Revista de investigación de sistemas e informática
16(1): 5-20 (2023)
DOI: https://doi.org/10.15381/risi.v16i1.25061

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Inteligencia artificial aplicado en la detección del discurso de odio en el contexto político social a principios del año 2023 en el Perú

Artificial intelligence applied in the detection of hate speech in the social political context in early 2023 in Peru

Luz Elena Torres Talaverano
https://orcid.org/0000-0001-6465-0430
luzelena.torres@unmsm.edu.pe

Paula Elianne Rojas Villanueva
https://orcid.org/0009-0006-8554-5406
paulaelianne.rojas@unmsm.edu.pe

Lucero Marysol Huamán Ampuero
https://orcid.org/0009-0005-6092-1607
lucero.huaman@unmsm.edu.pe

Ciro Rodriguez Rodriguez
https://orcid.org/0000-0003-2112-1349
crodriguezro@unmsm.edu.pe

Ivan Carlo Petrlik Azabache
https://orcid.org/0000-0002-1201-2143
ipetrlika@unmsm.edu.pe

Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática.
Lima, Perú

Recibido: 24/02/2023 - Aceptado: 26/04/2023 - Publicado: 21/08/2023

RESUMEN

La rápida propagación del discurso de odio a través de las redes sociales hacia un gran número de personas ha generado desestabilización en un contexto social y económico al inicio del año 2023. La finalidad de la investigación es detectar el discurso de odio a través de la Inteligencia Artificial con técnicas de Machine Learning en un contexto político social. Además, la investigación es de tipo aplicada en la que se llevó a cabo una búsqueda exploratoria y la población está comprendida por 1.970 tweets en las regiones del centro y sur del Perú. El muestreo es probabilístico, y se revisaron y etiquetaron manualmente 1.970 tweets de las regiones involucradas. Asimismo, la metodología aplicada fue SEMMA, ya que permite asegurar que los modelos sean precisos y se adapten a los datos de este estudio. Se realizó una comparación de los respectivos modelos para lograr un resultado óptimo, de las cuales el indicador principal fue la exactitud promedio en la que el modelo Multinomial Naive Bayes fue el mejor modelo con una puntuación del 73.87%. Finalmente, se concluye que este estudio permite acelerar el proceso de detección del discurso de odio verbal en redes sociales para el análisis de la opinión pública y la detección de presuntos incitadores de odio.

Palabras clave: Inteligencia artificial, Discurso de odio, Aprendizaje automático, Twitter.

ABSTRACT

The rapid spread of hate speech through social networks to a large number of people has generated destabilization in a social and economic context at the beginning of the year 2023. The purpose of the research is to detect hate speech through Artificial Intelligence with Machine Learning techniques in a social political context. In addition, the research is of an applied type in which an exploratory search was carried out and the population is comprised of 1,970 tweets in the central and southern regions of Peru. The sampling is probabilistic, and 1,970 tweets from the regions involved were reviewed and manually labeled. Likewise, the

methodology applied was SEMMA, as it allows to ensure that the models are accurate and fit the data of this study. A comparison of the respective models was performed to achieve an optimal result, of which the main indicator was the average accuracy in which the Multinomial Naive Bayes model was the best model with a score of 73.87%. Finally, it is concluded that this study allows accelerating the process of detecting verbal hate speech in social networks for the analysis of public opinion and the detection of suspected hate inciters.

Keywords: Artificial intelligence, Hate speech, Machine learning, Twitter.

I. INTRODUCCIÓN

Las redes sociales son una plataforma de comunicación más común y poderosa para compartir puntos de vista sobre cualquier tema o artículo, lo que en consecuencia conduce a conversaciones no estructuradas, tóxicas y llenas de odio (Chhabra, A., & Vishwakarma, D. K., 2023), que en otra palabras se traduce en un discurso de odio que según (Nockleby, 2000, como se citó en Parvaresh, V., 2023), lo considera a cualquier forma de comunicación que sirva para denigrar a individuos o grupos en base a una o más de sus características, incluyendo "raza, color, etnia, nacionalidad, religión", se ha visto como un problema dentro de las sociedades democráticas, exacerbando las redes sociales (Lilleker, D., & Pérez-Escolar, M, 2023). La incidencia del discurso de odio ha seguido aumentando, hasta el punto de que ahora ha alcanzado niveles extremadamente altos (Baider, 2020; Baider et al., 2017). En Perú, actualmente, hay una inestabilidad política de la cual hay grandes cambios producto del ascenso de líderes populistas y la aparición de movimientos ideológicos extremos. Según Romero-Rodríguez, L. M., Castillo-Abdul, B., & Cuesta-Valiño, P. (2023), el surgimiento de este fenómeno se debe, en gran medida, a la facilidad con la que estos actores políticos pueden difundir sus mensajes sin límites a través de las redes sociales, dejando de lado al antiguo "cuarto poder" de los medios como filtros y reinterpretadores de información. Generalmente, la fórmula utilizada por estos líderes y movimientos suele basarse en la división y polarización social simbólica a través de discursos de odio que permiten satanizar a sus adversarios mientras antagonizan a los emisores: un "nosotros" discursivo contra "ellos" basado en la violencia verbal para deshumanizar a un "exogrupo". A principios del año 2023, una gran cantidad de discursos de odio se manifestaron a través de las redes sociales en las regiones del centro y sur del Perú (Cusco, Arequipa, Puno y otros) con la finalidad de desestabilizar el régimen actual e incitar las protestas no pacíficas, que fueron catalogadas como la toma de Lima 2023, las cuales afectaron la economía del país.

Según Rodriguez, M. (2023), el INEI publicó un informe indicando que la producción nacional de enero de este año disminuyó un 1,12%, debido a conflictos sociales que interrumpieron las actividades económicas, como la paralización de labores, el bloqueo de carreteras, el cierre forzado de mercados y la restricción del libre tránsito de personas y mercancías. Sin embargo, las protestas se redujeron gradualmente desde mediados de febrero en donde se registró una mejora en el PBI a comparación de inicios de año.

La presente investigación se ha sustentado a través de una serie de antecedentes tanto nacionales como internacionales, los cuales se van a presentar de la siguiente manera:

Según Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., & Malik, S. H. (2022), la pandemia de COVID-19 ha afectado a todas las naciones, y el aislamiento social es el principal método de protección contra el coronavirus. La gente expresa a través de Facebook y Twitter desinformación y discursos de odio. Esta investigación pretende detectar el discurso del odio utilizando técnicas de aprendizaje automático y aprendizaje por conjuntos durante COVID-19. Los datos de Twitter se extrajeron utilizando su API con la ayuda de los hashtags que fueron tendencia durante la pandemia COVID-19. Los tweets se clasificaron manualmente en dos categorías en función de distintos factores. Las características se extrajeron utilizando TF/IDF, Bag of Words y Tweet Length. El estudio demostró la eficacia del clasificador de árbol de decisión. Comparado con otros clasificadores típicos de ML, tiene una precisión del 98%, una recuperación del 97%, una puntuación F1 del 97% y una exactitud del 97%. El clasificador Stochastic Gradient Boosting supera a todos los demás con un 99% de precisión, 97% de recuperación, 98% de puntuación F1 y 98,04% de exactitud.

