrollados por Rodriguez-Ibanez *et al.* (2020) y Ansari *et al.* (2020), quienes emplean exhaustividad, precisión y valor-F1; Chaudhry *et al.* (2021) quienes emplean exactitud y precisión; Ullah *et al.* (2020), Mohamed *et al.* (2020), Joseph (2019), y Arcila-Calderón *et al.* (2017) quienes aplican exactitud. La matriz de confusión, permite tener una imagen más clara en relación a los errores, predicciones correctas e incorrectas realizadas.

Asimismo, tanto como la exactitud, precisión y exhaustividad, se utilizaron los términos verdadero positivo (TP), verdadero negativo (TN), falso negativo (FN) y falso positivo (FP). Donde, TP representa el número de *tweets* etiquetados correctamente según su polaridad, TN, FP y FN representan el número de *tweets* etiquetados con otra polaridad que no corresponde a la suya. La métrica valor-F1 se da en términos de la precisión y la exhaustividad. Estas métricas se describen matemáticamente como:

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$

$$Precisión = \frac{TP}{TP + FP} * 100\%$$

$$Exhaustividad = \frac{TP}{TP + FN} * 100\%$$

$$Valor\ F1 = \frac{2 * Precisión * Exhaustividad}{Precisión + Exhaustividad}$$

La exactitud muestra en esencia con qué frecuencia un algoritmo de clasificación de *Machine Learning* es correcto en general, mientras que la exhaustividad muestra si el algoritmo puede encontrar todos los objetos de la clase objetivo (Liu, 2011). En tanto, la precisión muestra con qué frecuencia el algoritmo es correcto al predecir la clase objetivo. Ahora bien, para aplicar estas métricas, es clave que el *dataset* se encuentre equilibrado sobre todo para las métricas de exactitud y exhaustividad, sino fuera el caso es preciso incorporar las métricas de precisión y valor-F1.

CAPÍTULO IV
RESULTADOS Y DISCUSIÓN

En este capítulo se presentan los resultados obtenidos en relación al desarrollo de los objetivos de la investigación y su respectiva discusión. En la sección 4.1. se presenta la determinación de los algoritmos adecuados para el análisis de sentimientos en Twitter que permitan definir la intención de voto. En la sección 4.2. se explica cómo se realizó la construcción del *dataset* de *tweets* para el análisis de sentimientos en Twitter en relación a las Elecciones Presidenciales Perú 2021. En la sección 4.3. se muestra el diseño del modelo de análisis de sentimientos en Twitter utilizando técnicas de *Machine Learning*. En la sección 4.4. se expone el entrenamiento y evaluación del modelo de análisis de sentimientos en Twitter para determinar la intención de voto en las Elecciones Presidenciales Perú 2021.

4.1. Resultado conforme al primer objetivo específico

La revisión de la literatura sobre algoritmos *Machine Learning* utilizados para el análisis de sentimientos en el contexto político, permitió determinar los modelos más adecuados para el presente estudio, considerando para ello los resultados de evaluación obtenidos a nivel de desempeño y frecuencia de uso. La Tabla 1 muestra los autores con el detalle de las técnicas utilizadas y las métricas aplicadas para su evaluación.

Tabla 1

Técnicas de Machine Learning y métricas aplicadas para el análisis de sentimientos en el contexto político

N°	Referencia	Factor de medición	Fuente	Técnica	Métricas			
					Exactitud	Exhaustividad	Precisión	Valor-F1
1	Aramburo *et al.* (2022)	SJR	Science Direct - Procedia Computer Science	***Text Mining, Lexicon-based approach***				
2	Chaudhry *et al.* (2021)	SJR - JCR	MDPI - Electronics	***NB***	94.58%		93.19%	
3	Rodriguez-Ibanez *et al.* (2021)	SJR	IEEEXplore - IEEE Access	***Manifold Learning, autoencoders, t-SNE, lexicons***				
4	Rodriguez-Ibanez *et al.* (2020)	SJR	IEEEXplore - IEEE Access	***Lexicon-based approach***		85.00%	85.00%	85.00%
5	Khatua *et al.* (2020)	SJR	Science Direct - Applied Sof Computing Journal	RNN, LSTM, **Bi-LSTM**	87.00%			
6	Ansari *et al.* (2020)	SJR	Science Direct - Procedia Computer Science	*LR, SVM, DTC,* ***RF***, ***LSTM***		77.00%	77.00%	74.00%
7	Ullah *et al.* (2020)	JCR	IEEEXplore - WIE	*RF*, ***NB***, *ANN*	97.00%			
8	Mohamed *et al.* (2020)	SJR	World Scientific -International Journal of Computational Intelligence and Applications (IJCIA)	*NB,* ***SVM****, DTC, RF, LR, Discriminant Analysis, KNN, Stochastic Gradient Descent, AdaBoost, Gradient Boosting, ANN*	92.80%			
9	Hswen *et al.* (2020)	SJR	Science Direct - Heliyon	***Lexicon-based approach, VADER***				

10	Georgiadou *et al.* (2020)	SJR	Science Direct - Information Journal of Information Management	***Lexicon-based approach***				
11	Ahuja *et al.* (2019)	SJR	Science Direct - Procedia Computer Science	***LR**, SVM, DTC, RF, KNN*	57.00%	50.00%	57.00%	50.00%
12	Haryanto *et al.* (2019)	SJR	Science Direct - Procedia Computer Science	***NB***				
13	Joseph (2019)	JCR	IEEEXplore	***DTC***	97.00%			
14	Bansal y Srivastava (2018)	SJR	Science Direct - Procedia Computer Science	***Hybrid Topic Based Sentiment Analysis, HTBSA***				
15	Kušen y Strembeck (2018)	SJR	Science Direct - Online Social Networks and Media	***Lexicon-based approach, NRC emotion lexicon***				
16	Liu y Lei (2018)	SJR	Science Direct - Discourse, Context & Media	***Lexicon-based approach***				
17	Arcila-Calderón *et al.* (2017)	SJR - JCR	WoS, Scopus - Profesional de la información	***NB***	70.00%			

Por otro lado, se pueden apreciar los valores alcanzados para cada métrica, resaltándose en negrita las técnicas que lograron ese resultado; entre ellos los algoritmos NB, SVM y DTC mostraron un mejor desempeño en términos de exactitud.

Tabla 2

Resumen de técnicas Machine Learning utilizadas en el análisis de sentimientos en el contexto político

N°	Referencia	LR	NB	SVM	DTC	RF	*Lexicons*	Otros
1	Aramburo *et al.* (2022)						✓	✓
2	Chaudhry *et al.* (2021)		✓					
3	Rodriguez-Ibanez *et al.* (2021)						✓	✓
4	Rodriguez-Ibanez *et al.* (2020)						✓	
5	Khatua *et al.* (2020)							✓
6	Ansari *et al.* (2020)	✓		✓	✓	✓		✓
7	Ullah *et al.* (2020)		✓			✓		✓
8	Mohamed *et al.* (2020)	✓	✓	✓	✓	✓		✓
9	Hswen *et al.* (Hswen et al., 2020)						✓	
10	Georgiadou *et al.* (2020)						✓	
11	Ahuja *et al.* (2019)	✓		✓	✓	✓		✓
12	Haryanto *et al.* (2019)		✓					
13	Joseph (2019)				✓			
14	Bansal y Srivastava (2018)						✓	✓
15	Kušen y Strembeck (2018)						✓	
16	Liu y Lei (2018)						✓	
17	Arcila-Calderón *et al.* (2017)		✓					

En este contexto, Sudhir y Surech (2021) muestran las ventajas que poseen las técnicas tradicionales en relación a las basadas en *lexicons*, indicando que tienen un alto nivel de precisión en la clasificación de texto. Donde, son los algoritmos de SVM, KNN y DTC los que muestran un mayor porcentaje de exactitud. Además, las técnicas tradicionales utilizan las características sintácticas y/o lingüísticas para la clasificación de texto (Medhat *et al.*, 2014), más no requieren de diccionarios ni reglas creadas manualmente, los cuales forman parte de un conjunto de activos semánticos que son necesarios para las técnicas basadas en *lexicons*.

De acuerdo al análisis realizado, se observa que el 47% de los estudios usa técnicas más tradicionales (LR, NB, SVM, DTC, RF), el 41% emplean técnicas basadas en *lexicons*, el 6% utiliza técnicas híbridas, y el 6% restante trabaja con técnicas *Deep Learning*. Claramente, se observa que son las técnicas tradicionales las utilizadas con mayor frecuencia y en segundo lugar están las técnicas basadas en *lexicons*. De las técnicas tradicionales, el algoritmo NB es el más popular por su velocidad y exactitud (Arcila-Calderón *et al*., 2017; Chaudhry *et al.*, 2021; Haryanto *et al.*, 2019; Mohamed *et al.*, 2020; Ullah *et al.*, 2020), seguido por RF (Ahuja *et al.*, 2019; Ansari *et al.*, 2020; Mohamed *et al*., 2020; Ullah *et al.*, 2020) y DTC (Ahuja *et al*., 2019; Ansari *et al.*, 2020; Joseph, 2019; Mohamed *et al.*, 2020), luego LR y SVM (Ahuja *et al.*, 2019; Ansari *et al*., 2020; Mohamed *et al.*, 2020). Estos resultados se muestran en la Tabla 2.

4.2. Resultado conforme al segundo objetivo específico

Para alcanzar este objetivo se diseñó una arquitectura de procesamiento en la Nube para trabajar con el conjunto de datos obtenido, este esquema se muestra en la Figura 12. El primer paso consistió en la extracción de datos, utilizando ***snscrape***, que al igual que TWINT, es una herramienta desarrollada en Python, y no requiere el uso del API de Twitter para la extracción de la información. Esta herramienta es sencilla de instalar, pues no se necesita de *Keys*, *Tokens* o configuraciones adicionales (Aramburo *et al.*, 2022).

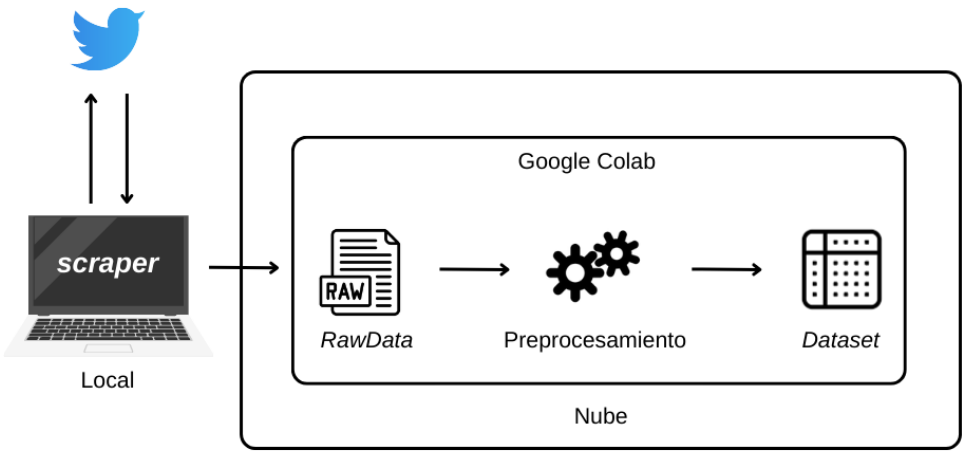

Figura 12. Arquitectura de procesamiento en la Nube utilizada

Antes de instalar ***snscrape***, se realizó su descarga desde un repositorio de GitHub https://github.com/JustAnotherArchivist/snscrape. La instalación se ejecutó de manera local en el disco duro del ordenador, y así como la instalación, se hizo mediante línea de comandos.

```
pip3 install git+https://github.com/JustAnotherArchivist/snscrape.git
```

Se recuperaron un total de 104,916 registros por 31 columnas o atributos iniciales para el análisis de sentimientos en Twitter en las Elecciones Presidenciales Perú 2021. Los datos fueron recuperados considerando el período de tiempo dado entre el 1 de enero y 11 de abril de 2021 (fecha en la que se realizaron las elecciones de primera vuelta). Asimismo, en la Tabla 3 se observa cómo se dividió el conjunto de datos total en dos grupos *RawData Users* y *RawData Tweets*. Para asegurar que la data esté relacionada únicamente con el contexto de las Elecciones Presidenciales Perú 2021, se realizó la búsqueda y extracción de datos en base a los *hashtags* específicos que estuvieron en tendencia durante el período tiempo indicado, considerando líderes y partidos políticos con los *hashtags* #EleccionesBicentenario y #Elecciones2021 (Ansari *et al.*, 2020; Georgiadou *et al.*, 2020; Khatua *et al.*, 2020). Se utilizaron esos hashtags como filtro para eliminar aquellos que no estuvieran relacionados al proceso electoral en cuestión, considerando además la ubicación, ya que en ese mismo período de tiempo se llevaban a cabo otros procesos electorales en países como México, Chile, Ecuador, Honduras, España, Bolivia y Argentina, donde también había otros *hashtags* en tendencia.

Tabla 3

DataFrames resultantes (cantidad de registros y atributos obtenidos)

RawData	**N° Registros**	**N° Atributos inicial**
Users	20,623	17
Tweets	104,916	31

En el proceso de arrastre o recuperación de *tweets* se obtuvieron todos los atributos (variables o columnas) que conforman un *tweet*, de los cuales se removieron los menos útiles y se renombraron aquellos para un mejor manejo y análisis. Posteriormente, se normalizaron los datos y se transformaron a *DataFrame*. Estos resultados se muestran en la Tabla 4.

Tabla 4

Variables o columnas de los RawData y DataFrame Users y Tweets

Entidad	Variables *RawData*	Variables *DataFrame*
Usuarios (*Users*)	_type username id displayname rawDescription renderedDescription descriptionLinks verified created followersCount friendsCount statusesCount, favouritesCount listedCount mediaCount location protected link profileImageUrl profileBannerUrl label url	Username userId displayname rawDescription created followersCount friendsCount statusesCount favouritesCount listedCount mediaCount userLocation profileUrl
Tuits (*Tweets*)	_type url date rawContent renderdContent id user replyCount retweetCount likeCount quoteCount conversationId lang source sourceUrl sourceLabel links media retweetedTweet quotedTweet inReplyToTweetId inReplyToUser metionedUsers coordinates place hashtags cashtags card	CreatedAt tweetId renderedContent retweetCount likeCount replyCount userId userName userScreenName userDescription userLocation (tweetLocation) hasthtags

Además, en el *DataFrame Tweets* se incluyeron algunos atributos del *DataFrame Users* para incluir los datos más relevantes del *DataFrame Users*, tener un conjunto de datos más completo y útil para su análisis.

En el preprocesamiento se realizaron varios ajustes, empezando por la limpieza del *DataFrame*, mediante la eliminación de duplicados y aquellos *tweets* que no pertenecen a Perú, lo que redujo considerablemente la cantidad de datos. Se conservaron solo los *tweets* en español, se eliminaron las URLs, signos de puntuación, caracteres especiales, menciones, espacios múltiples y palabras de tamaño menor a 3 caracteres, para evitar *tweets* sin sentido. En la Tabla 5 se muestran parte de los resultados obtenidos en el proceso de limpieza.

