



Universidad  
Nacional  
de Loja

## Universidad Nacional de Loja

### Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables

#### Carrera de Ingeniería en Sistemas

**Análisis de Sentimientos en Twitter para la Identificación de Depresión en  
Tiempos de COVID-19 en Ecuador**

**Sentiment Analysis on Twitter for the Identification of Depression in Times of  
COVID-19 in Ecuador**

**Trabajo de Titulación previo a  
la obtención del Título de  
Ingeniero en Sistemas**

**AUTOR:**

Byron Stalin Montaña Beltran

**DIRECTOR:**

Ing. Luis Antonio Chamba Eras, PhD.

Loja – Ecuador

2023

## Certificación

Loja, 23 de mayo de 2023

Ing. Luis Antonio Chamba Eras, PhD.  
**DIRECTOR DEL TRABAJO DE TITULACIÓN**

### **CERTIFICO:**

Que he revisado y orientado todo el proceso de elaboración del Trabajo de Titulación denominado: **Análisis de sentimientos en Twitter para la identificación de depresión en tiempos de COVID-19 en Ecuador**, previo a la obtención del título de **Ingeniero en Sistemas**, de la autoría del estudiante **Byron Stalin Montaña Beltran**, con cédula de identidad **Nro. 1105970014**, una vez que el trabajo cumple con todos los requisitos exigidos por la Universidad Nacional de Loja para el efecto, autorizo la presentación para la respectiva sustentación y defensa.

Ing. Luis Antonio Chamba Eras, PhD.  
**DIRECTOR DEL TRABAJO DE TITULACIÓN**

## **Autoría**

Yo, **Byron Stalin Montaña Beltran**, declaro ser autor del presente Trabajo de Titulación y eximo expresamente a la Universidad Nacional de Loja y a sus representantes jurídicos, de posibles reclamos o acciones legales, por el contenido del mismo. Adicionalmente acepto y autorizo a la Universidad Nacional de Loja la publicación de mi Trabajo de Titulación, en el Repositorio Digital Institucional - Biblioteca Virtual.

**Firma:** .....

**Cédula de Identidad:** 1105970014

**Fecha:** 31 de agosto del 2023

**Correo electrónico:** byron.montano@unl.edu.ec

**Teléfono:** 0959963333

**Carta de autorización por parte del autor, para consulta, reproducción parcial y/o total, publicación electrónica del texto completo, del Trabajo de Titulación**

Yo, **Byron Stalin Montaña Beltran** declaro ser autor del Trabajo de Titulación denominado: **Análisis de Sentimientos en Twitter para la Identificación de Depresión en Tiempos de COVID-19 en Ecuador**, como requisito para optar por el título de **Ingeniero en Sistemas**, autorizo al sistema bibliotecario de la Universidad Nacional de Loja para que, con fines académicos, muestre la producción intelectual de la universidad, a través de la visibilidad de su contenido en el Repositorio Digital Institucional.

Los usuarios pueden consultar el contenido de este trabajo en el Repositorio Institucional, en las redes de información del país y del exterior con las cuales tenga convenio la Universidad.

La Universidad Nacional de Loja, no se responsabiliza por el plagio o copia del Trabajo de Titulación que realice un tercero.

Para constancia de esta autorización, en la ciudad de Loja, a los treinta y un días del mes de agosto del dos mil veinte y tres.

**Firma:**

**Autor:** Byron Stalin Montaña Beltran

**Cédula de identidad:** 1105970014

**Dirección:** La Banda alto

**Correo electrónico:** byron.montano@unl.edu.ec

**Teléfono:** 0959963333

**DATOS COMPLEMENTARIOS:**

**Director de Trabajo de Titulación:** Ing. Luis Antonio Chamba Eras, PhD

## **Dedicatoria**

Dedico este Trabajo de Titulación de manera especial a mi madre, quien siempre ha sido mi apoyo incondicional en cada paso de mi vida, por enseñarme la importancia de la educación y por ser mi mayor motivación para seguir adelante. También dedico este logro a mis hermanos, quienes han sido mi sostén emocional y mis mejores amigos en todo momento, su aliento y confianza en mí me han dado la fortaleza para perseverar y alcanzar esta meta.