Para Toliyat, A., Levitan, S. I., Peng, Z., & Etemadpour, R. (2022), la pandemia por la COVID-19 provocó que los gobiernos de todo el mundo tomarán medidas severas para reducir la propagación de la

enfermedad. Estos sucesos conllevaron a que se genere racismo y odio hacia los asiáticos en algunos lugares, e inclusive crímenes de odio contra ellos. Por ello, este estudio tuvo como objetivo explorar la expresión de los usuarios en Twitter para presentar un enfoque de mejora en la clasificación de tweets de odio contra los asiáticos. Emplearon 10 millones de tweets a través de la API de Twitter V-2, seleccionaron una muestra de 3000 mil tweets para anotar. También, utilizaron métodos de aprendizaje automático como Random Forest, KNN, Máquina de vectores de soporte (SVM), Extreme Gradiente Boosting (XGBoost), Naive Bayes, Regresión Logística y modelos de aprendizaje profundo como LSTM, LSTM bidireccional, LSTM bidireccional con eliminación, BERT; todo ello con el fin de construir modelos predictivos de discurso de odio. El resultado que obtuvieron fueron que Regresión Logística (puntuación F1 de 0.72) y BERT (puntuación F1 de 0.85) lograron un mejor rendimiento estadístico en los modelos de aprendizaje automático y aprendizaje profundo, respectivamente.

Según, Amores, J., Blanco-Herrero, D., Sánchez-Holgado, P., Frías-Vázquez, M. (2021) el discurso de odio propagado a través de redes sociales como Twitter merece atención especial, ya que su incremento puede relacionarse con el aumento de crímenes de odio. De las 11 categorías de discriminación que contempla el Ministerio de Interior de España, la segunda en la que más delitos de odio se registran al año es la ideología. Sin embargo, esta categoría queda fuera de la mayor parte de los planes de acción para estudiar y combatir los delitos de odio. Lo mismo ocurre con los trabajos académicos, que se centran mayoritariamente en el odio en inglés y a nivel general. Los que estudian un único tipo de odio se han enfocado en el racismo, la xenofobia o la discriminación de género, pero nunca en la ideología política. Asimismo, los prototipos de detección desarrollados hasta ahora no usan bases de datos generadas manualmente por varios codificadores. Esta investigación busca superar estas limitaciones, desarrollando y evaluando un detector automático de discurso de odio por motivos ideológicos en Twitter en español a partir de técnicas de aprendizaje automático supervisado. Para ello, se ha desarrollado un total de ocho modelos predictivos a partir de un corpus de entrenamiento generado ad-hoc, y haciendo uso de modelado superficial y de aprendizaje profundo, lo que permite mejorar el rendimiento final del prototipo. El desarrollo del corpus permitió observar, además, que un 16,2% de la muestra, recogida en el otoño de 2019, incluyó algún tipo de odio ideológico.

En otra investigación, Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021), afirman que la libertad de expresión no siempre es respetuosa ya que en ocasiones se utiliza un lenguaje ofensivo o insultante ya sea en redes sociales, blogs y foros. Por ello, presentan un análisis comparativo de diferentes técnicas de Machine Learning (ML) en la clasificación de hate speech en las redes sociales. Ambos conjuntos de datos en español se recopilaron de HaterNet y HatEva. Para ello realizamos diferentes experimentos utilizando tres enfoques basados en dos técnicas de lenguaje natural y tres modelos de enfoque tradicional de Machine Learning.

Por otra parte, según Arcila Calderón, C., Blanco-Herrero, D., & Valdez Apolo, M. B. (2020) en su investigación tiene como objetivo detectar el rechazo a los extranjeros y refugiados en un entorno hispanohablante de información extraída de Twitter, la cual fue clasificada con relación al sentido o presencia de expresiones de rechazo y asociaciones negativas que justifican el rechazo. Además, se empleó la data para entrenar los modelos en donde se obtuvo como resultados los siguientes accuracy: Nave Bayes original, 75,36%; Naive Bayes para modelos multimodales, 74,64%; Naive Bayes para modelos multivariados Bernoulli, 72,15%; Regresión Logística, 74,53%; Regresión lineal con gradiente descendente estocástico, 71,84% y máquinas de vectores soporte (SVC) con 76,19% de accuracy. Por lo tanto, la investigación detectó rechazo hacia migrantes y refugiados, en donde finalmente se concluye que por un lado, la enorme variación que se puede observar en las expresiones de rechazo o aceptación de inmigrantes en las redes sociales en función de los últimos fenómenos mediáticos; y, por otro lado, la presencia de rechazo en redes sociales que en algunos casos puede estar basado en sentimientos y actitudes racistas o xenófobos y que en ocasiones es manifestada a través de discursos de odio.

Para Koushik, G., Rajeswari, K., & Muthusamy, S. K. (2019), el creciente uso de las redes sociales en estos últimos años ha generado un gran volumen de contenido realizado por los usuarios en línea, sin embargo, junto a ello ha habido un aumento en el discurso de odio y la intimidación en las plataformas de redes sociales conllevando a una mayor preocupación por la seguridad en línea y protección de los usuarios. Por ende, la investigación buscó desarrollar una variedad de técnicas de detección de discurso de odio en línea incluyendo preprocesamiento de lenguaje natural y aprendizaje automático para clasificar automáticamente los tweets en dos categorías:

discurso de odio o discurso no odio. Utilizaron conjunto de datos públicos de Twitter, aplicaron características de Bag of Words y TF/IDF para entrenar el clasificador de regresión logística. Los resultados del estudio mostraron que el clasificador tuvo un accuracy del 94.11%, esto permitió proporcionar una herramienta para la identificación y filtrado de discursos de odio en línea.

En base a lo mencionado, el discurso de odio se ha intensificado desde la pandemia de COVID-19 en las redes sociales, especialmente en Twitter, por ello se ha empleado el aprendizaje automático para detectar y clasificar el discurso de odio. Existen diferentes técnicas y modelos de aprendizaje automático utilizados para este propósito, los cuales han sido evaluados y comparados en términos de eficacia. Estos modelos tienen ciertas limitaciones, sin embargo, se mantiene la necesidad de abordar el discurso de odio en diferentes idiomas y contextos, incluyendo la ideología política. En el Perú, hubo una gran cantidad de discursos de odio en las redes sociales en las regiones peruanas de Cusco, Arequipa, Puno y otras, con el propósito de desestabilizar el régimen actual y promover protestas no pacíficas, incluyendo la toma de Lima 2023. Estas protestas afectaron la economía del país y se reflejaron en una disminución del PBI. Debido a lo mencionado, se propone la aplicación de la Inteligencia Artificial y técnicas de Machine Learning para la detección del discurso de odio en un contexto político social con el fin de identificar el modelo más apropiado para el análisis de la opinión pública en un corto periodo de tiempo.

II. MATERIALES Y MÉTODOS

A. Materiales

En este estudio se utilizaron dos herramientas fundamentales (Google Colaboratory y Twint) para la recolección, desarrollo y análisis de datos de Twitter:

1. GOOGLE COLABORATORY (GOOGLE COLAB)

Naik, Naik y Patil (2022), explican que Google Colaboratory, tambien conocido como Colab, permite a cualquier persona escribir y ejecutar código en Python a través del navegador, este es un servicio gratuito para crear modelos de machine learning y contiene módulos de herramientas para el análisis de ciencia de datos. Por lo tanto, se utilizó este recurso debido a que es una herramienta gratuita de Google que permitió la ejecución de código en Python y la creación de cuadernos en la nube. Asimismo, contribuyó a tener un mismo entorno de programación colaborativa y de compartir el código fuente dentro del equipo de trabajo.

2. TWINT

Silvia et al. (2022), señalan que existen algunas aplicaciones como Twint que ayudan en la recolección de información con fines de análisis. Twint es una herramienta desarrollada en Python que recibe parámetros de búsqueda con el fin de obtener tweets y devuelve los datos en formato JSON o CSV. A partir de ello, se eligió Twint ya que es una herramienta para recolectar los tweets de manera automatizada y eficiente, sin utilizar la API de Twitter.

A continuación, se va a presentar en la figura 1, el esquema de trabajo de las respectivas herramientas.

Según la figura 1, se observa el mecanismo de trabajo en el uso de las herramientas de recolección, desarrollo y análisis de datos, teniendo como punto de inicio la codificación en Python, que a través del entorno de desarrollo Google Colaboratory y la integración de la biblioteca Twint permitió tener un funcionamiento asíncrono *(nest_asyncio)* y con una debida configuración de parámetros (número máximo de tweets, idioma, popularidad, likes, fecha, y exportación de los *datasets* con formato *csv* de los datos del centro y sur de las regiones del Perú).