Tabla 5

Resultados obtenidos de la limpieza de datos

Index	RenderedContent	CleanContent
0	Al final posponer la emisión del documental “la revolución y la tierra” le costará más caro a la derecha bruta y achorada 🤷‍♀️ #EleccionesBicentenario https://t.co/Z8RBcunpYm	al final posponer la emisión del documental “la revolución la tierra” le costará más caro la derecha bruta achorada 🤷‍♀️ eleccionesbicentenario
1	Si Dios contigo , quien contra ti ??? #RafaelPresidente2021 #EleccionesBicentenario https://t.co/TFqCEzov6G	si dios contigo quien contra ti rafaelpresidente2021 eleccionesbicentenario
2	Si #HernandoDeSoto no es presidente me mato ptm😔 #EleccionesBicentenario	si hernandodesoto no es presidente me mato ptm😔 eleccionesbicentenario
3	Voy a llorar no quiero que #RafaelLopezAliaga sea presidente😭 #EleccionesBicentenario	voy llorar no quiero que rafaellopezaliaga sea presidente😭 eleccionesbicentenario
4	#Elecciones2021: ODPE Huancayo reparte material electoral para 270 locales de votación bit.ly/3dRfmmz https://t.co/CRaiYQVj7l	elecciones2021 odpe huancayo reparte material electoral para 270 locales de votación
5	#Elecciones2021 PE | JNE: salud, educación y economía, los temas con mayor cobertura informativa Otros temas importantes son seguridad ciudadana, reforma política o constitucional, empleo y corrupción. larepublica.pe/politica/2021/…	elecciones2021 PE jne salud educación economía los temas con mayor cobertura informativa otros temas importantes son seguridad ciudadana reforma política constitucional empleo corrupción

6	[¡Gustavo Gorriti apoya a Keiko Fujimori!] Con el fin de frenar el avance de Candidatos Radicales, Gorriti confirmó su apoyo a Keiko. Afirmó que: "Uno puede no estar de acuerdo con ella, pero su vocación democrática es evidente" #KeikoPresidenta #EleccionesBicentenario https://t.co/hQ7PRbwu8F	gustavo gorriti apoya keiko fujimori con el fin de frenar el avance de candidatos radicales gorriti confirmó su apoyo keiko afirmó que uno puede no estar de acuerdo con ella pero su vocación democrática es evidente keikopresidenta eleccionesbicentenario
7	Por más que muchos estén desilusionados o decepcionados, este 11 de abril tendremos la oportunidad de preservar ese sueño de un Perú mucho mejor. No dejemos ese sueño desaparecer. . . . #Elecciones2021 #SomosLibresSeamosloSiempre #LibertadDeElegir https://t.co/qTGbFSjohm	por más que muchos estén desilusionados decepcionados este 11 de abril tendremos la oportunidad de preservar ese sueño de un perú mucho mejor no dejemos ese sueño desaparecer elecciones2021 somoslibresseamoslosiempre libertaddeelegir
8	Mañana hagamos que valga la pena las doce horas que los miembros de mesa estaremos en los centros de votación. Vota consciente y respetando las medidas 🤲 #EleccionesBicentenario	mañana hagamos que valga la pena las doce horas que los miembros de mesa estaremos en los centros de votación vota consciente respetando las medidas eleccionesbicentenario
9	#Elecciones2021 Más de 900 miembros de la MML apoyarán durante la jornada electoral ►buff.ly/3tenT9G https://t.co/NjeMCGMPDX	elecciones2021 más de 900 miembros de la mml apoyarán durante la jornada electoral

Para el proceso de limpieza se utilizó un script en Python que permitió conservar los *emojis*, ya que estos también representan los sentimientos de los usuarios. Posteriormente, se realizó el proceso de tokenización y eliminación de palabras vacías. La tokenización se hizo mediante la clase *word_tokenize* de *nltk.tokenize*, y la eliminación de palabras vacías se hizo a través de *stopwords* de *nltk.corpus*. Como resultado se obtuvo la columna *ContentStopWords*, la cual se muestra en la Tabla 6.

Tabla 6

Ejemplo de los resultados obtenidos de la tokenización y eliminación de palabras vacías

Index	CleanContent	ContentStopWords
0	al final posponer la emisión del documental “la revolución la tierra” le costará más caro la derecha bruta achorada 🤷‍♀️ eleccionesbicentenario	final posponer emisión documental “ revolución tierra ” costará caro derecha bruta achorada 🤷‍♀️ eleccionesbicentenario
1	si dios contigo quien contra ti rafaelpresidente2021 eleccionesbicentenario	si dios contigo rafaelpresidente2021 eleccionesbicentenario
2	si hernandodesoto no es presidente me mato ptm😔 eleccionesbicentenario	si hernandodesoto presidente mato ptm😔 eleccionesbicentenario
3	voy llorar no quiero que rafaellopezaliaga sea presidente😭 eleccionesbicentenario	voy llorar quiero rafaellopezaliaga presidente😭 eleccionesbicentenario

4	elecciones2021 odpe huancayo reparte material electoral para 270 locales de votación	elecciones2021 odpe huancayo reparte material electoral 270 locales votación
5	elecciones2021 PE jne salud educación economía los temas con mayor cobertura informativa otros temas importantes son seguridad ciudadana reforma política constitucional empleo corrupción	elecciones2021 PE jne salud educación economía temas mayor cobertura informativa temas importantes seguridad ciudadana reforma política constitucional empleo corrupción
6	gustavo gorriti apoya keiko fujimori con el fin de frenar el avance de candidatos radicales gorriti confirmó su apoyo keiko afirmó que uno puede no estar de acuerdo con ella pero su vocación democrática es evidente keikopresidenta eleccionesbicentenario	gustavo gorriti apoya keiko fujimori fin frenar avance candidatos radicales gorriti confirmó apoyo keiko afirmó puede acuerdo vocación democrática evidente keikopresidenta eleccionesbicentenario
7	por más que muchos estén desilusionados decepcionados este 11 de abril tendremos la oportunidad de preservar ese sueño de un perú mucho mejor no dejemos ese sueño desaparecer elecciones2021 somoslibresseamoslosiempre libertaddeelegir	desilusionados decepcionados 11 abril oportunidad preservar sueño perú mejor dejemos sueño desaparecer elecciones2021 somoslibresseamoslosiempre libertaddeelegir
8	mañana hagamos que valga la pena las doce horas que los miembros de mesa estaremos en los centros de votación vota consciente respetando las medidas eleccionesbicentenario	mañana hagamos valga pena doce horas miembros mesa centros votación vota consciente respetando medidas eleccionesbicentenario
9	elecciones2021 más de 900 miembros de la mml apoyarán durante la jornada electoral	elecciones2021 900 miembros mml apoyarán jornada electoral

El siguiente paso corresponde al etiquetado de los *tweets*, con polaridad positiva, negativa o neutral. Esta tarea se realizó mediante *pysentimiento*, una librería desarrollada en Python de código abierto aplicada al análisis de sentimientos y emociones, con soporte para inglés y español (Pérez *et al.*, 2021). Esto permitió obtener el atributo *Polarity* de la Tabla 7. Todo este proceso dejó ver a detalle el paso a paso de los cambios a los que se sometieron los *tweets* para poder recibir una etiqueta de polaridad necesario para el entrenamiento del modelo.

Tabla 7

Ejemplo de los resultados obtenidos del etiquetado de polaridad

Index	CleanContent	ContentStopWords	Polarity	PolarityValue
0	al final posponer la emisión del documental “la revolución la tierra” le costará más caro la derecha bruta achorada 🤷‍♀️ eleccionesbicentenario	final posponer emisión documental “ revolución tierra ” costará caro derecha bruta achorada 🤷‍♀️ eleccionesbicentenario	NEG	-1
1	si dios contigo quien contra ti rafaelpresidente2021 eleccionesbicentenario	si dios contigo rafaelpresidente2021 eleccionesbicentenario	NEU	0

2	si hernandodesoto no es presidente me mato ptm😔 eleccionesbicentenario	si hernandodesoto presidente mato ptm😔 eleccionesbicentenario	NEG	-1
3	voy llorar no quiero que rafaellopezaliaga sea presidente😭 eleccionesbicentenario	voy llorar quiero rafaellopezaliaga presidente😭 eleccionesbicentenario	NEG	-1
4	elecciones2021 odpe huancayo reparte material electoral para 270 locales de votación	elecciones2021 odpe huancayo reparte material electoral 270 locales votación	NEU	0
5	elecciones2021 PE jne salud educación economía los temas con mayor cobertura informativa otros temas importantes son seguridad ciudadana reforma política constitucional empleo corrupción	elecciones2021 PE jne salud educación economía temas mayor cobertura informativa temas importantes seguridad ciudadana reforma política constitucional empleo corrupción	NEU	0
6	gustavo gorriti apoya keiko fujimori con el fin de frenar el avance de candidatos radicales gorriti confirmó su apoyo keiko afirmó que uno puede no estar de acuerdo con ella pero su vocación democrática es evidente keikopresidenta eleccionesbicentenario	gustavo gorriti apoya keiko fujimori fin frenar avance candidatos radicales gorriti confirmó apoyo keiko afirmó puede acuerdo vocación democrática evidente keikopresidenta eleccionesbicentenario	NEU	0
7	por más que muchos estén desilusionados decepcionados este 11 de abril tendremos la oportunidad de preservar ese sueño de un perú mucho mejor no dejemos ese sueño desaparecer elecciones2021 somoslibresseamoslosiempre libertaddeelegir	desilusionados decepcionados 11 abril oportunidad preservar sueño perú mejor dejemos sueño desaparecer elecciones2021 somoslibresseamoslosiempre libertaddeelegir	NEU	0
8	mañana hagamos que valga la pena las doce horas que los miembros de mesa estaremos en los centros de votación vota consciente respetando las medidas eleccionesbicentenario	mañana hagamos valga pena doce horas miembros mesa centros votación vota consciente respetando medidas eleccionesbicentenario	POS	1
9	elecciones2021 más de 900 miembros de la mml apoyarán durante la jornada electoral	elecciones2021 900 miembros mml apoyarán jornada electoral	NEU	0

Del análisis realizado, previo al preprocesamiento de datos, se obtuvo el *dataset* Elecciones Bicentenario 2021 Tweets.csv (almacenado en Google Drive), conformado por 49,916 registros y 12 columnas o atributos, los cuales fueron obtenidos con la combinación del período de tiempo y los hashtags seleccionados. El esquema del *dataset* resultante es similar al utilizado por Das *et al.* (2020), ya que los atributos utilizados en

el estudio son considerados adecuados e relevantes para conformar el *dataset* y luego ser analizados. La Tabla 8 muestra el detalle de los atributos que conforman el *dataset*.

Tabla 8

Atributos del dataset Elecciones Bicentenario 2021 Tweets

N°	Atributo	Descripción
1	CreatedAt	Fecha de creación del *tweet*
2	tweetId	Identificador de *tweet*
3	renderedContent	Contenido del *tweet*
4	retweetCount	Número de *retweet*
5	likeCount	Número de me gusta
6	replyCount	Número de respuestas
7	userId	Identificador del usuario
8	userName	Nombre del usuario
9	userScreenName	Nombre de usuario mostrado
10	userDescription	Descripción del usuario
11	tweetLocation	Localización
12	hasthtags	*Hashtags* que componen el *tweet*

4.3. Resultado conforme al tercer objetivo específico

Se evaluaron y entrenaron las técnicas de *Machine Learning* seleccionadas (LR, NB, SVM y DTC) sobre el conjunto de datos formado con los *tweets* publicados en relación a las Elecciones Presidenciales Perú 2021. En ese mismo año, la cantidad de usuarios en Twitter en Perú se incrementó, alcanzando para finales de año un total de 8 millones de usuarios, de los cuales el 60% estuvo representado hombres y el 48% por mujeres. Además, 81% de *tweets* provienen del departamento de Lima, seguido por La Libertad, Arequipa y Lambayeque (Zorrilla, 2022).

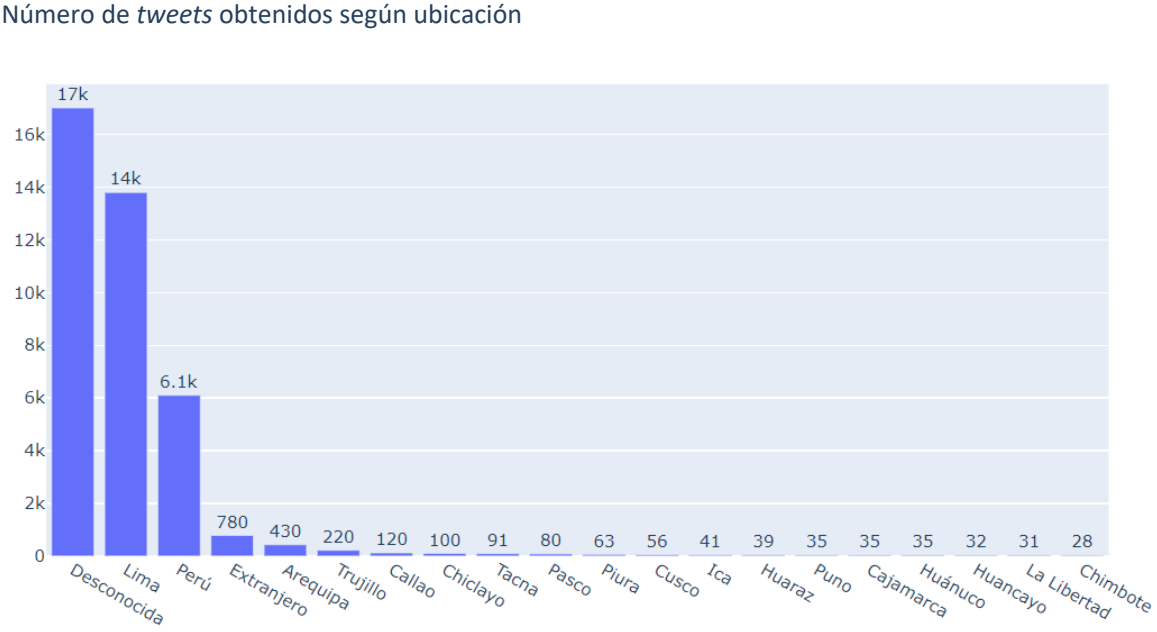

Figura 13. Número de *tweets* obtenidos según ubicación

En la Figura 13 se muestra la cantidad de publicaciones realizadas en Twitter en relación a las Elecciones Presidenciales Perú 2021 de acuerdo a las siguientes ubicaciones: Perú, Lima, provincias y ubicación desconocida. Esta clasificación se hizo agrupando los *tweets* en diferentes grupos: (1) Perú, considerando los *tweets* que provienen de "Perú", "Peru", "PERU", "Perú " y similares. (2) Lima, considerando los *tweets* que provienen de Lima, Lima-Perú, Lima – Perú y Lima – Peru. (3) Provincias, se muestran las que tienen mayor cantidad de publicaciones. (4) Extranjero, se encuentran aquellos *tweets* que fueron publicados en otro país. (5) Ubicación desconocida, en este grupo se encuentran aquellos *tweets* que se presumen que son de alguna parte del Perú o del extranjero. De este análisis, se tienen que, el 34% tienen una ubicación no conocida, el 28% provienen de Lima, el 12% provienen de Perú, el 1% tienen como origen el extranjero, y el 25% restante se distribuye entre las diferentes provincias del Perú. Por lo tanto, las opiniones publicadas en Twitter en relación a las Elecciones Presidenciales Perú 2021, muestran mayor inclinación por candidatos conocidos en Lima.

Del total de *tweets* procesados, 26,644 se etiquetaron como neutrales, 20,898 como negativos y 2,374 como positivos, lo que se refleja en las Figuras 14, 15 y 16, donde los *hashtags* y términos con más cantidad de apariciones en un sentido positivo, negativo y neutral fueron: #EleccionesBicentenario, #Elecciones2021, #LopezAliaga, #RafaelEnPrimeraVuelta, #AliagaPresidente, #DebatePresidencialJNE, #KeikoFujimori, #Lescano, "Pedro Castillo", " Rafael López", "Hernando Soto", entre otros.