Finalmente, dedico este Trabajo de Titulación a todas las personas que han impartido conocimiento sobre mí, me han brindado parte de su tiempo, y me han impulsado a seguir adelante.

***Byron Stalin Montaña Beltran***

## **Agradecimiento**

Agradecer a Dios por darme la fortaleza necesaria para superar los desafíos y no rendirme en cada etapa y propósito de mi vida. Además, agradezco la presencia y el apoyo constante de mis familiares y amigos, quienes han contribuido significativamente a mi desarrollo personal en cada paso del camino.

Al Ing. Luis Chamba, Director de mi Trabajo de Titulación, por su invaluable apoyo, orientación y por brindarme sus valiosos consejos, sugerencias y recomendaciones que me permitieron culminar este proyecto de manera exitosa

Así mismo, me gustaría expresar mi sincero agradecimiento a la Universidad Nacional de Loja, y en particular, a la Carrera de Ingeniería en Sistemas, por brindarme una excelente formación académica a lo largo de todos estos años. A cada uno de los docentes que me impartieron conocimientos, ya que su dedicación y vocación de enseñanza fueron clave para guiarme en mi aprendizaje.

A todas y cada una de las personas que me brindaron su ayuda en todo este proceso, a todos muchas gracias.

***Byron Stalin Montaña Beltran***

## Índice de contenidos

<b>Portada</b> .....	<b>I</b>
<b>Certificación</b> .....	<b>ii</b>
<b>Autoría</b> .....	<b>iii</b>
<b>Carta de autorización</b> .....	<b>iv</b>
<b>Dedicatoria</b> .....	<b>v</b>
<b>Agradecimiento</b> .....	<b>vi</b>
<b>Índice de contenidos</b> .....	<b>vii</b>
<b>Índice de tablas:</b> .....	<b>x</b>
<b>Índice de figuras:</b> .....	<b>xi</b>
<b>Índice de anexos:</b> .....	<b>xii</b>
<b>1. Título</b> .....	<b>1</b>
<b>2. Resumen</b> .....	<b>2</b>
2.1. Abstract .....	3
<b>3. Introducción</b> .....	<b>4</b>
<b>4. Marco teórico</b> .....	<b>6</b>
4.1. Salud mental.....	6
4.1.1. COVID-19 y salud mental .....	6
4.1.2. Depresión .....	7
4.2. Redes sociales .....	7
4.2.1. Twitter.....	8
4.3. Inteligencia artificial (IA).....	9
4.3.1. Minería de texto .....	9
4.3.1.1. Minería de texto vs Minería de datos .....	10
4.3.1.2. Procesamiento de Lenguaje Natural .....	11
4.3.1.3. Análisis de sentimientos .....	11
4.3.2. Machine Learning .....	13
4.3.2.1. Algoritmos de clasificación.....	13
4.3.2.2. Desequilibrio de clases.....	15
4.3.2.3. Validación cruzada .....	17
4.3.2.4. Métricas de evaluación .....	18
4.4. Herramientas para análisis de datos y minería de texto.....	20
4.4.1. Python .....	20
4.4.2. Jupyter Notebook.....	20
4.4.3. Twint.....	20
4.5. Trabajos relacionados.....	21
<b>5. Metodología</b> .....	<b>23</b>