El mecanismo de trabajo desarrollado con las herramientas Google Colaboratory y Twint posibilitó recolectar, preprocesar y analizar grandes cantidades de tweets, los cuales conllevó a construir y evaluar los modelos de Machine Learning de la presente investigación.

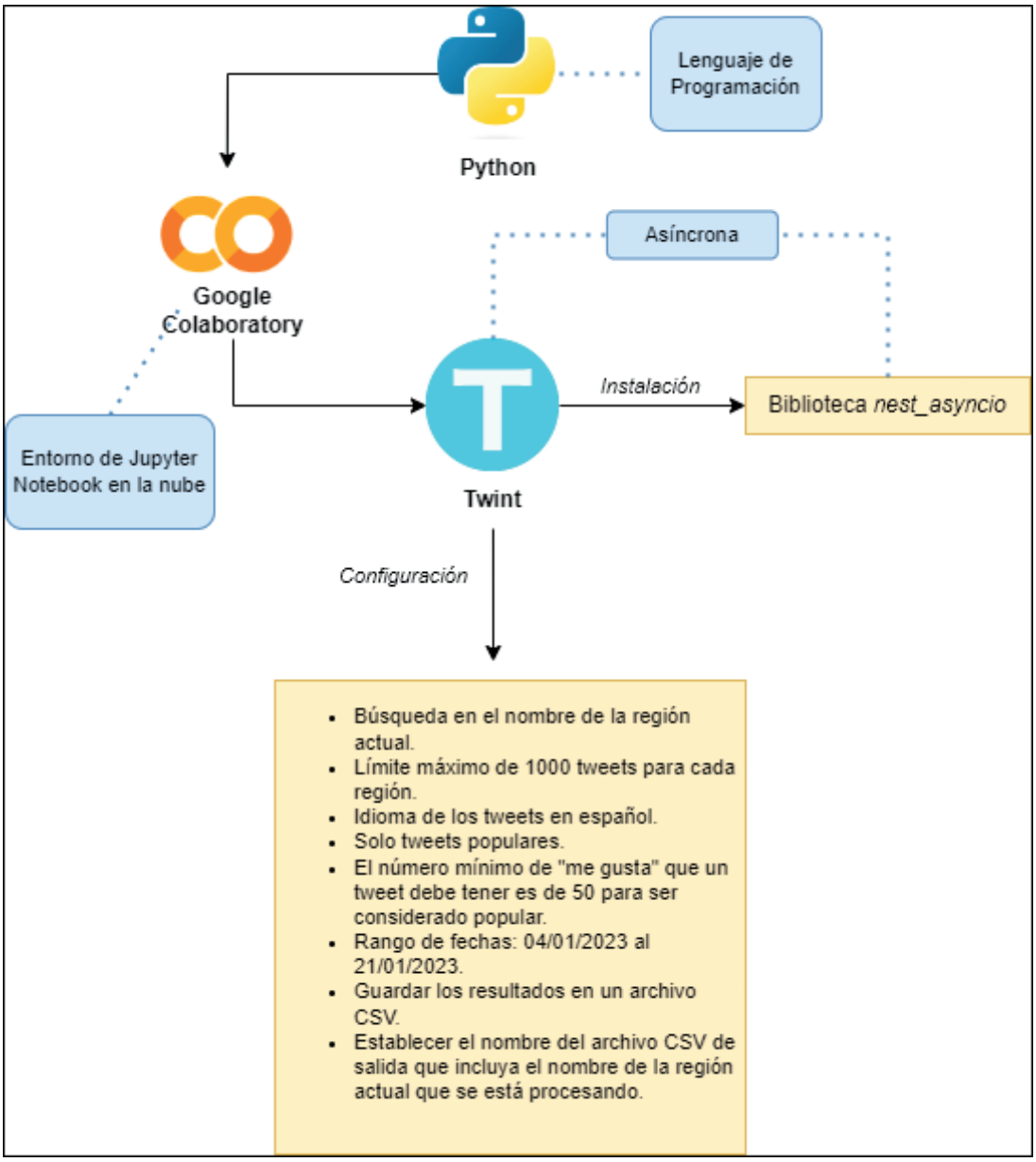
B. Métodos

En esta sección detallamos todo el proceso de la metodología SEMMA que según López-Torres et al. (2020), describen como un proceso principal de desarrollo de modelos de minería de datos ampliamente utilizada y aceptada.

Según, Rotondo y Quilligan (2020), esta metodología se basa en los pasos del proceso clásico de KDD, ya que existe un paralelismo con las tareas del modelo de esquema KDD. SEMMA comparado con KDD no necesita de requisitos previos en la comprensión comercial pero sí se puede incluir en la fase de muestra ya que los datos no pueden muestrearse hasta que haya una comprensión clara tanto del conjunto de datos como de los objetivos de los interesados (Tariq et al., 2019).

Figura 1

Esquema de trabajo en el uso de Google Colaboratory y Twint



Fuente: Elaboración propia

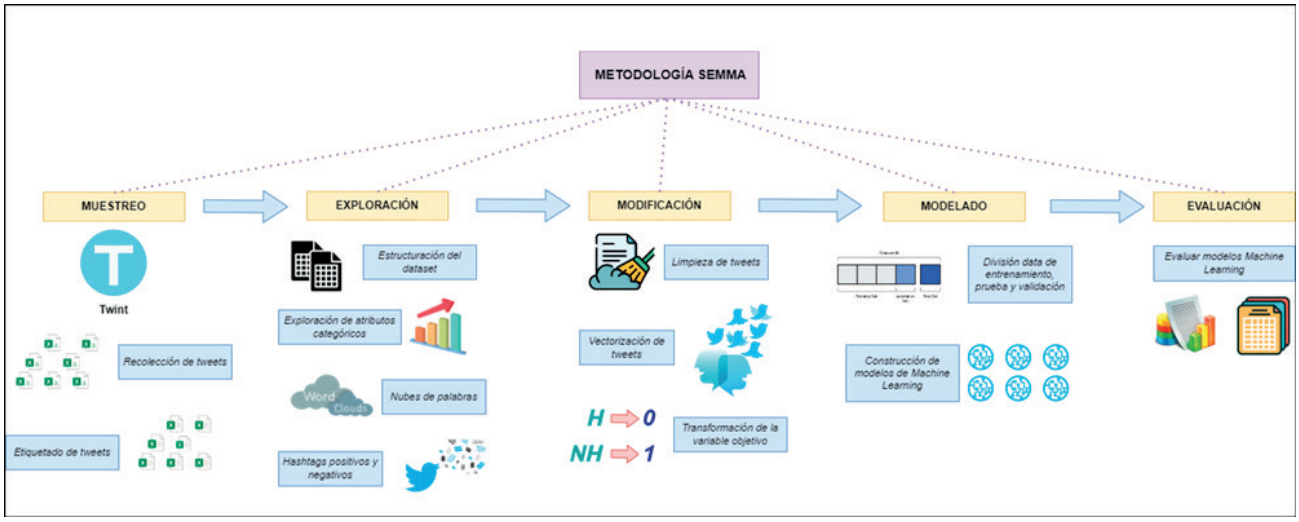
Por lo tanto, SEMMA brinda un enfoque estructurado y sistemático para trabajar de manera eficiente, flexible y efectiva para asegurar que los modelos resultantes sean lo más precisos posible y se adapten bien a los datos de este estudio.