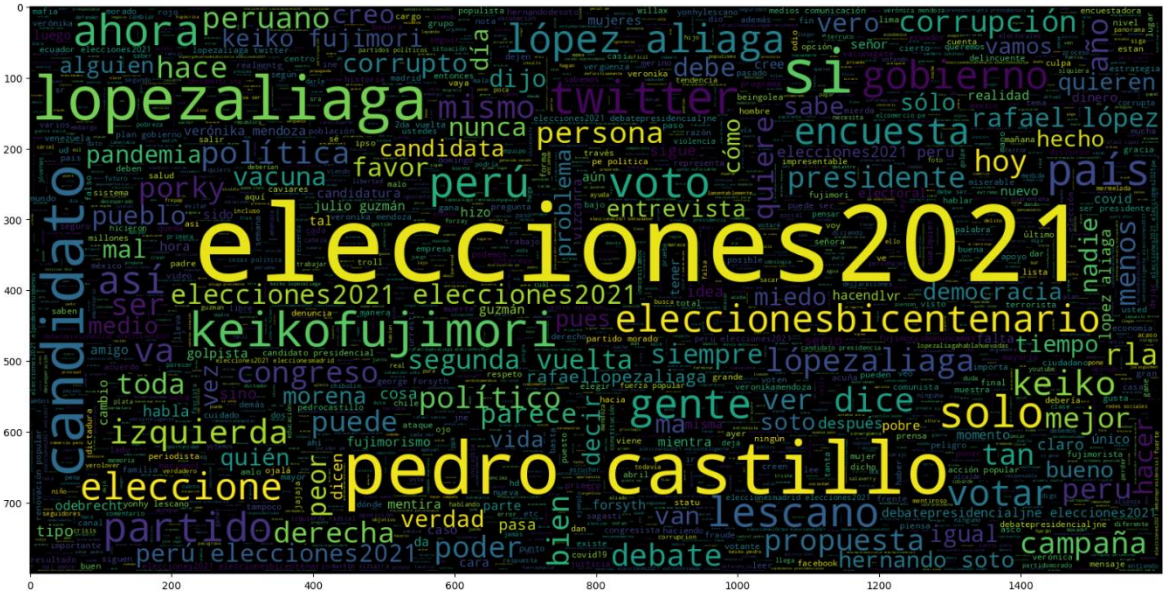

Figura 14. Nube de palabras para *tweets* etiquetados como positivos

Figura 15. Nube de palabras para *tweets* etiquetados como negativos

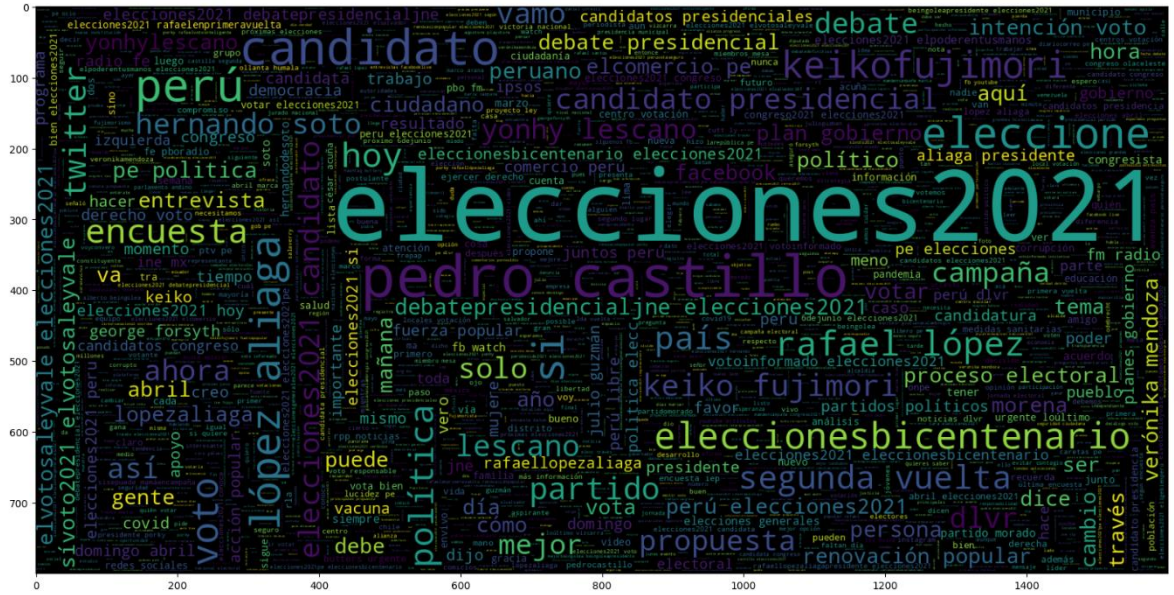

Figura 16. Nube de palabras para *tweets* etiquetados como neutrales

Posteriormente, para evaluar el desempeño de las técnicas *Machine Learning* seleccionadas, se estableció la misma configuración con los parámetros necesarios para su aplicación. Los experimentos se ejecutaron en una laptop con las siguientes características: procesador AMD RYZEN 7 2700U, tarjeta gráfica Radeon Vega Mobile Gfx (8 CPUs), 2.2GHz, memoria RAM de 8GB.

Primero, se codificó el atributo *Polarity* del *dataset* a valores enteros (1,-1,0) para sus equivalentes positivo, negativo y neutral respectivamente. Con este paso se generó un nuevo atributo denominado *PolarityValue* para almacenar esos valores. Luego, se dividió el conjunto de datos en dos grupos, considerando el 80% para datos de entrenamiento y 20% para datos de prueba.

Seguidamente, se extrajeron las características de los *tweets* obtenidos: número de oraciones, número de palabras, número de palabras únicas, número de letras, número de signos de puntuación, número de palabras y letras en mayúsculas, número de palabras vacías y longitud promedio de cada palabra. Luego, se generaron los N-gramas para tener un mejor panorama de los términos utilizados en los *tweets*. Las figuras 17, 18 y 19 muestran los N-gramas generados para 1 término, 2 términos y 3 términos respectivamente.

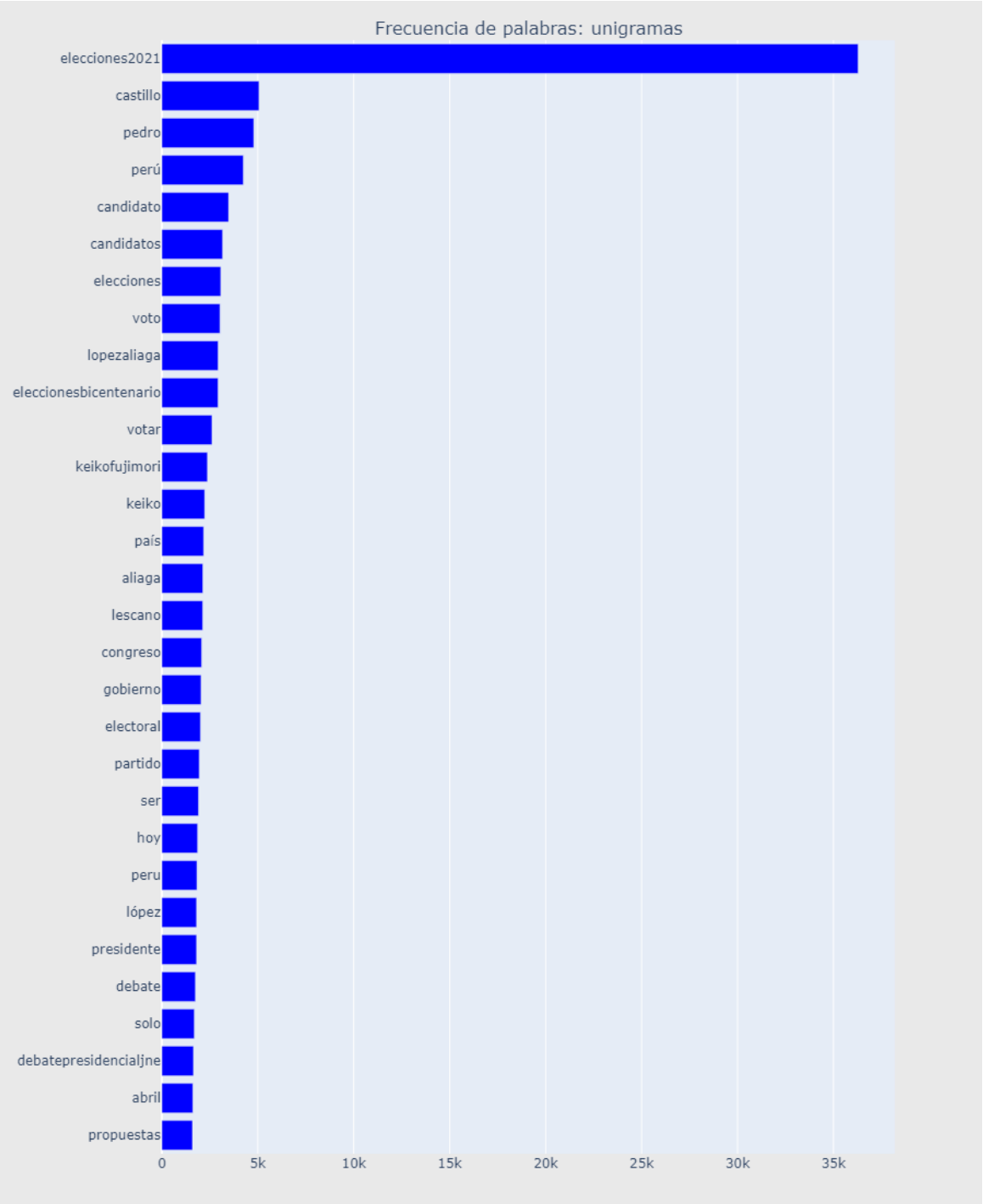

Figura 17. Frecuencia de palabras para unigramas

Como se observa en la Figura 17, el unigrama que se ubica en primer lugar es “Elecciones2021”, seguido por “castillo”, “pedro”, “perú”, “candidato”, “candidatos”, “elecciones”, “voto”, “lopezaliaga”, “eleccionesbicentenario”, entre otros.

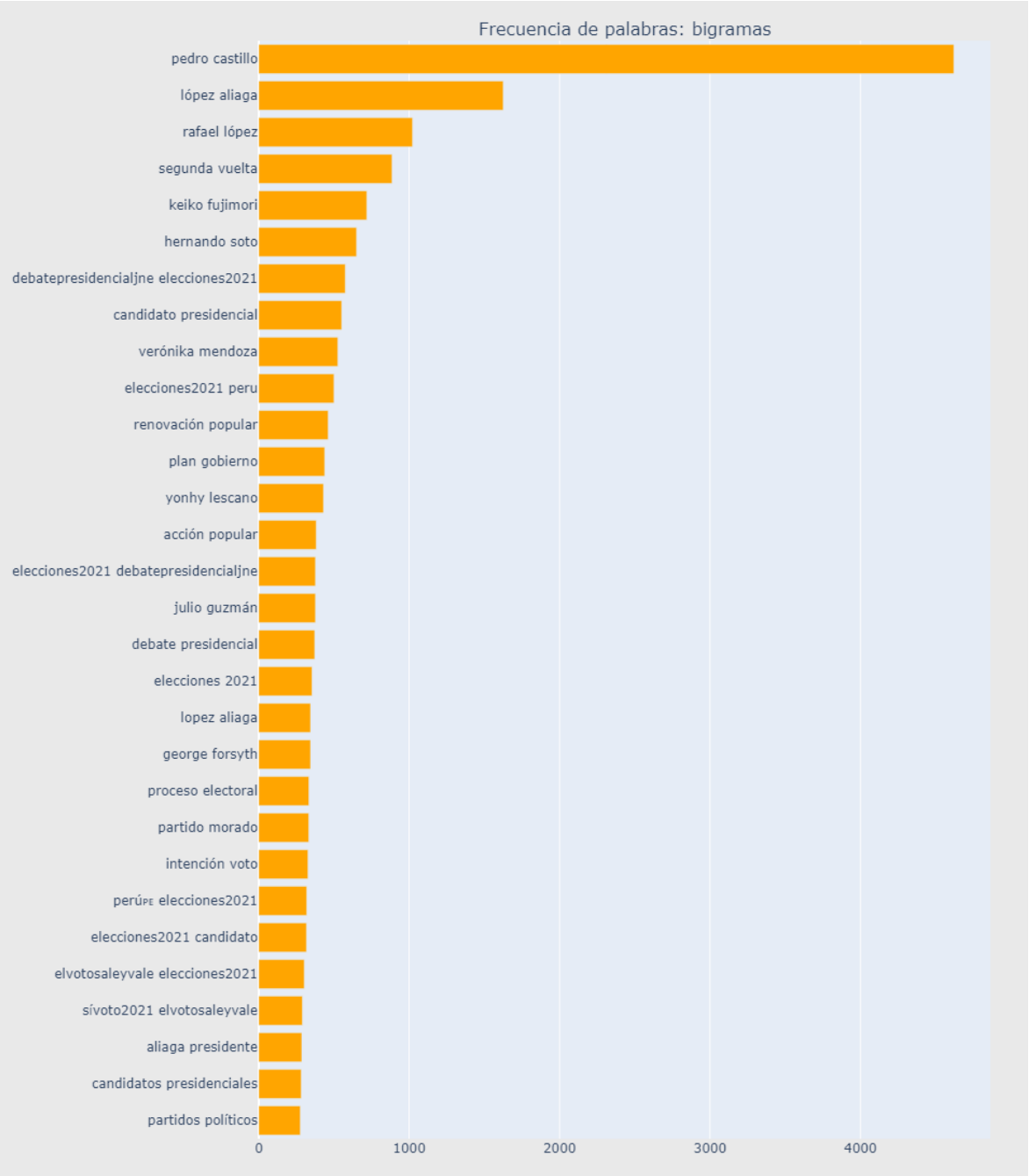

Figura 18. Frecuencia de palabras para bigramas

Para el caso de los bigramas, la Figura 18 muestra en primer lugar al bigrama “pedro castillo”, seguido por “lopez aliaga”, “rafael lópez”, “Keiko fujimori”, “hernando soto”, “debatepresidencialjne elecciones2021”, “candidato presidencial”, “elecciones 2021”, “verónica mendoza”, “elecciones2021 peru”, “renovación popular”, “yonhy lescano”, entre otros.