5.1.	Área de estudio.....	23
5.2.	Procedimiento.....	23
5.3.	Recursos .....	24
5.3.1.	Recursos científicos.....	24
5.3.1.1.	Método científico.....	24
5.3.1.2.	Método analítico .....	25
5.3.2.	Recursos técnicos .....	25
5.3.2.1.	Entrevista.....	25
5.3.2.2.	Encuesta.....	26
5.3.2.3.	Metodología KDT.....	26
5.4.	Participantes.....	27
<b>6.</b>	<b>Resultados.....</b>	<b>29</b>
6.1.	Objetivo 1: Construir un conjunto de datos a partir de las publicaciones de Twitter. 29	
6.1.1.	Tarea: Realizar una revisión de literatura sobre análisis de sentimientos en Twitter. 29	
6.1.1.1.	Análisis de sentimientos en Twitter .....	29
6.1.1.2.	Metodología de análisis de sentimientos.....	30
6.1.1.3.	Técnicas de extracción de características.....	33
6.1.1.4.	Herramientas usadas para el análisis de sentimientos.....	34
6.1.2.	Tarea: Definir intervalo de tiempo para la extracción de datos. ....	38
6.1.3.	Tarea: Recopilar tweets mediante una herramienta de scraping en Twitter. ..	39
6.2.	Objetivo 2: Aplicar el análisis de sentimientos mediante una técnica basada en Machine Learning.....	42
6.2.1.	Tarea: Realizar el Preprocesamiento de los datos obtenidos en la fase anterior. 42	
6.2.2.	Tarea: Extracción de características .....	50
6.2.2.1.	TF-IDF para unigramas.....	52
6.2.2.2.	TF-IDF para bigramas.....	54
6.2.2.3.	TF-IDF para trigramas .....	57
6.2.3.	Tarea: Detección de sentimiento.....	59
6.2.3.1.	Modelo de Máquinas de Soporte Vectorial (SVM).....	60
6.2.3.2.	Modelo de Random Forest.....	67
6.2.3.3.	Modelo de Naive Bayes .....	76
6.3.	Objetivo 3: Interpretar los resultados obtenidos en el análisis de sentimientos .....	84
6.3.1.	Tarea: Evaluar el desempeño de los algoritmos mediante métricas de precisión, accuracy, recall y F1 Score.....	85
6.3.2.	Tarea: Predecir contenido depresivo con tweets prepandemia .....	87



6.3.3.	Tarea: Realizar análisis univariado y bivariado de los datos para representar y comparar la cantidad de tweets depresivos. ....	88
<b>7.</b>	<b>Discusión</b> .....	<b>94</b>
7.1.	Objetivo 1: Construir un conjunto de datos a partir de las publicaciones de Twitter. 94	
7.2.	Objetivo 2: Aplicar el análisis de sentimientos mediante una técnica basada en Machine Learning.....	95
7.3.	Objetivo 3: Interpretar los resultados obtenidos en el análisis de sentimientos. ....	97
7.4.	Valoración técnica, económica, ambiental y social.....	99
7.4.1.	Valoración técnica .....	99
7.4.2.	Valoración económica.....	99
7.4.3.	Valoración ambiental .....	100
7.4.4.	Valoración social.....	100
<b>8.</b>	<b>Conclusiones</b> .....	<b>102</b>
<b>9.</b>	<b>Recomendaciones</b> .....	<b>103</b>
<b>10.</b>	<b>Bibliografía</b> .....	<b>104</b>
<b>11.</b>	<b>Anexos</b> .....	<b>110</b>