La metodología SEMMA consta de cinco fases: Muestrear (*Sample*), Explorar (*Explore*), Modificar (*Modify*), Modelar (*Model*) y Evaluar (*Access*) para la obtención de un control preciso de las actividades y validar la capacidad y calidad de la investigación propuesta (De-La-Hoz-Correa et al., 2019). A continuación, vamos a presentar la figura 2, el esquema completo de la metodología SEMMA aplicado en el presente proyecto de investigación.

Según la figura 2, se muestra de manera general todo el proceso de la metodología SEMMA aplicada en la investigación. En primer lugar, se llevó a cabo la fase de muestreo, la cual implicó la extracción y etiquetado de los tweets. Posteriormente, se procedió a la fase de exploración, en la cual se estructuró

Figura 2

Metodología SEMMA aplicada al presente estudio



Fuente: Elaboración propia

toda la data (*csv* de los datos del centro y sur de las regiones del Perú), se exploraron los atributos categóricos, se generaron nubes de palabras y se identificaron hashtags positivos y negativos. En la siguiente fase, se realizó la modificación, la cual incluyó la limpieza de los datos, la vectorización de los tweets y la transformación de la variable objetivo. Luego, se procedió a la fase de modelado, en la cual se dividió la data de entrenamiento en conjunto de validación y prueba, y se construyeron modelos de Machine Learning. Finalmente, se llevó a cabo la fase de evaluación de los modelos a través de reportes detallados logrando completar todo el proceso de la metodología. A continuación, vamos a detallar todo el proceso de la respectiva metodología de la siguiente manera:

1. *FASE DE MUESTREO*

Durante la primera fase, se extrajeron tweets (datos de siete regiones definidas en base a las tendencias en Perú) desde el 4 de enero de 2023 hasta el 21 de enero de 2023, cuando estaban aconteciendo las protestas al interior del país, utilizando la librería Twint en el entorno de desarrollo Google Colaboratory.

La figura 3, muestra las palabras o temas más mencionados en las siete regiones del país: Lima, Puno, Ayacucho, Cuzco, Arequipa, Tacna y Madre de Dios. Estos temas son evidenciados por los sucesos más relevantes que ocurrieron en cada región.

Figura 3

Las tendencias en Twitter en Perú ocurridas en a principios del mes enero

Fuente: Elaboración propia

Aplicando los parámetros establecidos en la configuración de Twint, se recopilaron 6.782 tweets, de los cuales se revisaron y etiquetaron manualmente 1.970 tweets. Es importante destacar que el etiquetado de los tweets como *"Hate"* o *"No Hate"* es una tarea compleja que requiere una cuidadosa consideración de diferentes factores.

A continuación, se muestra gráficamente el proceso del muestreo de los tweets aplicando la herramienta Twint en la presente investigación:

Según la figura 4 se observa que se llevó a cabo la extracción de tweets utilizando la herramienta Twint y su configuración correspondiente. Como resultado de este proceso, se generaron siete archivos con formato *csv* (uno para cada región). Posteriormente, se llevó a cabo la tarea de etiquetar manualmente 1970 tweets, con el objetivo de utilizarlos como datos de entrenamiento. Es importante destacar que solo se etiquetaron aquellos tweets relacionados con el contexto social y político de Perú, excluyendo cualquier otro tema.

2. *FASE DE EXPLORACIÓN*

En esta fase se buscó descubrir la información significativa y detectar posibles anomalías en los datos.

a. Estructuración de la dataset

Los datos proporcionados tienen tres atributos relacionados y todos están compuestos por categóricos.

Tabla 1

Descripción de la dataset del muestreo

Atributo	Tipo de dato	Descripción
tweet	*object*	Contenido del tweet
region	*object*	Lugar donde se realizó el tweet
detection	object	Etiqueta de detección de odio

Fuente: Elaboración propia

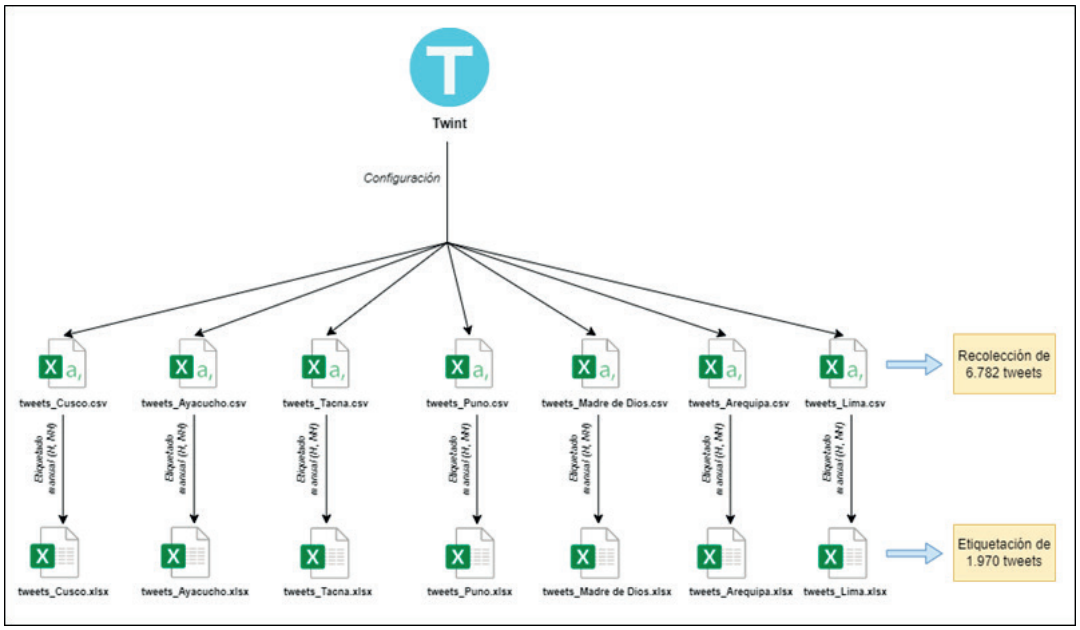
En primer lugar, el contenido del tweet proporciona información valiosa sobre el tipo de lenguaje utilizado en las publicaciones que fueron consideradas en la muestra. De esta manera, se pudo identificar la presencia de elementos discriminatorios o de odio en el discurso de los usuarios de la red social en cuestión.

En segundo lugar, la región donde se realizó el tweet permite contextualizar la información y entender cómo la expresión de odio varía de acuerdo con las características políticas y sociales de cada lugar. Esta variable es crucial para obtener un análisis más preciso y completo de los datos.

Por último, la detección de odio es el resultado del etiquetado manual que se realizó en la investigación,

Figura 4

Muestreo de tweets usando Twint



Fuente: Elaboración propia

en el cual se clasificaron los tweets en dos categorías: *"Hate"* y *"No Hate"*. Esta variable es fundamental para la fase de modelado y evaluación.

b. Exploración de atributos categóricos

En esta sección del estudio se realiza una exploración de los atributos categóricos. Se recopilan estadísticas de resumen para cada uno de los tres atributos, incluyendo la presencia de valores nulos en relación con su categoría correspondiente. Asimismo, se examina la distribución del atributo *detection* y del atributo *region*. Se hace un resumen de los valores nulos para cada atributo, con el fin de tener una mejor comprensión de la calidad de los datos y su influencia en los resultados del análisis.

Del análisis de las figuras 5 y 6 se visualiza que hay una gran cantidad de valores nulos para la categoría de detección, con solo 1970 filas que no son nulas. Debido a esto, se procedió a eliminar las instancias correspondientes, dejando un total de 1970 tweets distribuidos equitativamente entre las categorías de *"Hate"* y *"No Hate"*.

En esta investigación, además de visualizar la distribución de discurso de odio (*Hate*) y sin odio (*No Hate*) en general, se procedió a analizar la distribución por separado para cada una de las siete regiones consideradas en el análisis.