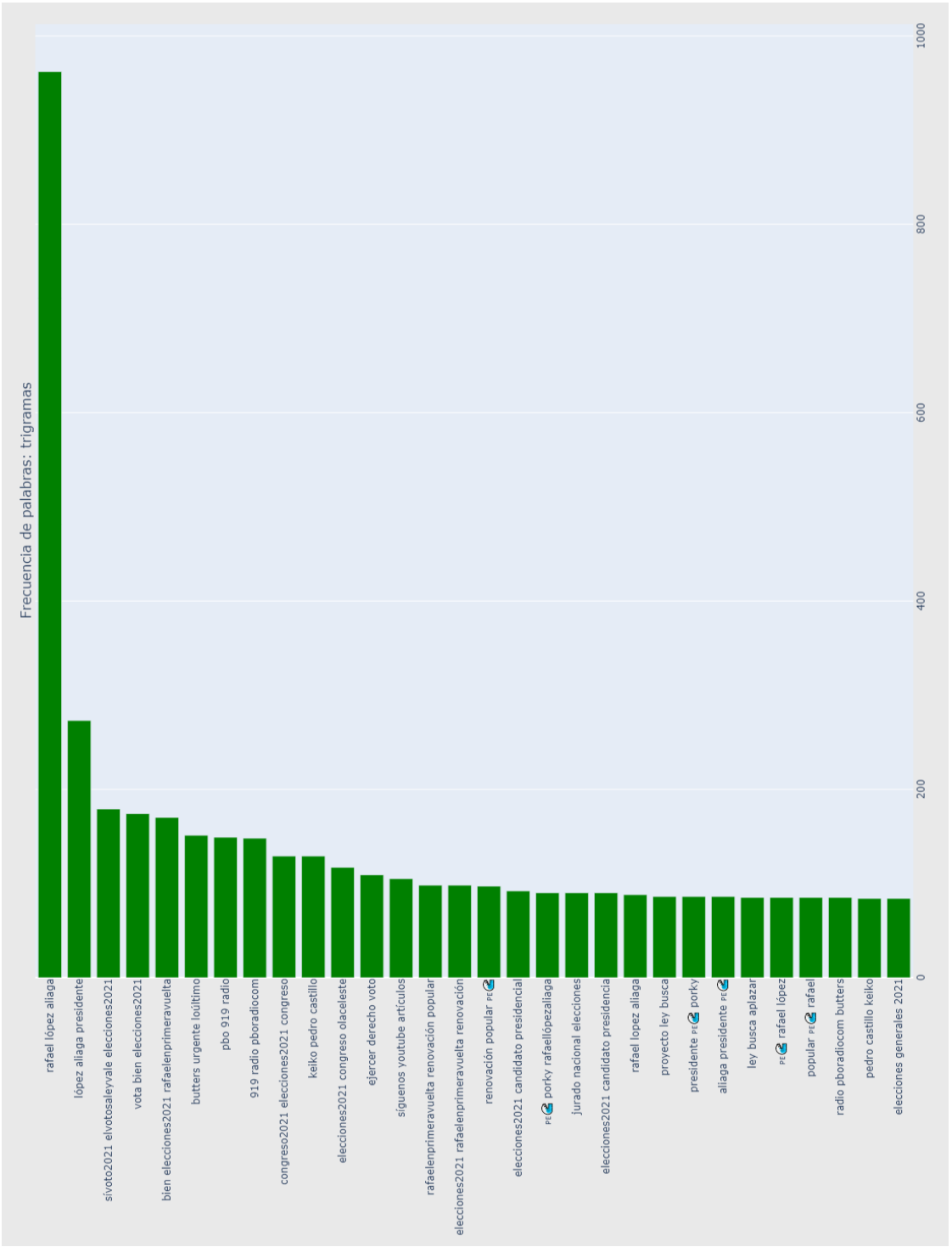

Figura 19. Frecuencia de palabras para trigramas

En Figura 19, el trigrama “Rafael López aliaga” está en primer lugar, seguido por “lópez aliaga presidente”, “vota bien elecciones2021”, entre otros.

En las Figuras 18 y 19, se observa que los candidatos presidenciales con mayor popularidad para un análisis a nivel de bigramas y trigramas son “Pedro Castillo” y “Rafael López Aliaga”. Este resultado también se refleja analizado la intención de voto mostrada en los *tweets* positivos, los cuales fueron verificados manualmente, considerando los *hashtags* más utilizados y el contenido de cada *tweet*. De acuerdo a este análisis se reconoce a “Rafael López” del partido Renovación Popular, como el candidato con mayor porcentaje de *tweets* positivos, alcanzando un 16.76%, seguido por los candidatos presidenciales “Pedro Castillo” del partido Perú Libre con 12.65%, “Alberto Beingolea” del partido PPC con 7.44%, “Keiko Fujimori” con 6.33%, “Verónika Mendoza” con 5.74%, “Julio Guzmán” del Partido Morado con 4.37% y “Hernando de Soto” del partido Avanza País con 3.78%. El resto de candidatos obtuvo cifras por debajo del 2%. Estos resultados confirman que el mayor porcentaje de usuarios y *tweets* provienen del departamento de Lima (Zorrilla, 2022), lo que explica la intención de voto por esos candidatos presidenciales.

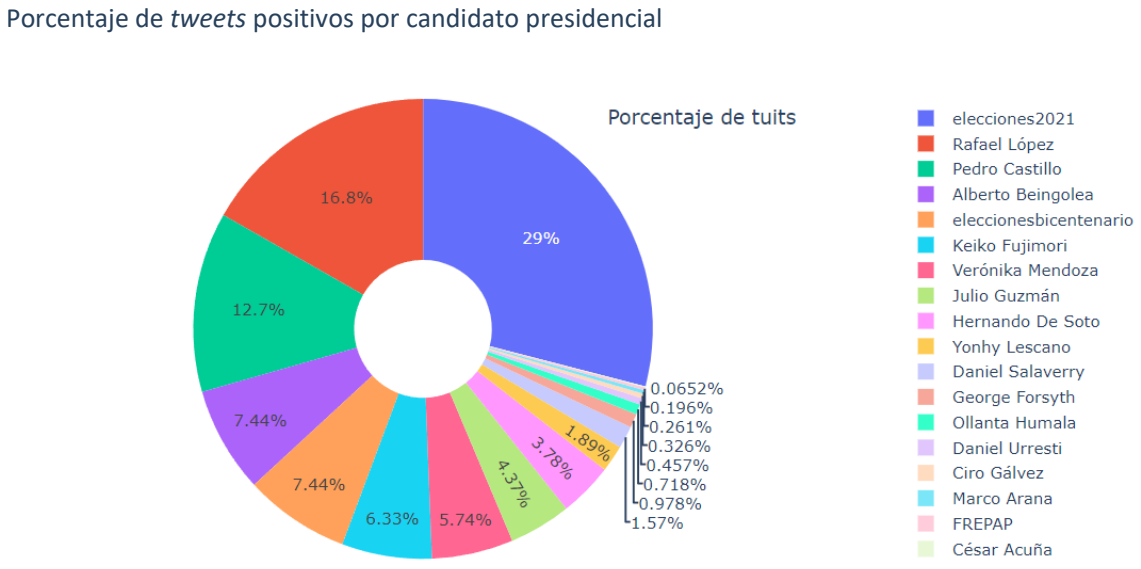

Figura 20. Porcentaje de *tweets* positivos por candidato presidencial

Otro resultado que se observa en la Figura 20, es el obtenido a partir de los *tweets* positivos, los cuales fueron etiquetados según el candidato presidencial al que se referían, considerando además los *hashtags* contenían. Los resultados demostraron que los *hashtags* “eleccionesbicentenario” y “elecciones2021” fueron tendencia, representando

un 29.03% y 7.44% respectivamente, haciendo un total de 36.47%. Este último, claro está, muestra las opiniones positivas en relación a las elecciones presidenciales Perú 2021 en términos generales, pero no específica una preferencia política, por lo que representa a los electores indecisos o votos en blanco o nulos. Resultados similares se obtuvieron a poco menos de un mes para las elecciones generales, en sondeos realizados por algunas encuestadoras como la Compañía Peruana de Estudios de Mercados y Opinión Pública (CPI), donde el 23.2% de los votantes aún no se decidían por un candidato y un 15.6% marcarían en blanco o viciado, haciendo un total de 38.8%; por otro lado, la encuestadora Datum reveló que el 21% eran indecisos y un 16% marcarían en blanco o viciado, sumando un 37%, en cambio, el Instituto de Estudios Peruanos (IEP) mostró que el 16.6% no votarían, un 2.2% lo harían en blanco o viciado y un 11% no sabían por quién votar, haciendo un total de 31% (Goberna Consultoría Política, 2021).

Adicionalmente, para cuantificar la temática de los *tweets*, así como la importancia de cada término que lo forma y reforzar la extracción de características, se utilizó TF-IDF. Sumado a esto, para crear una representación numérica de los *tweets*, identificando el conjunto de datos formado por todas las palabras que aparecen en los *tweets*, se utilizó BOW. Donde, para el caso de BOW, se utilizó la clase *CountVectorizer*, y para TF-IDF, se utilizó *TfidfVectorizer*. La extracción de características se aplicó a los grupos de prueba y entrenamiento, para cada una de las técnicas *Machine Learning* elegidas. En las Figuras 21 y 22, se muestran los ajustes realizados para trabajar con estas técnicas de extracción de características.

```
# Settings for Bag of words
cv = CountVectorizer (analyzer = 'word',
                      tokenizer = word_tokenize,
                      stop_words = stopword_list,
                      min_df = 5,
                      max_df = 0.9,
                      binary = False,
                      ngram_range = (1,3))

#transformed train tweets
cv_train_tweets=cv.fit_transform(x_train)
#transformed test tweets
cv_test_tweets=cv.transform(x_test)

print('BOW_cv_train:',cv_train_tweets.shape)
print('BOW_cv_test:',cv_test_tweets.shape)
```

Figura 21. Ajuste de parámetros para BOW y N-Gramas

Los ajustes realizados a la clase *CountVectorizer* convierten la columna ContentStopWords en una matriz en la que cada palabra es una columna cuyo valor es el número de veces que dicha palabra aparece en cada *tweet*, esto permite extraer y estructurar el contenido de cada *tweet* para poder analizarlo. Los parámetros que se configuran son: *analyzer*, se refiere al tipo de división que realiza el tokenizador, el cual es por palabras; *tokenizer*, se define la clase *word_tokenize* para convertir las cadenas de texto en una lista de palabras o tokens; *stop_words*, para eliminar las palabras vacías que no aportan valor sintáctico; *min_df*, para indicar el número mínimo de *tweets* en los que debe aparecer un término para no ser excluido en el tokenizado; *max_df*, para indicar la probabilidad de una palabra respecto al resto de palabras; *ngram_range*, para establecer el rango de n-gramas como (1,3) que incluyen unigramas, bigramas y trigramas. En cuanto a los ajustes de la clase *TfidfVectorizer*, son similares a los de la clase *CountVectorizer*.

```
# settings for TF-IDF
tv = TfidfVectorizer(analyzer = 'word',
                     tokenizer = word_tokenize,
                     stop_words = stopword_list,
                     min_df = 5,
                     max_df = 0.9,
                     use_idf = True,
                     ngram_range = (1,3))
#transformed train tweets
tv_train_tweets = tv.fit_transform(x_train)
#transformed test tweets
tv_test_tweets = tv.transform(x_test)
print('Tfidf_train:',tv_train_tweets.shape)
print('Tfidf_test:',tv_test_tweets.shape)
```

Figura 22. Ajuste de parámetros para TF-IDF y N-Gramas

Mediante las pruebas realizadas, la exactitud alcanzada por el modelo utilizando las diferentes técnicas *Machine Learning* depende de los ajustes realizados en el proceso de extracción de características para realizar la predicción de datos. De esta forma, en los resultados obtenidos para cada técnica utilizando BOW sobresalen LR y NB con un 75% y 76% de exactitud respectivamente, seguidos por SVM y DTC con un 74% y 62%. En cambio, para TF-IDF, destaca LR con un 79%, seguido por SVM con un 78%, luego NB con un 76% y finalmente DTC con un 63%. Estos resultados se muestran en la Tabla 9.

Tabla 9

Desempeño obtenido para cada técnica Machine Learning utilizando BOW y TF-IDF

Técnica *Machine Learning*	**BOW**	**TF-IDF**
Regresión Logística (LR)	0.75	0.79
Máquinas de Vectores de Soporte (SVM)	0.74	0.78
Naïve Bayes (NB)	0.76	0.76
Árboles de Decisión (DTC)	0.62	0.63

Mediante el análisis realizado, se demuestra que TF-IDF tiene mejor desempeño para cada una de las técnicas seleccionas. Por lo que el nivel de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 para cada uno de los candidatos será del 79% en términos de exactitud.

4.4. Resultado conforme al cuarto objetivo específico

Para evaluar el modelo de análisis de sentimientos en Twitter utilizado, se utilizaron las métricas descritas en el capítulo anterior. Las diferentes métricas fueron aplicadas a cada técnica o algoritmo de *Machine Learning* seleccionado, considerando que el modelo tiene más de dos clases (1,0,-1) para los *tweets* positivos, neutrales y negativos respectivamente, fue preciso calcular cada una de las métricas para cada clase, combinarlas y obtener una medida global.

4.4.1. Desempeño del algoritmo de Regresión Logística

Considerando BOW, el modelo predijo que 79% de los *tweets* eran positivos, de los cuales el 73% realmente lo son, resultando en un 76% para el valor-F1. Asimismo, se predijo que el 82% de los *tweets* eran negativos, de los cuales el 77% realmente lo son, obteniendo un 79% para el valor-F1. Estos resultados indicaron que el modelo predecía correctamente *tweets* positivos y negativos para BOW, con una exactitud del 75%. En cuanto a TF- IDF, el modelo predijo que 79% de los *tweets* eran positivos, de los cuales el 76% realmente lo son, resultando en un 78% para el valor-F1. Esto indicó que el modelo predecía correctamente *tweets* positivos y negativos para TF-IDF con una exactitud de 79%.

Es decir, el modelo de análisis de sentimientos basado en Regresión Logística logra un nivel de asertividad del 79% en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 para cada uno de los candidatos presidenciales en relación a los *tweets* positivos y negativos.

	precision	recall	f1-score	support
1	0.79	0.73	0.76	4106
-1	0.82	0.77	0.79	5429
0	0.29	0.68	0.41	457
accuracy			0.75	9992
macro avg	0.63	0.73	0.65	9992
weighted avg	0.78	0.75	0.76	9992

	precision	recall	f1-score	support
1	0.79	0.76	0.78	4106
-1	0.79	0.86	0.82	5429
0	0.75	0.18	0.29	457
accuracy			0.79	9992
macro avg	0.78	0.60	0.63	9992
weighted avg	0.79	0.79	0.78	9992

Figura 23. Desempeño del algoritmo de Regresión Logística para BOW y TF-IDF

4.4.2. Desempeño del algoritmo de Máquinas de Vectores de Soporte

Para el caso de BOW, el modelo predijo que 81% de los *tweets* eran positivos, de los cuales el 58% realmente lo son, resultando en un 68% para el valor-F1. Asimismo, se predijo que el 71% de los *tweets* eran negativos, de los cuales el 90% realmente lo son, obteniendo un 79% para el valor-F1. Estos resultados indicaron que el modelo predecía correctamente *tweets* positivos y negativos para BOW, con una exactitud del 74%. En cuanto a TF- IDF, el modelo predijo que el 79% de los *tweets* eran positivos, de los cuales el 74% realmente lo son, resultando en un 76% para el valor-Fl. Para los *tweets* negativos, se predijo que el 77% lo eran, de los cuales el 87% realmente lo son, obteniéndose un 81% para el valor-F1. Estos resultados indicaron que el modelo predecía correctamente *tweets* positivos y negativos para TF-IDF con una exactitud de 78%.