## Índice de tablas:

<b>Tabla 1.</b> Estructura de la matriz de confusión .....	19
<b>Tabla 2.</b> Comparación fases de Análisis de sentimientos y KDT. ....	32
<b>Tabla 3.</b> Bibliotecas de software de PLN y sus características .....	36
<b>Tabla 4.</b> Palabras clave usadas para la extracción de tweets. ....	39
<b>Tabla 5.</b> Atributos de los tweets recopilados en el dataset inicial. ....	41
<b>Tabla 6.</b> Ejemplos de eliminación de menciones, hashtags y urls en los tweets .....	44
<b>Tabla 7.</b> Ejemplos de limpieza de símbolos, espacios repetidos y conversión de emoticones en los tweets .....	45
<b>Tabla 8.</b> Ejemplos de depuración de los textos que fueron convertidos a partir de los emoticones.....	46
<b>Tabla 9.</b> Ejemplo de eliminación de tildes en un Tweet.....	46
<b>Tabla 10.</b> Ejemplo de conversión de tweet a minúsculas.....	46
<b>Tabla 11.</b> Ejemplo de un tweet tokenizado. ....	47
<b>Tabla 12.</b> Ejemplo de eliminación de stopwords de un tweet.....	48
<b>Tabla 13.</b> Ejemplo de texto aplicado la lematización.....	49
<b>Tabla 14.</b> Recuento de palabras (TF) del conjunto de datos.....	52
<b>Tabla 15.</b> Previa de los valores idf para unigramas del conjunto de datos .....	53
<b>Tabla 16.</b> Valores tf-idf de un tweet del conjunto de datos.....	54
<b>Tabla 17.</b> Recuento de bigramas (TF) del conjunto de datos.....	55
<b>Tabla 18.</b> Previa de los valores idf para bigramas del conjunto de datos .....	55
<b>Tabla 19.</b> Valores tf-idf (bigramas) de un tweet del conjunto de datos .....	56
<b>Tabla 20.</b> Recuento de trigramas (TF) del conjunto de datos.....	57
<b>Tabla 21.</b> Previa de los valores idf para trigramas del conjunto de datos.....	58
<b>Tabla 22.</b> Valores tf-idf (trigramas) de un tweet del conjunto de datos.....	58
<b>Tabla 23.</b> Validación cruzada del modelo SVM con unigramas.....	61
<b>Tabla 24.</b> Reporte de clasificación para el modelo SVM con unigramas.....	61
<b>Tabla 25.</b> Validación cruzada del modelo SVM con bigramas.....	63
<b>Tabla 26.</b> Reporte de clasificación para el modelo SVM con bigramas.....	63
<b>Tabla 27.</b> Validación cruzada del modelo SVM con trigramas. ....	65
<b>Tabla 28.</b> Reporte de clasificación para el modelo SVM con trigramas.....	65
<b>Tabla 29.</b> Validación cruzada del modelo RF con unigramas.....	68
<b>Tabla 30.</b> Reporte de clasificación para el modelo RF con unigramas.....	69
<b>Tabla 31.</b> Validación cruzada del modelo RF con bigramas.....	71
<b>Tabla 32.</b> Reporte de clasificación para el modelo RF con bigramas.....	71
<b>Tabla 33.</b> Validación cruzada del modelo RF con trigramas. ....	74
<b>Tabla 34.</b> Reporte de clasificación para el modelo RF con trigramas.....	74
<b>Tabla 35.</b> Validación cruzada del modelo NB con unigramas. ....	77
<b>Tabla 36.</b> Reporte de clasificación para el modelo NB con unigramas.....	78
<b>Tabla 37.</b> Validación cruzada del modelo NB con bigramas. ....	79
<b>Tabla 38.</b> Reporte de clasificación para el modelo NB con bigramas.....	80
<b>Tabla 39.</b> Validación cruzada del modelo NB con trigramas. ....	81
<b>Tabla 40.</b> Reporte de clasificación para el modelo NB con trigramas. ....	82
<b>Tabla 41.</b> Comparación de los modelos finales mediante métricas de evaluación. ....	86
<b>Tabla 42.</b> Tweets del 2019 clasificados como depresivos por el modelo RF.....	88
<b>Tabla 43.</b> Atributos del conjunto de datos con tweets depresivos. ....	89
<b>Tabla 44.</b> Obtención del lema de algunas palabras usando Spacy y Stanza. ....	96
<b>Tabla 45.</b> Presupuesto - recursos humanos .....	99
<b>Tabla 46.</b> Presupuesto - recursos HW, SW y TICs .....	100