Al analizar la figura 7 y figura 8, se observan que la categoría de detección (*detection*) tiene la misma cantidad de instancias tanto para el discurso de odio (*Hate*) como para el discurso sin odio (*No Hate*). Sin embargo, al evaluar la distribución de las instancias por región, se puede notar que en las ciudades de Arequipa y Lima el discurso de odio es mayor en comparación con las demás regiones consideradas en este estudio.

c. Nubes de palabras (Wordclouds)

En este estudio, se consideró que el atributo categórico "*tweet*" contenía información relevante que

Figura 5

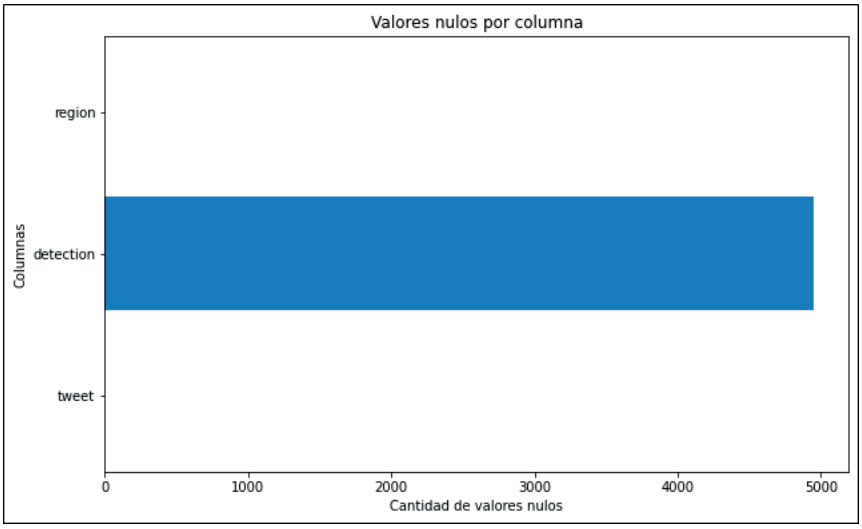
Resumen de valores nulos

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6918 entries, 0 to 1001
Data columns (total 3 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   tweet      6917 non-null   object
 1   detection  1970 non-null   object
 2   region     6918 non-null   object
dtypes: object(3)
memory usage: 474.2+ KB
```

Fuente: Elaboración propia

Figura 6

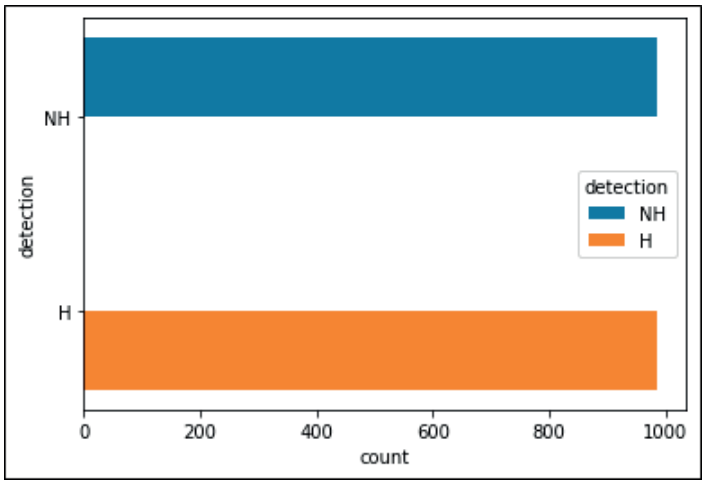
Gráfico del resumen de valores nulos



Fuente: Elaboración propia

Figura 7

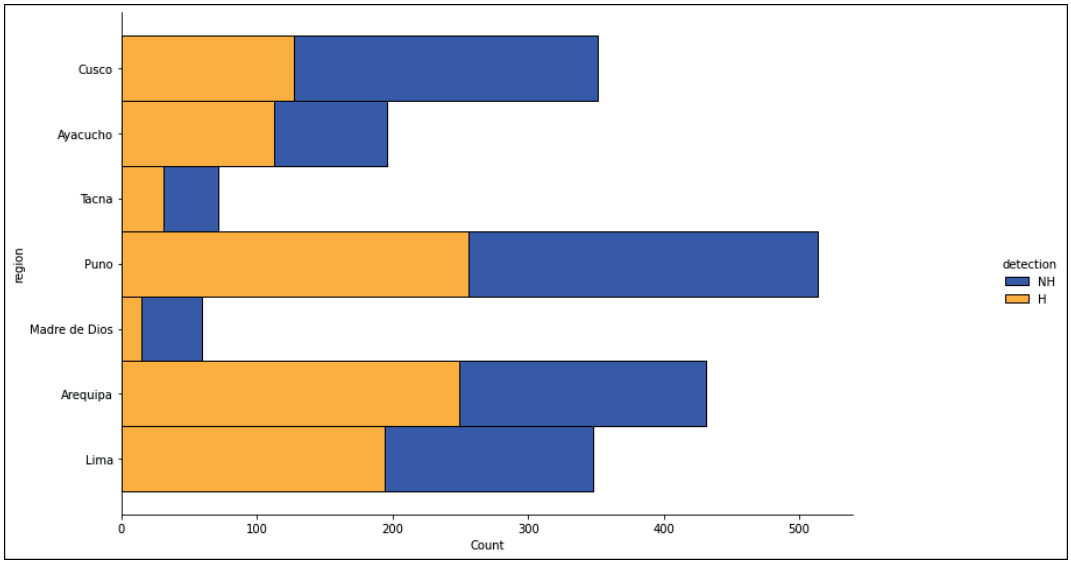
Gráfico de la distribución del atributo categórico detection



Fuente: Elaboración propia

Figura 8

Gráfico de distribución del atributo categórico region



Fuente: Elaboración propia

debía ser explorada. Para lograr esto, se utilizó una técnica de representación visual conocida como "nubes de palabras" (*Wordclouds*), para mostrar las palabras más frecuentes en cada una de las siete regiones analizadas. De esta manera, se pudo obtener una mejor comprensión del lenguaje utilizado en cada una de estas regiones y detectar patrones relevantes relacionados con el discurso de odio en el contexto político y social ocurrido en Perú. A continuación, vamos a mostrar palabras de Tweets representado a través de nubes de las siete regiones del Perú.

De las figuras 9, 10, 11, 12, 13, 14 y 15 mostradas, se observaron las palabras y personajes más mencionados en cada región, lo que permitió conocer el contexto social y político en el que se desenvuelve el discurso de odio. Entre las palabras más frecuentes se encuentran "protesta", "manifestación", "bloqueo" y "gas", entre otras. Sin embargo, también se encontraron palabras y calificativos de odio como "terrorista" e "izquierdista" en regiones como Arequipa *(a)*, Ayacucho *(b)* y Cuzco *(c)*.

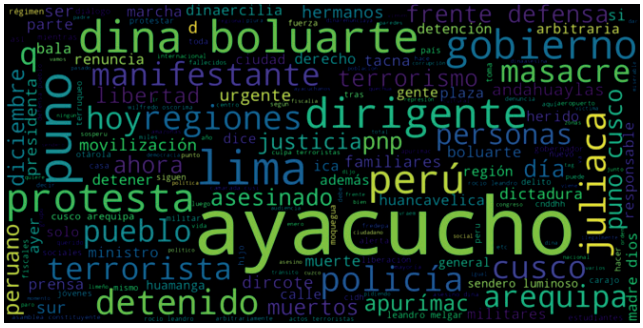
Figura 9

Nube de palabras de Tweets de la región de Arequipa

Fuente: Elaboración propia

Figura 10

Nube de palabras de Tweets de la región de Ayacucho



Fuente: Elaboración propia

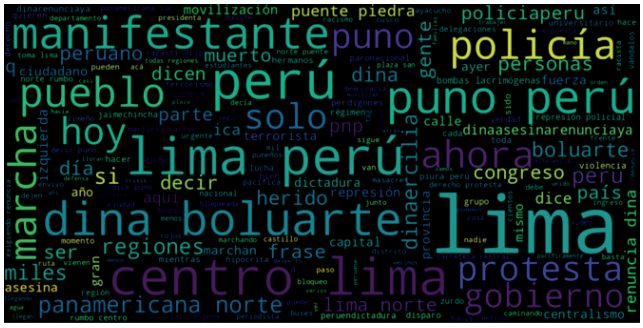
Figura 11

Nube de palabras de Tweets de la región de cuzco

Fuente: Elaboración propia

Figura 12

Nube de palabras de Tweets de la región de Lima



Fuente: Elaboración propia

Figura 13

Nube de palabras de Tweets de la región de Madre de Dios

Fuente: Elaboración propia

Figura 14

Nube de palabras de Tweets de la región de Puno

Fuente: Elaboración propia

Figura 15

Nube de palabras de Tweets de la región de Tacna

Fuente: Elaboración propia

Esta exploración de datos proporciona información valiosa sobre los temas y preocupaciones más frecuentes en cada una de las regiones. Además, pueden ser útiles para la implementación de políticas públicas y programas de prevención del discurso de odio en diferentes regiones.

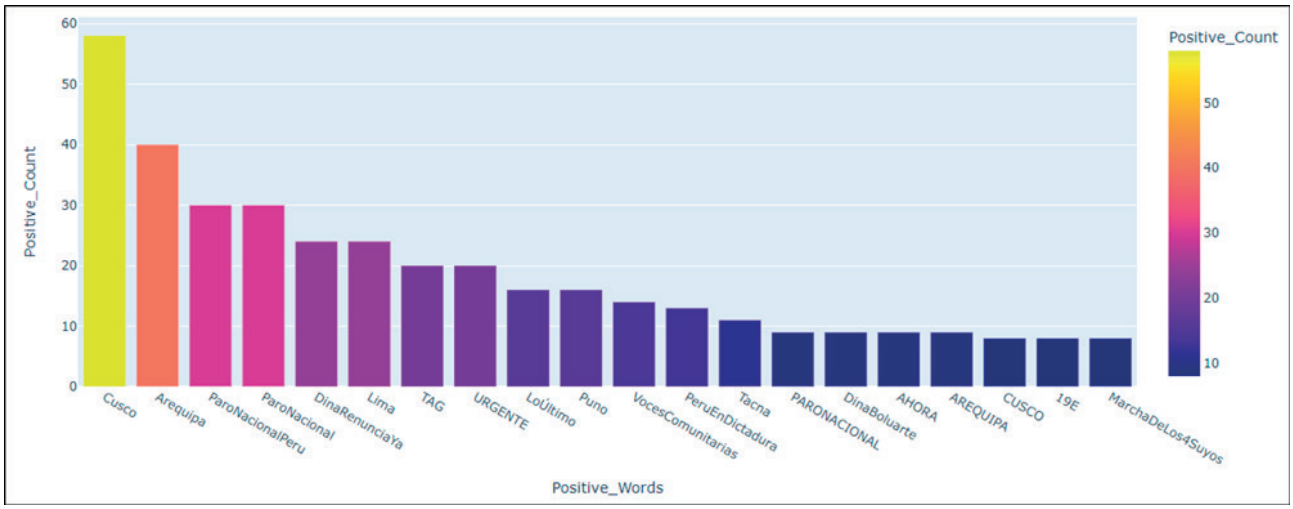
d. Hashtags positivos y negativos

Con el fin de explorar aún más cada uno de los tweets en el conjunto de datos, se extrajeron las palabras que contenían hashtags (#) y se almacenaron en una columna separada. Esta información fue utilizada para generar una representación gráfica de los hashtags positivos y negativos encontrados en los tweets. A partir de ello, este análisis permite identificar los temas y preocupaciones más frecuentes en el discurso de odio.

En las figuras 16 y 17, se puede observar que los hashtags positivos que son más frecuentes en comparación con los negativos. Este resultado sugiere que, en general, existe una tendencia hacia el uso de hashtags con una connotación positiva en los tweets analizados.

Figura 16

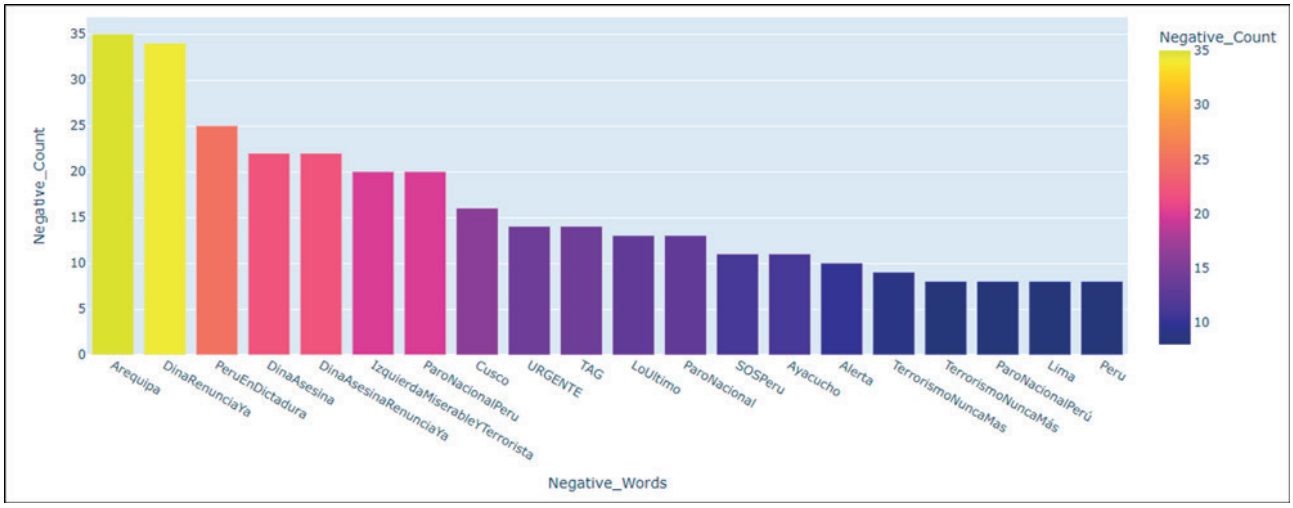
Hashtags de palabras positivas



Fuente: Elaboración propia

Figura 17

Hashtags de palabras negativas



Fuente: Elaboración propia

3. FASE DE MODIFICACIÓN

En esta fase, se lleva a cabo el proceso de limpieza y vectorización de los atributos necesarios. La finalidad es convertir todos los atributos categóricos con tipo de datos de caracteres en atributos numéricos para poder aplicar correctamente los modelos de aprendizaje automático. Este proceso implica eliminar información innecesaria o irrelevante, reducir palabras a su forma base y transformar los atributos en vectores de características numéricas. De esta manera, se garantiza que los datos sean aptos para el análisis y la aplicación de este presente estudio.

a. Limpieza de tweets

La limpieza de los tweets es una parte esencial del procesamiento de lenguaje natural. Para realizar esta tarea, se utilizaron varias librerías, como *"re"*, *"nltk"* y *"stopwords"*. Estas librerías son muy útiles para aplicar funciones de búsqueda y reemplazo, así como para importar una lista de *stopwords* y eliminarlas del texto.