Es decir, el modelo de análisis de sentimientos basado en Máquinas de Vectores de Soporte tiene un nivel de asertividad del 78% en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 para cada uno de los candidatos presidenciales en relación a los *tweets* positivos y negativos.

	precision	recall	f1-score	support
1	0.81	0.58	0.68	4106
-1	0.71	0.90	0.79	5429
0	0.41	0.16	0.23	457
accuracy			0.74	9992
macro avg	0.65	0.55	0.57	9992
weighted avg	0.74	0.74	0.72	9992

	precision	recall	f1-score	support
1	0.79	0.74	0.76	4106
-1	0.77	0.87	0.81	5429
0	0.88	0.05	0.10	457
accuracy			0.78	9992
macro avg	0.81	0.55	0.56	9992
weighted avg	0.78	0.78	0.76	9992

Figura 24. Desempeño del algoritmo de Máquinas de Vectores de Soporte para BOW y TF-IDF

4.4.3. Desempeño del algoritmo de *Naïve Bayes*

Aplicando la técnica de *Naïve Bayes*, primero considerando BOW, el modelo predijo que 71% de los *tweets* eran negativos, de los cuales el 83% realmente lo son, resultando en un 76% para el valor-F1. En relación a los *tweets* negativos, se predijo que un 81% lo eran, siendo realmente negativos el 76%, obtenido un valor-F1 del 79%. Esto indicó que el modelo predecía *tweets* positivos y negativos para BOW con una exactitud del 76%. En cambio, para TF-IDF, el modelo predijo que el 72% de los *tweets* eran positivos, de los cuales el 82% realmente lo son, resultando en un 77% para el valor-Fl. Para los *tweets* negativos, se predijo que el 80% lo eran, de los cuales el 78% realmente lo son, obteniéndose un 71% para el valor-F1. Estos resultados indicaron que el modelo predecía correctamente *tweets* positivos y negativos para TF-IDF con una exactitud de 76%.

Es decir, el modelo de análisis de sentimientos basado en *Naïve Bayes* alcanza un nivel de asertividad del 76% en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 para cada uno de los candidatos presidenciales en relación a los *tweets* positivos y negativos.

	precision	recall	f1-score	support
1	0.71	0.83	0.76	4106
-1	0.81	0.76	0.79	5429
0	0.54	0.06	0.10	457
accuracy			0.76	9992
macro avg	0.69	0.55	0.55	9992
weighted avg	0.76	0.76	0.75	9992

	precision	recall	f1-score	support
1	0.72	0.82	0.77	4106
-1	0.80	0.78	0.79	5429
0	0.60	0.01	0.03	457
accuracy			0.76	9992
macro avg	0.71	0.54	0.53	9992
weighted avg	0.76	0.76	0.75	9992

Figura 25. Desempeño del algoritmo de *Naïve Bayes* para BOW y TF-IDF

4.4.4. Desempeño del algoritmo de Árboles de Decisión

Utilizando el clasificador de Árboles de Decisión, y considerando BOW, el modelo predijo que 66% de los *tweets* eran negativos, de los cuales el 37% realmente lo es, resultando en un 46% para el valor-F1. En relación a los *tweets* negativos, se predijo que un 61% lo eran, siendo realmente negativos el 87%, obtenido un valor-F1 del 72%. Esto indicó que el modelo predecía *tweets* positivos y negativos para BOW con una exactitud del 62%. En cambio, para TF-IDF, el modelo predijo que el 67% de los *tweets* eran positivos, de los cuales el 37% realmente lo es, resultando en un 48% para el valor-Fl. Para los *tweets* negativos, se predijo que el 62% lo eran, de los cuales el 87% realmente lo son, obteniéndose un 72% para el valor-F1. Estos resultados indicaron que el modelo predecía correctamente *tweets* positivos y negativos para TF-IDF con una exactitud de 63%.

Es decir, el modelo de análisis de sentimientos basado en Árboles de Decisión obtiene un nivel de asertividad del 63% en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 para cada uno de los candidatos presidenciales en relación a los *tweets* positivos y negativos.

	precision	recall	f1-score	support
1	0.66	0.37	0.47	4106
-1	0.61	0.87	0.72	5429
0	0.14	0.00	0.00	457
accuracy			0.62	9992
macro avg	0.47	0.41	0.40	9992
weighted avg	0.61	0.62	0.59	9992

	precision	recall	f1-score	support
1	0.67	0.37	0.48	4106
-1	0.62	0.87	0.72	5429
0	0.20	0.00	0.00	457
accuracy			0.63	9992
macro avg	0.49	0.42	0.40	9992
weighted avg	0.62	0.63	0.59	9992

Figura 26. Desempeño del algoritmo de Árboles de Decisión para BOW y TF-IDF

4.4.5. Comparación de resultados para *tweets* positivos, negativos y neutrales

La Tabla 10 muestra el desempeño alcanzado de cada técnica utilizada para los *tweets* positivos, negativos y neutrales, utilizando BOW. Se realizó el análisis de los mejores resultados en cada técnica teniendo en cuenta la precisión, exhaustividad y valor-F1, marcando en negrita el valor con mejor asertividad. Los *tweets* positivos arrojaron resultados altos para las 4 técnicas utilizadas con valores igual a 0.79, 0.73, 0.76, 0.81, 0.71, 0.83 y 0.76. Le siguen los *tweets* negativos con valores igual a 0.82, 0.77, 0.79, 0.71, 0.90, 0.79, 0.81, 0.76, 0.79, 0.87 y 0.72. Entonces, puede afirmarse que las técnicas de LR, SVM y NB son adecuados para el contexto, siendo DTC la única técnica con valores por debajo a los obtenidos por las demás.

Tabla 10

Resultados obtenidos para tweets positivos, negativos y neutrales utilizando BOW

	LR			**SVM**			**NB**			**DTC**		
	Precisión	**Exhaustividad**	**Valor-F1**	**Precisión**	**Exhaustividad**	**Valor-F1**	**Precisión**	**Exhaustividad**	**Valor-F1**	**Precisión**	**Exhaustividad**	**Valor-F1**
1	**0.79**	**0.73**	**0.76**	**0.81**	0.58	0.68	**0.71**	**0.83**	**0.76**	0.66	0.37	0.47
-1	**0.82**	**0.77**	**0.79**	**0.71**	**0.90**	**0.79**	**0.81**	**0.76**	**0.79**	0.61	**0.87**	**0.72**
0	0.29	0.68	0.41	0.41	0.16	0.23	0.54	0.06	0.10	0.14	0.00	0.00

La Tabla 11 muestra el desempeño alcanzado de cada técnica utilizada para los *tweets* positivos, negativos y neutrales, utilizando TF-IDF. Se realizó el análisis de los mejores resultados en cada técnica teniendo en cuenta la precisión, exhaustividad y valor-F1, marcando en negrita el valor con mejor asertividad. Los *tweets* positivos arrojaron resultados altos para las 4 técnicas utilizadas con valores igual a 0.79, 0.76, 0.78, 0.79, 0.74, 0.76, 0.72, 0.82 y 0.77. Le siguen los *tweets* negativos con valores igual a 0.79, 0.86, 0.82, 0.77, 0.87, 0.81, 0.80, 0.78, 0.79, 0.87 y 0.72. Entonces, puede afirmarse que las técnicas de LR, SVM y NB son adecuados para el contexto. siendo DTC la única técnica con valores por debajo a los obtenidos por las demás, al igual que con BOW.

Tabla 11

Resultados obtenidos para tweets positivos, negativos y neutrales utilizando TF-IDF

	LR			SVM			NB			DTC		
	Precisión	**Exhaustividad**	**Valor-F1**	**Precisión**	**Exhaustividad**	**Valor-F1**	**Precisión**	**Exhaustividad**	**Valor-F1**	**Precisión**	**Exhaustividad**	**Valor-F1**
1	**0.79**	**0.76**	**0.78**	**0.79**	**0.74**	**0.76**	**0.72**	**0.82**	**0.77**	0.67	0.37	0.48
-1	**0.79**	**0.86**	**0.82**	**0.77**	**0.87**	**0.81**	**0.80**	**0.78**	**0.79**	0.62	**0.87**	**0.72**
0	**0.75**	0.18	0.29	**0.88**	0.05	0.10	0.60	0.01	0.03	0.20	0.00	0.00

En cuanto al porcentaje de votos obtenidos en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 y el porcentaje de votos emitidos en el proceso electoral se ubican a los candidatos presidenciales Rafael López, Pedro Castillo y Keiko Fujimori en los primeros puestos (ONPE, 2021), estos resultados se muestran en la Tabla 12. Si bien estos no son iguales, se acercan mucho e identifican el porcentaje de votos para los candidatos que pasaron a segunda vuelta, a saber, Pedro Castillo y Keiko Fujimori. Siendo esta última una sorpresa, sin embargo refleja la intención de voto de Lima, donde se concentra el 34.7% de la población total (IEP, 2021), además, el 81% de tweets provienen de también de Lima (Zorrilla, 2022). Entonces, se comprende que los candidatos presidenciales más populares en la zona de Lima sean Rafael López, Keiko Fujimori y Hernando de Soto.

Tabla 12

Resultados obtenidos en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021

Candidato Presidencial	% Votos en la predicción de intención de voto	% Votos emitidos en el proceso electoral
Rafael López	**16.76%**	**9.553%**
Pedro Castillo	**12.65%**	**15.382%**
Keiko Fujimori	**6.33%**	**10.900%**
Verónika Mendoza	5.74%	6.394%
Yonhy Lescano	1.89%	7.374%
Hernando de Soto	3.78%	**9.451%**
George Forsyth	0.98%	4.598%
Cesar Acuña	0.07%	4.895%
Otros	**15.35%**	**12.749%**
Indecisos, votos en blanco o viciados	**36.47%**	**18.704%**
Total de votos	**100.00%**	**100.00%**

4.5. Discusión

Las medidas de desempeño usadas en las Tablas 10 y 11 indican que los algoritmos de Regresión Logística, Máquinas de Vectores de Soporte y *Naïve Bayes* demostraron ser mejores logrando alcanzar un valor de 79% de exactitud. Sin embargo, tras los experimentos realizados es el algoritmo de Regresión Logística el que demostró mejores resultados en relación a los otros. La comparación del algoritmo de Regresión Logística con otros trabajos se muestra en la Tabla 13.

Tabla 13

Comparación de desempeño del algoritmo de Regresión Logística con otros trabajos en el análisis de sentimientos en Twitter

N°	Dataset	Nro. de Tweets	Características	Algoritmo	Exactitud	Referencia
1	Propio	4242	TF-IDF, BOW y N-gramas	LR	57%	Ahuja *et al.* (2019)
2	Propio	3896	TF-IDF	LSTM	77%	Ansari *et al.* (2020)
				LR	71%	
3	Propio	18,432,811	TF-IDF, N-gramas	NB	94.58%	Chaudhry *et al.* (2021)
4	Elecciones Bicentenario 2021 Tweets	49,916	TF-IDF, BOW	LR	79%	Modelo propuesto

En la determinación de las técnicas de *Machine Learning* más adecuadas para el análisis de sentimientos en Twitter en el contexto político, se eligieron aquellas que tienen mayor popularidad en el área de estudio, en razón al nivel de predicción y las ventajas que tienen frente a otras técnicas como lo son las basadas en *lexicons* (Arcila-Calderón *et al.*, 2017; Sudhir & Suresh, 2021). Asimismo, se compararon los resultados rescatados a partir de los datos publicados en el estado del arte, donde se utilizaron diferentes algoritmos de *Machine Learning* para el análisis de sentimientos en Twitter en el contexto político, y es el algoritmo de Regresión Logística el que obtuvo resultados sobresalientes (Ahuja *et al.*, 2019; Ansari *et al.*, 2020; Mohamed *et al.*, 2020) alcanzando un 79% de exactitud. Además, los resultados obtenidos por Ahuja *et al.* (2019), donde también se consideraron las características TF-IDF, BOW N-gramas en el análisis de sentimientos en Twitter, muestran que TF-IDF tiene mejores resultados (3-4%) comparados a los de BOW, al igual que en el modelo propuesto donde se obtuvo 75% para BOW y 79% para TF-IDF en términos de exactitud, demostrándose que RL proporciona las mejores predicciones de sentimientos al brindar el máximo rendimiento para los cuatro parámetros de evaluación (exactitud, exhaustividad, precisión y valor-F1).

Ahora bien, en comparación con el estudio realizado por Ansari *et al.* (2020) donde se emplean exactitud, exhaustividad, precisión y valor-F1 para la evaluación de los algoritmos de clasificación de sentimientos, destacan LSTM, seguido por RF y LR. Sin embargo, aunque LSTM tiene mejor desempeño en términos de exactitud, implica un alto costo computacional con un tiempo de entrenamiento de 82.7816 segundos, mientras que

RF tarda 0.8587 segundos y LR tarda solo 0.0519 segundos, siendo este último el mejor en este aspecto. Además, comparando los valores alcanzados por LR para exhaustividad (73%), precisión (72%) y valor-F1 (70%), con los obtenidos en el presente estudio, para exhaustividad (86%), precisión (79%) y valor-F1 (82%), se observa una notable mejoría en los resultados alcanzados.

Gracias a los resultados obtenidos, puede señalarse que LR proporciona uno de los mejores resultados en predicción de sentimientos relacionados a la intención de voto para las métricas señaladas, exactitud, exhaustividad, precisión y valor-F1.

Sin embargo, comparando los resultados obtenidos con el trabajo desarrollado por Chaudry *et al.* (2021), donde los niveles de asertividad alcanzados por NB en relación a la exactitud es de 94.58% y para la precisión es del 93.19%, es preciso señalar que el estudio se realiza en torno a dos candidatos presidenciales, Joe Biden y Donald Trump, se realizó en un contexto mucho más grande que el de Perú, con un tamaño de datos también más numeroso, y durante un período de tiempo más corto y cercano al proceso electoral (del 28 de septiembre al 20 de noviembre de 2020). En cambio, para el caso peruano, se consideran más de 20 candidatos presidenciales (ONPE, 2021), por lo que los resultados se segmentan con mayor facilidad. Así mismo, el período de tiempo es más largo y lejano al proceso electoral (del 1 de enero al 11 de abril de 2021).

Por otro lado, en relación a la intención de voto, las encuestadoras CPI, Datum e IEP, señalaron que los primeros puestos estaban ocupados por Yonhy Lescano, Rafael López, George Forsyth, Keiko Fujimori y Verónika Mendoza a poco menos de un mes del proceso electoral (Goberna Consultoría Política, 2021). Sin embargo, los resultados cambiaron luego de los esperados debates presidenciales, donde se observó un ascenso del candidato presidencial Pedro Castillo, considerándolo dentro de los 7 candidatos con mayor intención de voto junto a Keiko Fujimori, Hernando de Soto, Rafael López, Yonhy Lescano, Verónika Mendoza y George Forsyth (IEP, 2021). Estos resultados se reflejan en la Tabla 12, donde se muestran los candidatos presidenciales con mayor intención de voto en las Elecciones Presidenciales Perú 2021.

4.6. Prueba de hipótesis

La prueba de hipótesis del modelo de análisis de sentimientos en Twitter basado en técnicas de *Machine Learning* para determinar la intención de voto, se planteó de la siguiente manera:

Planteamiento de la prueba de hipótesis

$$H_0: \mu \leq \mu_0$$

$$H_1: \mu > \mu_0$$

La hipótesis nula H_0 señala que μ es menor o igual que μ_0 y la hipótesis alterna nos indica que μ es mayor que μ_0.