<b>Tabla 47.</b> Presupuesto total del PTT.....	100
---	-----

## Índice de figuras:

<b>Figura 1.</b> Mapa conceptual del Marco teórico, organizado en cinco categorías vinculadas con el objeto de estudio. ....	6
<b>Figura 2.</b> Distribución de usuarios de Twitter en Ecuador por rango de edad. ....	8
<b>Figura 3.</b> Diagrama de Venn que muestra la intersección de la minería de texto con seis campos relacionados. ....	10
<b>Figura 4.</b> Comparación enfoques de análisis de sentimientos. ....	13
<b>Figura 5.</b> Ejemplo de Máquinas de Soporte Vectorial para dos clases. ....	14
<b>Figura 6.</b> Funcionamiento del algoritmo Bosque aleatorio. ....	15
<b>Figura 7.</b> Submuestreo y sobremuestreo en clases desequilibradas. ....	16
<b>Figura 8.</b> Representación de sobremuestreo usando SMOTE.....	17
<b>Figura 9.</b> Validación cruzada con 5 iteraciones .....	18
<b>Figura 10.</b> Área de estudio (Universidad Nacional de Loja).....	23
<b>Figura 11.</b> Fases de la metodología KDT .....	31
<b>Figura 12.</b> Flujo detallado del análisis de sentimientos planteado. ....	32
<b>Figura 13.</b> Promedio de muertes por Covid-19 en Ecuador.....	38
<b>Figura 14.</b> Mapa de mosaicos H3 cubriendo territorio de Ecuador (azul), delimitado por la línea verde. ....	40
<b>Figura 15.</b> Flujo de trabajo del preprocesamiento de datos. ....	43
<b>Figura 16.</b> Nube de palabras de los tweets aleatorios. ....	49
<b>Figura 17.</b> Nube de palabras de los tweets depresivos .....	49
<b>Figura 18.</b> Tweets clasificados como aleatorios y depresivos.....	51
<b>Figura 19.</b> Conjunto de datos con las clases equilibradas .....	51
<b>Figura 20.</b> División de los datos en entrenamiento y prueba para el modelo SVM con Unigramas .....	60
<b>Figura 21.</b> Matriz de confusión del modelo SVM con unigramas. ....	61
<b>Figura 22.</b> División de los datos en entrenamiento y prueba para el modelo SVM con Bigramas. ....	62
<b>Figura 23.</b> Matriz de confusión del modelo SVM con bigramas. ....	63
<b>Figura 24.</b> División de los datos en entrenamiento y prueba para el modelo SVM con Trigramas .....	64
<b>Figura 25.</b> Matriz de confusión del modelo SVM con trigramas. ....	65
<b>Figura 26.</b> Comparación del rendimiento de los 3 modelos SVM. ....	66
<b>Figura 27.</b> División de los datos en entrenamiento y prueba para el modelo RF con Unigramas. ....	67
<b>Figura 28.</b> Rendimiento del modelo RF con unigramas en distintas cantidades de árboles. ....	68
<b>Figura 29.</b> Matriz de confusión del modelo RF con unigramas. ....	69
<b>Figura 30.</b> División de los datos en entrenamiento y prueba para el modelo RF con Bigramas. ....	70
<b>Figura 31.</b> Rendimiento del modelo RF con bigramas en distintas cantidades de árboles. ....	70
<b>Figura 32.</b> Matriz de confusión del modelo RF con bigramas. ....	72
<b>Figura 33.</b> División de los datos en entrenamiento y prueba para el modelo RF con Trigramas. ....	73
<b>Figura 34.</b> Rendimiento del modelo RF con trigramas en distintas cantidades de árboles. ....	73
<b>Figura 35.</b> Matriz de confusión del modelo RF con trigramas. ....	75
<b>Figura 36.</b> Comparación del rendimiento de los 3 modelos Random Forest. ....	76

<b>Figura 37.</b> División de los datos en entrenamiento y prueba para el modelo NB con Unigramas. ....	77
<b>Figura 38.</b> Matriz de confusión del modelo NB con unigramas. ....	78
<b>Figura 39.</b> División de los datos en entrenamiento y prueba para el modelo NB con Bigramas. ....	79
<b>Figura 40.</b> Matriz de confusión del modelo NB con bigramas. ....	80
<b>Figura 41.</b> División de los datos en entrenamiento y prueba para el modelo NB con Trigramas. ....	81
<b>Figura 42.</b> Matriz de confusión del modelo NB con trigramas. ....	82
<b>Figura 43.</b> Comparación del rendimiento de los 3 modelos Naive Bayes. ....	83
<b>Figura 44.</b> Diagrama de flujo del proceso de entrenamiento de modelos y predicción con los nuevos datos. ....	84
<b>Figura 45.</b> Cantidad de tweets del conjunto de datos de prueba por cada clase. ....	85
<b>Figura 46.</b> Cantidad de tweets predichos por cada modelo. ....	85
<b>Figura 47.</b> Mapa de calor de la distribución de tweets depresivos en Ecuador. ....	89
<b>Figura 48.</b> Distribución de publicaciones depresivas por hora. ....	90
<b>Figura 49.</b> Publicaciones depresivas por meses de los años 2020 y 2021. ....	91
<b>Figura 50.</b> Cantidad de tweets depresivos de los años 2019, 2020 y 2021. ....	92
<b>Figura 51.</b> Variación porcentual en la cantidad de tweets entre 2019 y 2020. ....	93
<b>Figura 52.</b> Comparación del rendimiento del modelo Random Forest con otros modelos en trabajos relacionados. ....	97
<b>Figura 53.</b> Geodatos de Ecuador obtenidos mediante la herramienta geoson.io ....	117
<b>Figura 54.</b> Captura parcial del dataset inicial de los tweets recopilados relacionados a depresión. ....	118
<b>Figura 55.</b> Captura parcial del dataset inicial de los tweets aleatorios recopilados. ....	118