En la tabla 2 se muestra las librerías consideradas en este proceso en donde se eliminaron elementos no relevantes como las menciones, URLs, caracteres especiales y signos de puntuación que no aportan significado al texto. Además, se eliminaron las palabras vacías o *stopwords*, que son palabras comunes que no tienen un significado relevante para el análisis de los tweets.

En la figura 18, se observa la columna *"clean_tweet"*, la cual contiene palabras clave con sentido después de llevar a cabo la limpieza y tokenización de la columna *"tweet"*.

b. Vectorización

En esta investigación, se utilizó la técnica de vectorización *CountVectorizer* de la librería *sklearn (scikit-learn)*. Dicha técnica permitió ajustar el modelo a los datos de entrenamiento, en este caso el atributo *"tweet"*, y transformarlos en una matriz de características.

En la figura 19, se presenta una lista de las palabras únicas encontradas en los tweets, la cual se encuentra ordenada alfabéticamente.

En la figura 20, se muestra la transformación de la variable objetivo mediante la función *"dummies"*. Esta función toma una variable categórica y la convierte en varias variables binarias correspondientes a cada categoría. En el caso de esta investigación, las categorías fueron *"Hate"* y *"No Hate"*, siendo transformadas a los valores 0 y 1, respectivamente.

4. FASE DE MODELADO

a. División data de entrenamiento *(train)*, prueba *(test)* y validación *(valid)*

En la presente investigación, es fundamental dividir la data en conjuntos de entrenamiento, validación y prueba, ya que permite obtener modelos más robustos al evaluar su capacidad de generalización en datos no vistos previamente.

En la tabla 3 se muestra la división de la data. Para ello, se utilizó una muestra de datos de 1477 instancias. Asimismo, se destinó un porcentaje del 15% de los datos totales, es decir, 295 instancias, para llevar a cabo la validación del modelo. Por último,

Tabla 2

Librerías Python para la limpieza de tweets

Librerías	Uso
re	Remover las expresiones regulares que contienen URLs, caracteres especiales y signos de puntuación.
nltk	Conjunto de librerías para preprocesamiento de texto.
stopwords	Remover las palabras vacías (*stopwords*) del texto para mejorar la calidad del análisis y el desempeño del modelo de aprendizaje automático

Fuente: Elaboración propia

Figura 18

Limpieza y tokenización de la columna tweet

	tweet	clean_tweet
0	La impresionante movilización de ciudadanos re...	impresionante movilización ciudadanos resident...
1	Una familia d "pudientes cusqueños" descendien...	familia d pudientes cusqueños descendientes d ...
17	Gracias a la #IzquierdaMiserable ,el turismo e...	gracias izquierdamiserable turismo cusco quebr...
18	¿Cómo fue que, de un imperio incaico repleto d...	cómo imperio incaico repleto genios organizaci...
20	Los q hicieron campaña y pedían "NO VIAJAR A...	q hicieron campaña pedían viajar cusco jodan h...

Fuente: Elaboración propia

Figura 19

Vectorización de tweets

	000	03	06	07	10	100	1000	100mil	108	1095	...	Panamericana	Fuera	Las	en	indigenas	la	movilizacion	mujeres	nacional	presentes
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1965	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1966	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1967	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1968	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1969	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

Fuente: Elaboración propia

Figura 20

Transformación del atributo detection

detection	region	Hash words	clean_tweet
0	Cusco	No hashtags	impresionante movilización ciudadanos resident...
1	Cusco	No hashtags	familia d pudientes cusqueños descendientes d ...
1	Cusco	#IzquierdaMiserable	gracias izquierdamiserable turismo cusco quebr...
1	Cusco	No hashtags	cómo imperio incaico repleto genios organizaci...
1	Cusco	No hashtags	q hicieron campaña pedían viajar cusco jodan h...

Fuente: Elaboración propia

Tabla 3

Partición de datos

Data	% de Data	Número de instancias
Entrenamiento *(train)*	75 %	1477
Validación *(validation)*	15 %	295
Prueba *(test)*	10 %	198

Fuente: Elaboración propia

se destinó un 10% del total de los datos, lo que equivale a 198 instancias, para realizar las pruebas pertinentes al modelo. Cabe resaltar que el proceso de partición de los datos en diferentes conjuntos permitió evaluar la capacidad de generalización del modelo.

b. Modelos de Machine Learning

En la investigación se utilizaron modelos de clasificación binaria para predecir el discurso de odio en los tweets. Específicamente, se entrenaron dos tipos de modelos: uno sin validación cruzada y otro con validación cruzada. La validación cruzada es una técnica que consiste en dividir el conjunto de entrenamiento en varias partes y utilizarlas para validar el modelo en diferentes iteraciones. Se realizaron estas dos versiones de los modelos para comparar sus resultados del *accuracy* en la predicción del discurso de odio en los tweets de la data de validación.

1. Multinomial Naive Bayes

El modelo fue construido utilizando la función *MultinomialNB()* de la librería *sklearn.naive_bayes*, en donde es necesario importarla previamente. Dicha función hizo entrenar el modelo y ajustar los parámetros utilizando el conjunto de datos de entrenamiento. Una vez que este modelo fue entrenado para realizar predicciones sobre la detección de odio en los tweets, sobre el conjunto de datos de prueba, se generó un reporte de clasificación para obtener la exactitud.