Nivel de significancia

Para comprobar la prueba de hipótesis, se consideró un nivel de significancia de 0,05 o 5% de error:

$$\propto = 0.05 = 5\%$$

Se utilizó la distribución *t-Student* para el área de una cola en relación al grado de libertad GL, n – 1 = 2, done n es la muestra, por lo tanto, el valor de $t_{\propto}$ es:

$$t_{\propto} = t_{0,05} = 2.92$$

Se empleó la desviación estándar de muestra s, donde x es el promedio de asertividad alcanzado para cada clase de *tweet* (positivo, negativo y neutral), $\bar{x}$ es el promedio de asertividad alcanzado por los cuatro modelos de *Machine Learning* utilizados, y n es el número de clases de *tweets*. Siendo s igual a:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{166}{2}} = 9.11$$

Zona de rechazo y regla de decisión

Se utilizó el estadístico de prueba *t-Student*, considerando un $n = 3$; siendo $n < 30$.

$$\mu = \left(\bar{\chi} - t_{\alpha}\frac{s}{\sqrt{n}}\right) = \left(74 - 2.92\frac{9.11}{\sqrt{3}}\right) = 58.64$$

Estadística de prueba

Se aplicó el estadístico de prueba *t-Student* para cada uno de los modelos de *Machine Learning* utilizados:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Respecto al modelo de Regresión Logística

H_0: El modelo de análisis de sentimientos en Twitter basado en Regresión Logística no obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

H_1: El modelo de análisis de sentimientos en Twitter basado en Regresión Logística obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Tabla 14

Prueba de muestra única en función del modelo de Regresión Logística

Prueba de muestra única				
Modelo	**Nivel de asertividad**			
RL	N especies	X (%) exactitud	S (%)	μ (%)
	3	79	9.11	58.64

$t = \frac{79-58.64}{\frac{9.11}{\sqrt{3}}} = 3.87$; se usaron los datos de la Tabla 14 en el reemplazo de la ecuación. Rechazándose la hipótesis nula, pues el valor de t = 3.87 está dentro de la zona de rechazo. Entonces, se acepta la hipótesis alterna. Lo que señala que el modelo de Regresión Logística obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Respecto al modelo de Máquinas de Vectores de Soporte

H_0: El modelo de análisis de sentimientos en Twitter basado en Máquinas de Vectores de Soporte no obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

H_1: El modelo de análisis de sentimientos en Twitter basado en Máquinas de Vectores de Soporte obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Tabla 15

Prueba de muestra única en función del modelo de Máquinas de Vectores de Soporte

Prueba de muestra única				
Modelo	**Nivel de asertividad**			
SVM	N especies	X (%) exactitud	S (%)	μ (%)
	3	78	9.11	58.64

$t = \frac{78-58.64}{\frac{9.11}{\sqrt{3}}} = 3.68;$ se usaron los datos de la Tabla 15 en el reemplazo de la ecuación. Rechazándose la hipótesis nula, pues el valor de t = 3.68 está dentro de la zona de rechazo. Entonces, se acepta la hipótesis alterna. Lo que señala que el modelo de Máquinas de Vectores de Soporte obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Respecto al modelo de *Naïve Bayes*

H_0: El modelo de análisis de sentimientos en Twitter basado en *Naïve Bayes* no obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

H_1: El modelo de análisis de sentimientos en Twitter basado en *Naïve Bayes* obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Tabla 16

Prueba de muestra única en función del modelo de Naïve Bayes

Prueba de muestra única				
Modelo	**Nivel de asertividad**			
NB	N especies	X (%) exactitud	S (%)	μ (%)
	3	76	9.11	58.64

$t = \frac{76-58.64}{\frac{9.11}{\sqrt{3}}} = 3.3$; se usaron los datos de la Tabla 16 en el reemplazo de la ecuación. Rechazándose la hipótesis nula, pues el valor de t = 3.3 está dentro de la zona de rechazo. Entonces, se acepta la hipótesis alterna. Lo que señala que el modelo de *Naïve Bayes* obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Respecto al modelo de Árboles de Decisión

H_0: El modelo de análisis de sentimientos en Twitter basado en Árboles de Decisión no obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

H_1: El modelo de análisis de sentimientos en Twitter basado en Árboles de Decisión obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Tabla 17

Prueba de muestra única en función del modelo de Regresión Logística

Prueba de muestra única				
Modelo	**Nivel de asertividad**			
DTC	N especies	X (%) exactitud	S (%)	μ (%)
	3	63	9.11	58.64

$t = \frac{63-58.64}{\frac{9.11}{\sqrt{3}}} = 0.83$; se usaron los datos de la Tabla 17 en el reemplazo de la ecuación. Aceptándose la hipótesis nula, pues el valor de t = 0.83 está dentro de la zona de aceptación. Entonces, se acepta la hipótesis nula y se rechaza la hipótesis alterna. Lo que señala que el modelo de Regresión Logística no obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

Finalmente, de las Tablas 14 – 17, se observa que el modelo de Regresión Logística obtiene los mejores resultados de asertividad en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021.

CONCLUSIONES

1. La técnica *Machine Learning* apropiada para un modelo de análisis de sentimientos en Twitter que obtuvo los mejores resultados en la predicción de intención de voto en las Elecciones Presidenciales Perú 2021 ha sido Regresión Logística, alcanzando un 75% de exactitud para BOW y un 79% de exactitud para TF-IDF, seguido por el algoritmo de Máquinas de Vectores de Soporte con un 74% de exactitud para BOW y 78% para TF-IDF, y finalmente, el algoritmo de *Naïve Bayes* con un 76% de exactitud tanto para BOW como para TF-IDF. Asimismo, los candidatos con mayor intención de voto identificados a través del estudio fueron a Rafael López con un 16.76%, Pedro Castillo con un 12.65%, Alberto Beingolea con un 7.44%, Keiko Fujimori con un 6.33% y Verónika Mendoza con un 5.74%.

2. Se determinó que los algoritmos tradicionales de *Machine Learning* como Regresión Logística, *Naïve Bayes*, Máquinas de Vectores de Soporte y Árboles de Decisión, son los más adecuados para el análisis de sentimientos en Twitter en un contexto político, considerando su desempeño y popularidad en la revisión de la literatura realizada.

3. Se conformó el *dataset* Elecciones Bicentenario 2021 Tweets, compuesto por 49,916 *tweets* por 12 columnas o atributos. Este conjunto de datos está formado por *tweets* que incluyen los hashtags en tendencia durante las Elecciones Presidenciales Perú 2021, como #Elecciones2021 y #EleccionesBicentenario, los cuales sirvieron de filtro para la selección de *tweets*. Asimismo, se logró un ajuste del *dataset* eliminando los *tweets* que contenían hashtags ajenos a las Elecciones Presidenciales Perú 2021 y cuya ubicación estuviera relacionada a otro proceso electoral, logrando el esquema y atributos necesarios para su análisis mediante las técnicas de *Machine Learning* utilizadas.

4. Se diseñó un modelo de análisis de sentimientos en Twitter utilizando las técnicas más adecuadas de *Machine Learning* para determinar la intención de voto en las Elecciones Presidenciales Perú 2021, logrando recuperar un total de 104,916 *tweets* para luego afinarlo y reducirlo a un *dataset* de 49,916 *tweets*. Asimismo, para el entrenamiento y prueba del modelo, se realizaron los ajustes necesarios utilizando BOW y TF-IDF para cada técnica de *Machine Learning*. Los ajustes fueron pertinentes, porque las técnicas obtuvieron resultados sobresalientes en los distintos parámetros de evaluación.

5. Se fijaron las métricas necesarias para evaluar el desempeño de los cuatro algoritmos *Machine Learning* seleccionados para el estudio. Los algoritmos fueron evaluados en términos de exactitud, precisión, exhaustividad y valor-F1, considerando, además las técnicas BOW y TF-IDF, las cuales fueron aplicadas a cada uno de los modelos. Las métricas utilizadas permitieron conocer de manera comparativa y detallada el desempeño de cada algoritmo, siendo el algoritmo de Regresión Logística el más asertivo.

RECOMENDACIONES

Se recomienda en trabajos posteriores relacionados con el análisis de sentimientos en Twitter para predecir la intención de voto utilizar diferentes enfoques como los híbridos incorporando *lexicons* o técnicas de *Deep Learning* para fortalecer y/o mejorar el desempeño del modelo.

Se recomienda usar el conjunto de datos Elecciones Bicentenario Tweets Data conformado y aplicar en él otras técnicas de *Machine Learning* u otros enfoques de análisis de sentimientos. Asimismo, se sugiere ajustar más el *dataset* revisando los *tweets*, debido a que un gran porcentaje no tiene una ubicación conocida y podría haber aún existencias de *tweets* relacionados a otros procesos electorales, lo que afectaría el desempeño del modelo.

Se recomienda realizar cambios en el modelo de análisis de sentimientos propuesto, incorporando tareas de lematización y/o *steaming* para comprobar si los resultados obtenidos en términos de exactitud mejoran.

Se recomienda el uso de técnicas de codificación para variables categóricas, a través del uso de librerías como *Label Encoder, Ordinal Encoder* o *category_encoders*, y de esta forma mejorar el desempeño del algoritmo de Árboles de Decisión en términos de exactitud, ya que tiene un mejor funcionamiento con variables numéricas que categóricas.

BIBLIOGRAFÍA

Abdullah, M., & Hadzikadic, M. (2018). Sentiment analysis of twitter data: Emotions revealed regarding donald trump during the 2015-16 primary debates. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, *2017-Novem*, 760–764. https://doi.org/10.1109/ICTAI.2017.00120

Ahuja, R., Chug, A., Kohli, S., Gupta, S., & Ahuja, P. (2019). The impact of features extraction on the sentiment analysis. *Procedia Computer Science*, *152*, 341–348. https://doi.org/10.1016/j.procs.2019.05.008

Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., & Singh, K. P. (2020). Analysis of Political Sentiment Orientations on Twitter. *Procedia Computer Science*, *167*, 1821–1828. https://doi.org/10.1016/j.procs.2020.03.201

Ansari, M. Z., Siddiqui, A. F., & Anas, M. (2021). Inferring Political Preferences from Twitter. *Lecture Notes in Networks and Systems*, *164*, 581–589. https://doi.org/10.1007/978-981-15-9774-9_54

Aramburo, R. F. P., Moreira, M. Â. L., Fávero, L. P. L., De Araújo Costa, I. P., & Dos Santos, M. (2022). Data Science in Social Politics with Particular Emphasis on Sentiment Analysis. *Procedia Computer Science*, *214*(C), 420–427. https://doi.org/10.1016/j.procs.2022.11.194

Arcila-Calderón, C., Ortega-Mohedano, F., Jiménez-Amores, J., & Trullenque, S. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático. *El Profesional de La Información*, *26*(5), 973. https://doi.org/10.3145/epi.2017.sep.18

Aung, K. Z., & Myo, N. N. (2017). Sentiment analysis of students' comment using lexicon based approach. *Proceedings - 16th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2017*, 149–154. https://doi.org/10.1109/ICIS.2017.7959985

Bansal, B., & Srivastava, S. (2018). On predicting elections with hybrid topic based sentiment analysis of tweets. *Procedia Computer Science*, *135*, 346–353. https://doi.org/10.1016/j.procs.2018.08.183

Bermingham, A., & Smeaton, A. F. (2011). On Using Twitter to Monitor Political Sentiment and Predict Election Results. *Proceedings of the Workshop on Sentiment*

Analysis Where AI Meets Psychology (SAAIP 2011), 2–10. Recuperado de: https://aclanthology.org/W11-3702

Berumen, E. (2022). *Intención de voto y estimación de voto ¿cuál es la diferencia?* Recuperado de: https://berumen.com.mx/intencion-de-voto-y-estimacion-de-voto-cual-es-la-diferencia/

Bohorquez, V., Mendez, C., Altube, L., & Santana, E. (2019). Identificación del sentimiento usando redes sociales en política. *AMCIS 2019 Proceedings, Twenty-Fifth Americas Conference on Information Systems*. Recuperado de: https://aisel.aisnet.org/amcis2019/spanish_portuguese_latin_america/spanish_portuguese_latin_america/41/

Britzolakis, A., Kondylakis, H., & Papadakis, N. (2021). Athppa: A data visualization tool for identifying political popularity over twitter. *Information (Switzerland)*, *12*(8), 1–24. https://doi.org/10.3390/info12080312

Cerón-Guzmán, J. A., & León-Guzmán, E. (2016). A sentiment analysis system of Spanish tweets and its application in Colombia 2014 presidential election. *Proceedings - 2016 IEEE International Conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom 2016*, 250–257. https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.47

Chamorro Alvarado, V. L. (2018). *Clasificación de tweets mediante modelos de aprendizaje supervisado* [Universidad Complutense de Madrid]. Recuperado de: https://eprints.ucm.es/id/eprint/49774/1/TFM Veronica Chamorro Alvarado.pdf

Chaudhry, H. N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z. I., Shoaib, U., & Janjua, S. H. (2021). Sentiment analysis of before and after elections: Twitter data of U.S. election 2020. *Electronics (Switzerland)*, *10*(17), 1–26. https://doi.org/10.3390/electronics10172082

Choy, M., Cheong, M. L. F., Laik, M. N., & Shung, K. P. (2011). *A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction*. http://arxiv.org/abs/1108.5520

Congreso Constituyente Perú. (1997). Ley Orgánica de Elecciones Perú. *Congreso Nacional*, *26859*.

Das, B. C., Anwar, M. M., & Sarker, I. H. (2020). Reducing Social Media Users' Biases to Predict the Outcome of Australian Federal Election 2019. *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2020*. https://doi.org/10.1109/CSDE50874.2020.9411633

El Comercio. (2021, April 15). *Elecciones 2021: ¿Qué pasa si voto en blanco o viciado?* 1–10. Recuperado de: https://elcomercio.pe/respuestas/elecciones-2021-que-pasa-si-voto-en-blanco-o-viciado-onpe-votos-segunda-vuelta-pedro-castillo-keiko-fujimori-comicios-revtli-noticia/

Fernández, M. (2019). *Análisis de opinión del microblogging Twitter por la clasificación al mundial de fútbol Rusia - 2018 de la selección peruana de fútbol, usando el framework Spark* [Universidad Nacional del Altiplano]. Recuperado de: http://repositorio.unap.edu.pe/handle/20.500.14082/13506

Figueroa, C. A. (2018). *Método de predicción electoral a través de modelo basado en análisis de sentimientos en Twitter* [Pontificia Universidad Católica de Valparaíso]. Recuperado de: http://repositorio.ucv.cl/handle/10.4151/93220

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. *International Journal of Information Management*, *51*(July), 102048. https://doi.org/10.1016/j.ijinfomgt.2019.102048

Goberna Consultoría Política. (2021). *La indecisión y el desinterés ocupan el primer lugar de cara a las elecciones de abril en Perú*. Recuperado de: https://goberna.pe/la-indecision-y-el-desinteres-ocupan-el-primer-lugar-de-cara-a-las-elecciones-de-abril-en-peru/

Gorodnichenko, Y., Pham, T., & Talavera, O. (2021). Social media, sentiment and public opinions: Evidence from #Brexit and #USElection. *European Economic Review*, *136*, 103772. https://doi.org/10.1016/j.euroecorev.2021.103772

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in US Election 2016 outcomes – Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, *145*(August), 438–460. https://doi.org/10.1016/j.techfore.2018.09.009

Haryanto, B., Ruldeviyani, Y., Rohman, F., Julius Dimas, T. N., Magdalena, R., &

Muhamad Yasil, F. (2019). Facebook analysis of community sentiment on 2019 Indonesian presidential candidates from Facebook opinion data. *Procedia Computer Science*, *161*, 715–722. https://doi.org/10.1016/j.procs.2019.11.175

Hernández, R., & Mendoza, C. (2018). Metodología de la investigación. In *Mc Graw Hill* (Vol. 1, Issue Mexico).