## Índice de anexos:

<b>Anexo 1.</b> Entrevista dirigida a Psicóloga Clínica. ....	110
<b>Anexo 2.</b> Encuesta dirigida al Psicólogo de la Unidad de Bienestar Estudiantil de la UNL. ....	115
<b>Anexo 3.</b> Geodatos del territorio de Ecuador. ....	117
<b>Anexo 4.</b> Dataset recopilado de tweets depresivos y tweets aleatorios. ....	118
<b>Anexo 5.</b> Certificado de traducción. ....	119

## **1. Título**

**Análisis de Sentimientos en Twitter para la Identificación de Depresión en Tiempos de COVID-19 en Ecuador**

## 2. Resumen

La pandemia de COVID-19 cambió drásticamente la forma en que vivían las personas, causando devastaciones importantes en temas como la economía y la salud, esto sumado a las restricciones como el encierro, distanciamiento social, además del aumento de casos de muerte, afectó de forma negativa la vida de las personas. La salud mental se ha convertido en una preocupación creciente debido a la terminación del empleo, la pérdida de ingresos, el estrés familiar y otras incertidumbres que han llevado a desencadenar consecuencias graves como la depresión, que es considerada por la OMS como el principal factor que contribuye a la discapacidad mundial. Las redes sociales como Twitter se han convertido actualmente en un medio universal de comunicación, expresión de opiniones y sentimientos, por lo que han jugado un papel fundamental durante los cierres y cuarentenas. Es por esto que el presente Trabajo de Titulación se centra en desarrollar modelos de aprendizaje automático que permitan identificar patrones lingüísticos obtenidos a partir de publicaciones de Twitter y determinar si un usuario está deprimido o no. Para esto se recolectaron tweets del año 2020 y 2021 en temporada de pandemia utilizando palabras clave a través de la herramienta de scraping Twint, y se etiquetaron como depresivos y aleatorios; además, se realizó una limpieza manual para garantizar la calidad del conjunto de datos recolectado. Se utilizaron los algoritmos de Máquinas de Soporte Vectorial, Random Forest y Naive Bayes aplicados en la representación tf-idf de los datos de texto y se obtuvo la característica de los textos conjuntamente con N-gramas para evaluar los modelos con mejor rendimiento por cada tipo de algoritmo. Se comparó el rendimiento de los mejores modelos aplicando las métricas de evaluación Precision, Recall, F1-score y Accuracy. Los resultados mostraron que el clasificador Random Forest presenta resultados aceptables, registrando una puntuación F1 de 95.4% y una exactitud (accuracy) de 95,6%. Finalmente, el modelo con mejor rendimiento se utilizó para predecir los sentimientos en los tweets sin etiquetar del año 2019, en donde se determinó que desde el año 2019 al año 2020 hubo un incremento porcentual en las publicaciones depresivas de un 31.17%.