2. Random Forest

El modelo fue construido utilizando la librería *sklearn.ensemble* ya que proporciona la función *RandomForestClassifier() para el ajuste del modelo* de clasificación a conjuntos de datos de entrenamiento. Esta función admite varios parámetros que permiten ajustar y personalizar el modelo, entre ellos el número de árboles utilizados para el proceso de entrenamiento *(n_estimators)*. En el caso particular de este conjunto de datos, se ha utilizado un valor de 100 para este parámetro, debido al tamaño del conjunto de datos (1477 instancias y 8233 características).

3. Logistic Regression

El modelo fue construido utilizando la función *LogisticRegression()* la cual proviene de la librería *sklearn*, ya que permite entrenar, predecir y evaluar la exactitud del modelo. Los parámetros principales del modelo incluyen el parámetro de regularización (C) con el fin de evitar reajuste. Sin embargo, en este estudio no se incluyeron, debido a que previamente se hizo una selección de características y una posterior validación cruzada para reducir la complejidad del modelo.

4. Support Vector Classifier

El modelo fue construido utilizando la función *LinearSVC()* de la biblioteca *scikit-learn*, que provee una implementación del clasificador de Vectores de Soporte Lineal (SVM). En este caso, se decidió inicializar el modelo con números aleatorios *(random_state=42)*.

5. AdaBoost

El modelo fue construido utilizando *AdaBoostClassifier* en la cual se consideró dos parámetros principales: El estimador base, que en este caso es *DecisionTreeClassifier()*, y el número de estimadores, que es 200. Es decir, se construyó un conjunto de 200 clasificadores débiles basados en árboles de decisión simples con una profundidad máxima de 1. El algoritmo de conjunto de AdaBoost luego combina estos clasificadores débiles para formar un clasificador fuerte y más preciso.

6. Gradient Boosting

El modelo fue construido utilizando *GradientBoostingClassifier()* en la cual se consideró varios parámetros con valores predeterminados. En este caso particular, se ajustaron 100 clasificadores débiles (árboles de decisión) con una profundidad máxima de 3, utilizando la tasa de aprendizaje predeterminada de 0.1 y un submuestreo aleatorio del 100% de los datos en cada iteración.

5. FASE DE EVALUACIÓN

La fase de evaluación es esencial para determinar la exactitud promedio de los modelos de aprendizaje automático. En este caso, se presentan los resultados de dos evaluaciones distintas para diferentes modelos.

Tabla 4

Resultados sin validación cruzada

Modelos	Exactitud Promedio (%)
Multinomial Naive Bayes	71,18%
Random Forest	71,52%
Logistic Regression	71,86%
Support Vector Classifier	71,18%
AdaBoost	67,45%
Gradient Boosting	70,84%

Fuente: Elaboración propia

En la tabla 4, se observa que los modelos con mayor exactitud promedio (%) son Regresión Logística con 71.86% y Random Forest con 71.52%.

Tabla 5

Resultados con validación cruzada

Modelos	Exactitud Promedio (%)
Multinomial Naive Bayes	73,87%
Random Forest	71,84%
Logistic Regression	73,12%
Support Vector Classifier	73,39%
AdaBoost	66,08%
Gradient Boosting	68,12%

Fuente: Elaboración propia

Se analiza que la técnica de validación cruzada ha mejorado significativamente la fiabilidad y representatividad de los resultados obtenidos en la tabla 5, en comparación con los resultados sin validación cruzada en la tabla 4. En lugar de dividir los datos en conjuntos de entrenamiento y prueba, se dividen en k=10 subconjuntos. El modelo se entrena k=10 veces con diferentes subconjuntos como prueba y los demás como entrenamiento. Al final, se promedian los resultados para obtener una estimación del rendimiento del modelo.

III. RESULTADOS

Se recolectó información de Twitter, la cual se clasificó en dos categorías: *'Hate'* para tweets con opiniones negativas, discriminatorias y/o despectivas, y *'No Hate'* para comentarios donde se expresa una opinión sin agresividad u hostilidad. A partir de ello, se seleccionaron 985 tweets de cada categoría que fueron limpiados y procesados para entrenar seis modelos de aprendizaje automático: Multinomial Naive Bayes, Random Forest, Logistic Regression, Support Vector Classifier, AdaBoost y Gradient Boosting. La finalidad fue determinar cuál de los modelos es el más efectivo para el conjunto de datos. De acuerdo con la fase de evaluación, observamos que los resultados en la tabla 5 fueron los más óptimos y confiables, debido a que se utilizó la función *cross_val_score* de la biblioteca *Scikit-learn* para calcular la exactitud mediante validación cruzada con 10 folds estratificados y con una división aleatoria de los datos en cada fold. Para cada modelo, se creó una instancia y se utilizó la función *cross_val_score* para calcular la exactitud promedio en todos los folds.

Por lo tanto, de los seis modelos de Machine Learning construidos, vemos que en su mayoría son valores superiores al 70%, destacando el modelo Multinomial Naive Bayes con un 73,87% y seguido de este el modelo Support Vector Classifier con 73,39%.

IV. RECOMENDACIONES

Para mejorar el alcance y la calidad del estudio, se sugiere tomar en cuenta las siguientes recomendaciones:

En primer lugar, se recomienda ampliar el rango de fechas y fuentes de extracción de datos. Esto permitiría obtener una muestra más diversa y representativa de la información verbal y no verbal que circula en las redes sociales. En segundo lugar, es importante destacar que la vectorización de tweets usando *CountVectorizer* es solo el primer paso para el procesamiento de lenguaje natural. A menudo, es necesario realizar preprocesamiento adicional, como la eliminación de *stopwords*, antes de vectorizar los tweets para obtener mejores resultados en la clasificación y el análisis de texto. En tercer lugar, se recomienda incluir el análisis de información no verbal, como los emojis, en el estudio. Estos elementos también pueden ser indicadores de odio y, por lo tanto, es importante considerarlos en el análisis para obtener un análisis completo de la situación.

Por lo tanto, se sugiere destacar la importancia de los resultados obtenidos en el estudio y cómo estos pueden ser aplicados en el desarrollo de futuros sistemas de detección de odio en tiempo real en diferentes redes sociales. Esto permitiría resaltar el impacto y la relevancia de la investigación en el campo de la detección de odio en línea. Siguiendo estas recomendaciones, se espera tener una mayor relevancia y aporte a la Inteligencia Artificial.

V. CONCLUSIONES

En este artículo se presenta los resultados obtenidos de diferentes modelos de Machine Learning para detectar el discurso de odio relacionado con el contexto político social peruano, estos modelos fueron entrenados a partir de data extraída de Twitter con ayuda de la librería Twint, la cual fue clasificada en discursos con odio y no odio. De los seis modelos realizados el modelo con la mayor exactitud promedio es Multinomial Naive Bayes, seguido por Support Vector Classifier.

Finalmente, podemos destacar que la metodología aplicada y resultados obtenidos sirven para acelerar el proceso de detección del discurso de odio verbal en redes sociales y, a su vez, permite identificar el modelo más adecuado para el análisis de la opinión pública en un corto periodo de tiempo, la cual refleja que el estado actual de un país influye en el comportamiento de sus integrantes y por ende, en la economía del sector peruano, además de que permite la detección de presuntos incitadores de odio por medio de la red social Twitter.

VI. REFERENCIAS

[1] Amores, J.,Blanco-Herrero, D., Sánchez-Holgado, P., Frías-Vázquez, M.(2021). ”Detectando el odio ideológico en Twitter. Desarrollo y evaluación de un detector de discurso de odio por ideología política en tuits en español”. Universidad de Salamanca (España). Cuadernos.info, 49, 98–124. https://doi.org/10.7764/cdi.49.27817

[2] De-La-Hoz-Correa, E., Mendoza-Palechor, F. E., De-La-Hoz-Manotas, A., Morales-Ortega, R. C., & Adriana, S. H. B. (2019). Obesity Level Estimation Software based on Decision Trees. *Journal of Computer Science*, *15*(1), 67-77. https://doi.org/10.3844/jcssp.2019.67.77

[3] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., & Malik, S. H. (2022). Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques.

International Journal of Information Management Data Insights, *2*(2), 100120. https://doi.org/10.1016/j.jjimei.2022.100120

[4] López-Torres, S., López-Torres, H., Rocha-Rocha, J., Aziz Butt, S., Imran Tariq, M., Collazos-Morales, C., & Piñeres-Espitia, G. (2020). IoT Monitoring of Water Consumption for Irrigation Systems Using SEMMA Methodology. *Intelligent Human Computer Interaction*, 222-234. https://doi.org/10.1007/978-3-030-44689-5_20

[5] Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166(114120), 114120. https://doi.org/10.1016/j.eswa.2020.114120

[6] Chhabra, A., & Vishwakarma, D. K. (2023). A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*. https://doi.org/10.1007/s00530-023-01051-8

[7] Lilleker, D., & Pérez-Escolar, M. (2023). Demonising migrants in contexts of extremism: Analysis of hate speech in UK and Spain. *Politics and governance*, *11*(2). https://doi.org/10.17645/pag.v11i2.6302

[8] Naik, P., Naik, G., & Patil, M. B. (2022). Conceptualizing Python in Google COLAB.

[9] Parvaresh, V. (2023). Covertly communicated hate speech: A corpus-assisted pragmatic study. *Journal of Pragmatics*, *205*, 63–77. https://doi.org/10.1016/j.pragma.2022.12.009

[10] Rodríguez, M. (15 de marzo de 2023). INEI: Economía peruana cayó 1,12% en enero debido a protestas. *El Comercio.* https://elcomercio.pe/economia/peru/economia-peruana-cayo-112-en-enero-por-protestas-noticia/

[11] Rotondo, A., & Quilligan, F. (2020). Evolution Paths for Knowledge Discovery and Data Mining Process Models. *SN Computer Science*, *1*(2). https://doi.org/10.1007/s42979-020-0117-6

[12] Romero-Rodríguez, L. M., Castillo-Abdul, B., & Cuesta-Valiño, P. (2023). The process of the transfer of hate speech to demonization and social polarization. *Politics and governance*, *11*(2). https://doi.org/10.17645/pag.v11i2.6663

[13] Silva, H., Andrade, E., Araújo, D., & Dantas, J. (2022). Análise de Sentimentos de Tweets Relacionados ao SUS Antes e Durante a Pandemia do COVID-19. *IEEE Latin America Transactions, 20(1),* 6-13.

[14] Tariq, H. I., Sohail, A., Aslam, U., & Batcha, N. K. (2019). Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA). *Journal of Computational and Theoretical Nanoscience*, *16*(8), 3489-3503. https://doi.org/10.1166/jctn.2019.8313

Fuentes de financiamiento:

Propia.

Conflictos de interés:

Los autores declaran no tener conflictos de interés.