Hswen, Y., Qin, Q., Williams, D. R., Viswanath, K., Subramanian, S. V., & Brownstein, J. S. (2020). Online negative sentiment towards Mexicans and Hispanics and impact on mental well-being: A time-series analysis of social media data during the 2016 United States presidential election. *Heliyon*, *6*(9), e04910. https://doi.org/10.1016/j.heliyon.2020.e04910

IEP. (2021). *IEP Informe de Opinión –Abril 2021 Intención de voto- Elecciones Generales 2021*. *1215*. Recuperado de: https://iep.org.pe/wp-content/uploads/2021/04/Informe-IEP-OP-abril-I-2021.pdf

Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2019). Predicting elections from social media: a three-country, three-method comparative study. *Asian Journal of Communication*, *29*(3), 252–273. https://doi.org/10.1080/01292986.2018.1453849

Joseph, F. J. J. (2019). Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree. *Proceedings of 2019 4th International Conference on Information Technology: Encompassing Intelligent Technology and Innovation Towards the New Era of Human Life, InCIT 2019*, 50–53. https://doi.org/10.1109/INCIT.2019.8911975

Kemp, S. (2021a). *Digital 2021: Peru*. DataReportal. Recuperado de: https://datareportal.com/reports/digital-2021-global-overview-report

Kemp, S. (2021b). *Global Social Media Stats*. DataReportal. Recuperado de: https://datareportal.com/social-media-users

Khatua, A., Khatua, A., & Cambria, E. (2020). Predicting political sentiments of voters from Twitter in multi-party contexts. *Applied Soft Computing Journal*, *97*, 106743. https://doi.org/10.1016/j.asoc.2020.106743

Kušen, E., & Strembeck, M. (2018). Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections. *Online Social Networks and Media*, *5*, 37–50.

https://doi.org/10.1016/j.osnem.2017.12.002

Ligthart, A., Catal, C., & Tekinerdogan, B. (2021). Systematic reviews in sentiment analysis: a tertiary study. In *Artificial Intelligence Review* (Vol. 54, Issue 7). Springer Netherlands. https://doi.org/10.1007/s10462-021-09973-3

Liu, B. (2011). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. In *Global Journal of Pure and Applied Mathematics* (Vol. 11, Issue 5). Springer.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining Text Data* (Vol. 9781461432, pp. 415–463). Springer Science & Bussiness Media. https://doi.org/10.1007/978-1-4614-3223-4

Liu, D., & Lei, L. (2018). The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election. *Discourse, Context and Media*, *25*, 143–152. https://doi.org/10.1016/j.dcm.2018.05.001

Long, C. (2015). Advanced analytical theory and methods: Time Series Analysis. In *Data Science and Big Data Analytics Discovering, analyzing, visualizing and presenting data*.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., & McCullough, D. (2012). On building a reusable Twitter corpus. *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1113–1114. https://doi.org/10.1145/2348283.2348495

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mee, A., Homapour, E., Chiclana, F., & Engel, O. (2021). Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit[Formula presented]. *Knowledge-Based Systems*, *228*, 107238. https://doi.org/10.1016/j.knosys.2021.107238

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *Fifth International AAAI Conference on Weblogs and Social Media*. https://doi.org/https://doi.org/10.1609/icwsm.v5i1.14168

Mohamed, E. H., Moussa, M. E. S., & Haggag, M. H. (2020). An Enhanced Sentiment

Analysis Framework Based on Pre-Trained Word Embedding. *International Journal of Computational Intelligence and Applications*, *19*(4). https://doi.org/10.1142/S1469026820500315

ONPE. (2021). *Presentación de resultados. Elecciones Generales y Parlamento Andino 2021*. Recuperado de: https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T

Paliwal, S., Khatri, S. K., & Sharma, M. (2018). Sentiment Analysis and Prediction Using Neural Networks. *International Conference on Advanced Informatics for Computing Research*, *1*(July), 1035–1042. https://doi.org/10.1007/978-981-13-3140-4

Paul, J., Parameswar, N., Sindhani, M., & Dhir, S. (2021). Use of microblogging platform for digital communication in politics. *Journal of Business Research*, *127*(February), 322–331. https://doi.org/10.1016/j.jbusres.2021.01.046

Peng, M., Zhang, Q., Jiang, Y. G., & Huang, X. (2018). Cross-domain sentiment classification with target domain specific information. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 2505–2513. https://doi.org/10.18653/v1/p18-1233

Pérez, J. M., Giudici, J. C., & Luque, F. (2021). *pysentimiento: A Python Toolkit for Sentiment Analysis and SocialNLP tasks*. 1–4. http://arxiv.org/abs/2106.09462

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). *Sentiment Analysis in Social Networks* (M. Kaufmann (ed.)). Todd Green.

Qi, L., Li, R., Wong, J., Tavanapong, W., & Peterson, D. A. M. (2017). Social media in state politics: Mining policy agendas topics. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, 274–277. https://doi.org/10.1145/3110025.3110097

Rodriguez-Ibanez, M., Gimeno-Blanes, F. J., Cuenca-Jimenez, P. M., Munoz-Romero, S., Soguero, C., & Rojo-Alvarez, J. L. (2020). On the Statistical and Temporal Dynamics of Sentiment Analysis. *IEEE Access*, *8*, 87994–88013. https://doi.org/10.1109/ACCESS.2020.2987207

Rodriguez-Ibanez, M., Gimeno-Blanes, F. J., Cuenca-Jimenez, P. M., Soguero-Ruiz, C., & Rojo-Alvarez, J. L. (2021). Sentiment Analysis of Political Tweets from the 2019 Spanish Elections. *IEEE Access*, *9*, 101847–101862.

https://doi.org/10.1109/ACCESS.2021.3097492

Saleiro, P., Gomes, L., & Soares, C. (2018). Sentiment aggregate functions for political opinion polling using microblog streams. *ACM International Conference Proceeding Series*, 44–50. https://doi.org/10.1145/2948992.2949022

Seckin, T., & Kilimci, Z. H. (2020). The Evaluation of 5G technology from Sentiment Analysis Perspective in Twitter. *Proceedings - 2020 Innovations in Intelligent Systems and Applications Conference, ASYU 2020*. https://doi.org/10.1109/ASYU50717.2020.9259900

Sobrino, J. C. (2018). *Análisis de sentimientos en Twitter* [Universitat Oberta de Catalunya]. Recuperado de: http://hdl.handle.net/10609/81435

Sudhir, P., & Suresh, V. D. (2021). Comparative study of various approaches, applications and classifiers for sentiment analysis. *Global Transitions Proceedings*, *2*(2), 205–211. https://doi.org/10.1016/j.gltp.2021.08.004

Ullah, M. A., Hasnayeen, M. A., Shan-A-Alahi, A., Rahman, F., & Akhter, S. (2020). A Search for Optimal Feature in Political Sentiment Analysis. *Proceedings of 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, WIECON-ECE 2020*, 340–343. https://doi.org/10.1109/WIECON-ECE52138.2020.9397966

Zhai, C., & Massung, S. (2016). Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. In *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. https://doi.org/10.1145/2915031

Zorrilla, D. (2022). *5 datos para comprender el impacto de Twitter en las comunicaciones digitales*. Quantico. Recuperado de: https://www.quanticotrends.com/estudios/estudio-de-uso-de-twitter-en-el-peru/

ANEXOS

Anexo 1. Scripts para la limpieza y normalización de datos

```
# Read JSON file containing tweets data and remove tweets not in Spanish
raw_tweets = pd.read_json(r'/content/EleccionesBicentenario2021.json', lines=True)
raw_tweets = raw_tweets[raw_tweets['lang']=='es']
print("Shape: ", raw_tweets.shape)
raw_tweets.head(5)

# Normalize 'user' field on tweets
users = json_normalize(raw_tweets['user'])
users.drop(['_type', 'verified', 'renderedDescription', 'descriptionLinks', 'profileImag
eUrl', 'profileBannerUrl', 'label', 'label._type', 'label.description', 'label.url', 'la
bel.badgeUrl', 'label.longDescription', 'link._type', 'link.text', 'link.tcourl', 'link.
indices','protected', 'link', 'link.url'], axis=1, inplace=True)
users.rename(columns={'id':'userId', 'url':'profileUrl', 'location':'userLocation'}, inp
lace=True)
users.head(5)

# Create DataFrame for Users and remove duplicates
users = pd.DataFrame(users)
users.drop_duplicates(subset=['userId'], inplace=True)
print("Shape: ", users.shape)
users.head(5)

# Export dataframe Users into a CSV
users.to_csv('EleccionesPeru_UsersDataFrame2021.csv', sep=',', index=False)

# Transform 'raw_tweets' to DataFrame Tweets
# Add column for userId, username, location and creation date
user_id = []
user_name = []
user_screen_name = []
user_description = []
user_location = []
for user in raw_tweets['user']:
  uid = user['id']
  uname = user['username']
  uscreenname = user['displayname']
  udescription = user['rawDescription']
  ulocation = user['location']

  user_id.append(uid)
  user_name.append(uname)
  user_screen_name.append(uscreenname)
  user_description.append(udescription)
  user_location.append(ulocation)
raw_tweets['userId'] = user_id
raw_tweets['userName'] = user_name
raw_tweets['userScreenName'] = user_screen_name
```

```
raw_tweets['userDescription'] = user_description
raw_tweets['userLocation'] = user_location

# Delete some columns and keep only the most important
cols = ['date', 'id', 'renderedContent', 'retweetCount', 'likeCount', 'replyCount', 'use
rId', 'userName', 'userScreenName', 'userDescription', 'userLocation', 'hashtags']
tweets = raw_tweets[cols]
tweets.rename(columns = {'id':'tweetId', 'date':'createdAt'}, inplace = True)
print("Shape: ", tweets.shape)
tweets.head(5)

# Convert to DataFrame, remove duplicates and keep only Spanish tweets
tweets = pd.DataFrame(tweets)
tweets.drop_duplicates(subset=['tweetId'], inplace=True)
print("Shape: ", tweets.shape)
tweets.head(5)

# Remove duplicates from rederedContent column in tweets DataFrame
tweets.drop_duplicates(subset=['renderedContent'], inplace=True)
print("Shape: ", tweets.shape)
tweets.head(5)

# Delete tweets where location is different from Peru, because there were others
elections processes around the world:
Ecuador, Bolivia, Argentina y Perú El Salvador, Honduras
tweets.drop(tweets.index[(tweets["userLocation"] == "México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ciudad de México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Puebla, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"]=="San Luis Potosí, México")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Puebla")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Sinaloa, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mexico")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mérida, Yucatán")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Querétaro")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Culiacán, Sinaloa")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Monterrey, Nuevo León")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guayaquil - Ecuador")],inplace=True
)
tweets.drop(tweets.index[(tweets["userLocation"] == "Long Beach, CA")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Hidalgo, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guadalajara, Jalisco")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "La Habana, Cuba")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "WTC Cd. de México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Puebla Mexico")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Aguascalientes, Ags.")],inplace=Tru
e)
```

```
tweets.drop(tweets.index[(tweets["userLocation"] == "Ciudad de México, MX")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Misiones, Argentina")],inplace=True
)
tweets.drop(tweets.index[(tweets["userLocation"] == "Camargo, Chihuahua")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Toluca")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "México, D.F.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Celaya, Gto.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Bolivia")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Baja California")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Groningen, Nederland")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ciudad Juárez")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Colima")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Queretaro, Mexico")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Michoacan")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cuenca, Ecuador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "España")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Querétaro, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tlaxcala")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "León, Guanajuato")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Torreón")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Puyo Pastaza, Ecuador")],inplace=Tr
ue)tweets.drop(tweets.index[(tweets["userLocation"]=="Mérida, Yucatán, México")],inplace
=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "León, Guanajuato.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"]=="México, San Luis Potosí")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Victoria, Tamaulipas")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Marbella")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "León, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Huauchinango, Puebla")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mazatlán, Sinaloa")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Xalapa, Veracruz")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] =="Monterrey, N.L. México")],inplace=Tr
ue)tweets.drop(tweets.index[(tweets["userLocation"] == "Oaxaca, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guayaquil, Ecuador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Quito")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Quito, Ecuador EC")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Colchane, Chile")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Hidalgo, México.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mazātlān")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Toluca, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] =="Aguascalientes, México")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Morelia, Michoacán")],inplace=True)
```

```
tweets.drop(tweets.index[(tweets["userLocation"] == "Irapuato, Guanajuato")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tecomán Colima. Mx")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tlaxcala, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Playas de Rosarito")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Veracruz, Vearcruz")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ensenada, Mx")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "santiago de Chile")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Pichincha, Ecuador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Villahermosa, Tabasco")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tehuacán")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "San Luis Potosi")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Quito Ecuador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Morelos")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Coyoacán-
Querétaro.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Atlixco, Puebla")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Quito-Ecuador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Chihuahua")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ibarra, Ecuador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guadalajara, Jal.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "CDMX, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "CDMX")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Xalapa, Ver.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Zacatecas, Mexico")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tepic Nayarit")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Asunción, Paraguay")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guanajuato, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ciudad De México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Salamanca, Guanajuato")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mérida Yucatán")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "San Luis Potosí")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Durango, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "México, DF")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Zacatecas, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Chiapas, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Hermosillo, Sonora")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mexico ")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guanajuato, Gto.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Quito, Ecuador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Chihuahua, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Oaxaca")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Quintana Roo, México")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Sonora")],inplace=True)
```

```
tweets.drop(tweets.index[(tweets["userLocation"] == "Tegucigalpa, Honduras")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Morelia, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tlaxcala ")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Veracruz ")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "El Salvador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "San Salvador")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] =="Puebla, Puebla, México")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Colima, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tepic, Nayarit")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "CDMX ")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "EL SALVADOR")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Chiapas, Mexico")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Puebla, Pue")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Puebla ")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "México ")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Morelos, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cuernavaca, Morelos")],inplace=True
)
tweets.drop(tweets.index[(tweets["userLocation"] == "Xalapa")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Veracruz")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Jalisco, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Quintana Roo")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cuernavaca Morelos")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tamaulipas, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Coahuila")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tlaquepaque, Jalisco")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cd. Juarez / Chih")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "México DF")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] =="Santa Ana, El Salvador")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Veracruz MX")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cuautla, Morelos")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Veracruz, MX")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Poblano")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Latinoamérica")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Washington, DC")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Madrid")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Campeche, México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Colima, Colima")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cancún")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "TEPIC NAYARIT")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Las Condes, Chile")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Estado de México")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Zacatecas, Zacatecas.")],inplace=Tr
ue)
tweets.drop(tweets.index[(tweets["userLocation"] == "Dallas Texas USA ")],inplace=True)
```

```
tweets.drop(tweets.index[(tweets["userLocation"] == "Monterrey")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "La Plata, Argentina")],inplace=True
)
tweets.drop(tweets.index[(tweets["userLocation"] == "Centroamérica")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guerrero")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Viña del Mar, Chile")],inplace=True
)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guayaquil")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ciudad Imagen")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Los Angeles, CA")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "La Serena, Chile")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cancún, Quintana Roo")],inplace=Tru
e)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mexico City")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Pachuca, Hidalgo")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cerrillos")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Concepción, Chile")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guayaquil, EC")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "México D.F.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "El mundo")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ciudad de Mexico")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Salta, Argentina")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Santiago, chile")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Buenos Aires")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Miami, FL")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Valparaíso, Chile")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "MEXICO ")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "MEXICO")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Durango Mex.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Cuenca ECU.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Querétaro.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Mexico, ME")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Ecuador- Uio")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Bijnor, India")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Puebla, Pue.")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Tijuana")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Montevideo, Uruguay")],inplace=True
)
tweets.drop(tweets.index[(tweets["userLocation"] == "Nicaragua")],inplace=True)
tweets.drop(tweets.index[(tweets["userLocation"] == "Guayas, Ecuador")],inplace=True)
# Create some columns to save data from other location
def location(df) :
    if (df["userLocation"] == "Lima" or df["userLocation "] == "Lima-
Perú" or df["userLocation "] == "Lima Perú" or df["userLocation "] == "Lima - Perú. "):
        return "Lima"
    elif (df["userLocation "] =="Chile" or df["userLocation
"] == "Santiago, Chile" or df["userLocation "] == "Santiago de Chile"):
        return "Chile"
    else :
```

```
        return df["userLocation "]
tweets["location2"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location2"] == "Lima" or df["location2"] == "Lima, Perú." or df["location2"]
 == "Lima / CDMX / Toronto" or df["location2"] == "Valle de Lurín, Lima - Perú "):
        return "Lima"
    elif (df["location2"] =="Argentina" or df["location2"] == "Buenos Aires, Argentina")
:
        return "Argentina"
    else :
        return df["location2"]
tweets["location3"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location3"] == "Lima" or df["location3"] == "Lima - Peru"  or df["location3"
] == "Lima " or df["location3"] == "Lima Peru" or df["location3"] == "Santiago de Surco,
 Peru"):
        return "Lima"
    elif (df["location3"] =="Arequipa, Peru" or df["location3"] == "Arequipa, Perú" or d
f["location3"] == "Arequipa - Perú"):
        return "Arequipa"
    else :
        return df["location3"]
tweets["location4"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location4"] == "Lima" or df["location4"] == "Lima e ICA"  or df["location4"]
 == "Lima, Perú " or df["location4"] == "Lima,Perú" or df["location4"] == "lima"):
        return "Lima"
    elif (df["location4"] =="Perú" or df["location4"] == "Worldwide - Peru" or df["locat
ion4"] == "peru"):
        return "Perú"
    else :
        return df["location4"]
tweets["location5"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location5"] == "Lima" or df["location5"] == "Lima, Perú PE"  or df["location5
"] == "Lima, Perú PE " or df["location5"] == "La Molina, Peru" or df["location5"] == "Lim
a-Peru" or df["location5"] == "Miraflores, Peru"):
        return "Lima"
    elif (df["location5"] =="Ecuador" or df["location5"] == "Chile" or df["location5"] =
= "Argentina" or df["location5"] == "Latinoamérica" or df["location5"] =="Venezuela"):
        return "Extranjero"
    else :
        return df["location5"]
tweets["location6"] = tweets.apply(lambda tweets:location(tweets),axis = 1)
```

```
def location(df) :
    if (df["location6"] == "Lima" or df["location6"] == "Miraflores, Lima, Perú"  or df[
"location6"] == "Las faldas de un cerro de Lima" or df["location6"] == "La Victoria, Lim
a, Perú" or df["location6"] == "Lima 1" or df["location6"] == "San isidro, Peru"):
        return "Lima"
    elif (df["location6"] =="Extranjero" or df["location6"] == "Washington, DC" or df["l
ocation6"] == "Mount Greenwood Chicago , US" or df["location6"] == "Madrid" or df["locat
ion6"] =="Global"):
        return "Extranjero"
    else :
        return df["location6"]
tweets["location7"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location7"] == "Lima" or df["location7"] == "Jesús Maria, Peru" or df["locat
ion7"] == "Las Palomas 430,Surquillo,Lima"):
        return "Lima"
    elif (df["location7"] =="Extranjero" or df["location7"] == " Sverige"  or df["userLo
cation"]=="Los Angeles, CA"):
        return "Extranjero"
    else :
        return df["location7"]
tweets["location8"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location8"] == "Lima" or df["location8"] == "Llima peru" or df["location8"]
== "CDMX COLONIA NAPOLES- LIMA"):
        return "Lima"
    elif (df["location8"] =="Trujillo, Peru" or df["location8"] == "Trujillo"):
        return "Trujillo"
    else :
        return df["location8"]
tweets["location9"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location9"] == "Lima" or df["location9"] == " Lima Peru" or df["location9"]
== "San Isidro, Lima, Perú" or df["location9"] == "LIMA-
PERU" or df["location9"] == "Lima - Los Angeles"):
        return "Lima"
    elif (df["location9"] =="Perú" or df["location9"] == "Tu Corazon" or df["location9"]
 == "Perú " or df["location9"] =="IG @viviperu"):
        return "Perú"
    else :
        return df["location9"]
tweets["location10"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location10"] == "Lima" or df["location10"] == "Lima - Perú "):
        return "Lima"
```

```
    elif (df["location10"] == "Extranjero" or df["location10"] == "Seoul, South Korea "
or df["location10"] == "Nueva York, USA"):
        return "Extranjero"
    else :
        return df["location10"]
tweets["location11"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location11"] == "Lima" or df["location11"] == "LIMA" or df["location11"] ==
"San Juan de Miraflores"):
        return "Lima"
    elif (df["location11"] == "Perú" or df["location11"] == "Getting out of here"):
        return "Perú"
    elif (df["location11"] == "Extranjero" or df["location11"] == "Quito - Ecuador" or d
f["location11"] == "Ciudad Autónoma de Buenos Aire" or df["location11"] == "Antofagasta,
 Chile" or df["location11"] == "Texas, USA" or df["location11"] == "Universidad Nacional
 Autónoma " or df["location11"] == "Miami, FL"):
        return "Extranjero"
    else :
        return df["location11"]
tweets["location12"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location12"] == "Lima" or df["location12"] == "Surco, Peru" or df["location1
2"] == "Carabayllo, Peru" or df["location12"] == "lima, peru" or df["location12"] == "Li
ma- Perú" or df["location12"] == "LIMA - PERU"):
        return "Lima"
    elif (df["location12"] == "Perú" or df["location12"] == "Contigo este Verano."):
        return "Perú"
    elif (df["location12"] == "Trujillo" or df["location12"] == "Trujillo - Perú"):
        return "Trujillo"
    elif (df["location12"] == "Extranjero" or df["location12"] == "Montevideo, Uruguay"
or df["location12"] == "América Latina y el Caribe"):
        return "Extranjero"
    else :
        return df["location12"]
tweets["location13"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

def location(df) :
    if (df["location13"] == "Lima" or df["location13"] == "Barranco, Peru" or df["locati
on13"] == "Lima - Madrid ✈"):
        return "Lima"
    elif (df["location13"] == "Chiclayo, Peru"):
        return "Chiclayo"
    elif (df["location13"] == "Callao, Peru" or df["location13"] == "Callao"):
        return "Callao"
    elif (df["location13"] == "Pasco, Peru"):
        return "Pasco"
    elif (df["location13"] == "Tacna, Peru" or df["location13"] == "Tacna Perú"):
```

```
        return "Tacna"
    elif (df["location13"] == "Cusco, Peru"):
        return "Cusco"
    elif (df["location13"] == "Huaraz, Peru"):
        return "Huaraz"
    elif (df["location13"] == "Piura, Peru"):
        return "Piura"
    elif (df["location13"] == "Puno, Peru"):
        return "Puno"
    elif (df["location13"] == "Huanuco, Peru"):
        return "Huánuco"
    elif (df["location13"] == "Cajamarca, Peru"):
        return "Cajamarca"
    elif (df["location13"] == "Huancayo, Peru"):
        return "Huancayo"
    elif (df["location13"] == "Ica, Peru"):
        return "Ica"
    else :
        return df["location13"]
tweets["location14"] = tweets.apply(lambda tweets:location(tweets),axis = 1)

# Remove columns created to filter location
tweets.drop(['location', 'location2', 'location3', 'location4', 'location5', 'location6'
, 'location7', 'location8', 'location9', 'location10', 'location11', 'location12', 'loca
tion13'], axis=1, inplace = True)
df.drop(['userLocation'], axis=1, inplace = True)
# Rename column created to filter location
df.rename(columns = {'location14':'tweetLocation'}, inplace = True)
# Complete rows with null values
df.tweetLocation.fillna("Desconocida",inplace = True)

def clean_tweets(text):
    # All text to lowercase
    new_tweet = text.replace('\r', '').replace('\n', ' ').replace('\n', ' ').lower()
    # Remove links and common URLs
    new_tweet = re.sub(r"(?:\@|https?\://|bit.ly\/|buff.ly\/|ow.ly\/|cutt.ly\/|shorturl.
at\/|youtu.be\/|youtube.com\/|larepublica.pe\/|elcomercio.pe\/|dlvr.it\/|twitter.com\/|f
acebook.com\/|instagram.com\/|ptv.pe\/|tinyurl.com\/)\S+", "", new_tweet)
    # Remove punctuation, add Spanish symbols
    regex = '[\\¡\\!\\"\\#\\$\\%\\&\\\'\\(\\)\\*\\+\\,\\-
\\.\\/\\:\\;\\<\\=\\>\\¿\\?\\@\\[\\\\\\]\\^_\\`\\{\\|\\}\\~]'
    new_tweet = re.sub(regex , '', new_tweet)
    # Remove empty spaces
    new_tweet = re.sub("\\s+", ' ', new_tweet)
    # Tokenization and remove RT
    new_tweet = new_tweet.split(sep = ' ')
    new_tweet = [word for word in new_tweet if (word != 'rt')]
        return(new_tweet)
```

Anexo 2. Scripts para la eliminación de palabras vacías, tokenización y etiquetado de *tweets*

```
# remove stopwords
def remove_stopwords(text, is_lower_case=False):
    tokens = word_tokenize(text)
    tokens = [token.strip() for token in tokens]
    if is_lower_case:
        filtered_tokens = [token for token in tokens if (token not in stopword_list and
len(token)>2)] #remove stopwords and words <2
    else:
        filtered_tokens = [token for token in tokens if (token.lower() not in stopword_l
ist and len(token)>2)]
    filtered_text = ' '.join(filtered_tokens)
    return filtered_text

# Tagging task for tweets
pip install pysentimiento

from pysentimiento import create_analyzer
analyzer = create_analyzer(task="sentiment", lang="es")

def sent_analyzer(row):
    res = analyzer.predict(row["cleanContent"])
    return pd.Series({'polarity': res.output, **res.probas})

df1 = df1.join(df1.apply(sent_analyzer, axis=1))
df2 = df2.join(df2.apply(sent_analyzer, axis=1))

# Merge of two dataframes
frames = [df1, df2]
df = pd.concat(frames)

# function to decode polarity in tweets
def polarity_rating (score):
    if score == 'POS':
        return 1
    elif score == 'NEG':
        return -1
    else:
      return 0
```

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

DECLARACIÓN JURADA DE AUTENTICIDAD DE TESIS

Por el presente documento, Yo ALODIA FLORES ARNAO,
identificado con DNI 43201154 en mi condición de egresado de:

☐ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☒ **Programa de Maestría o Doctorado**

MAESTRÍA EN INGENIERÍA DE SISTEMAS,

informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:
" INTENCIÓN DE VOTO A TRAVÉS DE UN MODELO DE ANÁLISIS DE SENTIMIENTOS EN TWITTER BASADO EN TÉCNICAS DE MACHINE LEARNING
"

Es un tema original.

Declaro que el presente trabajo de tesis es elaborado por mi persona y **no existe plagio/copia** de ninguna naturaleza, en especial de otro documento de investigación (tesis, revista, texto, congreso, o similar) presentado por persona natural o jurídica alguna ante instituciones académicas, profesionales, de investigación o similares, en el país o en el extranjero.

Dejo constancia que las citas de otros autores han sido debidamente identificadas en el trabajo de investigación, por lo que no asumiré como suyas las opiniones vertidas por terceros, ya sea de fuentes encontradas en medios escritos, digitales o Internet.

Asimismo, ratifico que soy plenamente consciente de todo el contenido de la tesis y asumo la responsabilidad de cualquier error u omisión en el documento, así como de las connotaciones éticas y legales involucradas.

En caso de incumplimiento de esta declaración, me someto a las disposiciones legales vigentes y a las sanciones correspondientes de igual forma me someto a las sanciones establecidas en las Directivas y otras normas internas, así como las que me alcancen del Código Civil y Normas Legales conexas por el incumplimiento del presente compromiso

Puno 20 de diciembre del 2023

FIRMA (obligatoria)

Huella

Universidad Nacional del Altiplano Puno | Vicerrectorado de Investigación | Repositorio Institucional

AUTORIZACIÓN PARA EL DEPÓSITO DE TESIS O TRABAJO DE INVESTIGACIÓN EN EL REPOSITORIO INSTITUCIONAL

Por el presente documento, Yo ALODIA FLORES ARNAO,
identificado con DNI 43201154 en mi condición de egresado de:

☐ **Escuela Profesional,** ☐ **Programa de Segunda Especialidad,** ☒ **Programa de Maestría o Doctorado**

MAESTRÍA EN INGENIERÍA DE SISTEMAS,
informo que he elaborado el/la ☒ **Tesis** o ☐ **Trabajo de Investigación** denominada:

" INTENCIÓN DE VOTO A TRAVÉS DE UN MODELO DE ANÁLISIS DE SENTIMIENTOS EN TWITTER BASADO EN TÉCNICAS DE MACHINE LEARNING "

para la obtención de ☒**Grado,** ☐ **Título Profesional** o ☐ **Segunda Especialidad.**

Por medio del presente documento, afirmo y garantizo ser el legítimo, único y exclusivo titular de todos los derechos de propiedad intelectual sobre los documentos arriba mencionados, las obras, los contenidos, los productos y/o las creaciones en general (en adelante, los "Contenidos") que serán incluidos en el repositorio institucional de la Universidad Nacional del Altiplano de Puno.

También, doy seguridad de que los contenidos entregados se encuentran libres de toda contraseña, restricción o medida tecnológica de protección, con la finalidad de permitir que se puedan leer, descargar, reproducir, distribuir, imprimir, buscar y enlazar los textos completos, sin limitación alguna.

Autorizo a la Universidad Nacional del Altiplano de Puno a publicar los Contenidos en el Repositorio Institucional y, en consecuencia, en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, sobre la base de lo establecido en la Ley N° 30035, sus normas reglamentarias, modificatorias, sustitutorias y conexas, y de acuerdo con las políticas de acceso abierto que la Universidad aplique en relación con sus Repositorios Institucionales. Autorizo expresamente toda consulta y uso de los Contenidos, por parte de cualquier persona, por el tiempo de duración de los derechos patrimoniales de autor y derechos conexos, a título gratuito y a nivel mundial.

En consecuencia, la Universidad tendrá la posibilidad de divulgar y difundir los Contenidos, de manera total o parcial, sin limitación alguna y sin derecho a pago de contraprestación, remuneración ni regalía alguna a favor mío; en los medios, canales y plataformas que la Universidad y/o el Estado de la República del Perú determinen, a nivel mundial, sin restricción geográfica alguna y de manera indefinida, pudiendo crear y/o extraer los metadatos sobre los Contenidos, e incluir los Contenidos en los índices y buscadores que estimen necesarios para promover su difusión.

Autorizo que los Contenidos sean puestos a disposición del público a través de la siguiente licencia:

Creative Commons Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional. Para ver una copia de esta licencia, visita: https://creativecommons.org/licenses/by-nc-sa/4.0/

En señal de conformidad, suscribo el presente documento.

Puno 20 de diciembre del 2023

FIRMA (obligatoria)

Huella