**Palabras clave:** salud mental, cuarentena, twint, aprendizaje automático, tf-idf, N-gramas.

## 2.1. Abstract

The COVID-19 pandemic drastically changed the way people lived, wreaking further devastation on issues like the economy and health, coupled with restrictions like lockdown and social distancing. In addition to the increase in death toll, it negatively affected people's lives. Mental health has become a growing concern due to the termination of employment, loss of income, family stress and other uncertainties that have led to serious consequences such as depression, which is considered by the WHO as the main factor that contributes to global disability. Social networks such as Twitter have currently become a universal means of communication and expression of opinions and feelings, which is why they have played a fundamental role during closures and quarantines. For this reason, this Thesis is focused on developing automatic learning models that make it possible to identify linguistic patterns obtained from Twitter posts and determine if a user is depressed or not. For this purpose, tweets from the year 2020 and 2021 were collected during the pandemic, using keywords through the Twint scraping tool, and they were labeled as depressive and random (not depressive); in addition, a manual cleaning was carried out to guarantee the quality of the collected data set. Support Vector Machines, Random Forest and Naive Bayes algorithms were applied to the tf-idf representation of the text data and the characteristic of the texts was obtained together with N-grams to evaluate the models with the best performance for each algorithm type. The performance of the best models was compared by applying the Precision, Recall, F1-score and Accuracy evaluation metrics. The results showed that the Random Forest classifier presents acceptable results, registering an F1 score of 95.4% and an accuracy of 95.6%. Finally, the model with the best performance was used to predict the feelings in the unlabeled tweets of the year 2019, and it was possible to determine that from the year 2019 to the year 2020 there was a percentage increase in depressive publications of 31.17%.

**Keywords:** *mental health, quarantine, twint, machine learning, tf-idf, N-grams.*

### 3. Introducción

La depresión es considerada como la principal causa de discapacidad en todo el mundo y puede afectar a cualquier persona en algún momento de la vida y traer consigo consecuencias devastadoras [1]. Las personas con esta condición suelen utilizar las plataformas de redes sociales como Twitter para expresar sus sentimientos y pensamientos de manera un poco más activa [2], ya que estas redes sociales permiten expresar sus opiniones de un tema en específico, sin censura y con mayor libertad. Según la Organización Mundial de la Salud (OMS), aproximadamente 280 millones de personas padecen depresión en todo el mundo [1] [3], lo que equivale a un 4.4% de la población mundial, además ha clasificado a la depresión como el principal factor que contribuye a la discapacidad mundial (más del 7.5% de todos los años vividos con discapacidad en el 2015) y en la sexta posición se encuentran los trastornos de ansiedad [4]. En el Ecuador, durante el año 2020 se realizó un estudio para determinar los efectos de la salud mental y los resultados indican que un número preocupante de personas informó niveles severos o extremadamente severos de depresión (10.3%), ansiedad (19.4%) y estrés (13.5%) [5]. Además, es importante agregar que, tras la declaración del brote de COVID-19 como una preocupación pública internacional por la OMS el 30 de enero de 2020, seguido por restricciones impuestas por el gobierno al movimiento (encierro), el distanciamiento social y el aumento de casos de muerte amenazó no solo la salud física de las personas, sino también afectó su salud mental, especialmente en términos de emociones [6].

En base a esto, las técnicas de inteligencia artificial como machine learning y procesamiento de lenguaje natural (PLN) pueden proporcionar una alternativa para mitigar este problema, ya que ayuda a las computadoras a comprender, interpretar y manipular el lenguaje humano. Y con la cantidad de información existente en las redes sociales como Twitter, ésta puede ser analizada mediante el uso de éstas técnicas de una forma más óptima y rápida que cualquier usuario normal podría hacerlo. Una de estas técnicas es el análisis de sentimientos que utiliza el lenguaje natural y la lingüística computacional para extraer sistemáticamente emociones, sentimientos, opiniones, es decir, la información subjetiva en una pieza de datos textuales [3].

Por lo expuesto, se planteó el desarrollo de un análisis de sentimientos basado en machine learning que permita identificar si una publicación en Twitter tiene contenido depresivo, lo cual permitirá conocer si desde que inició la pandemia de COVID-19, específicamente durante los dos primeros años, hubo un impacto en la salud mental de los usuarios de Twitter en Ecuador mediante sus publicaciones.

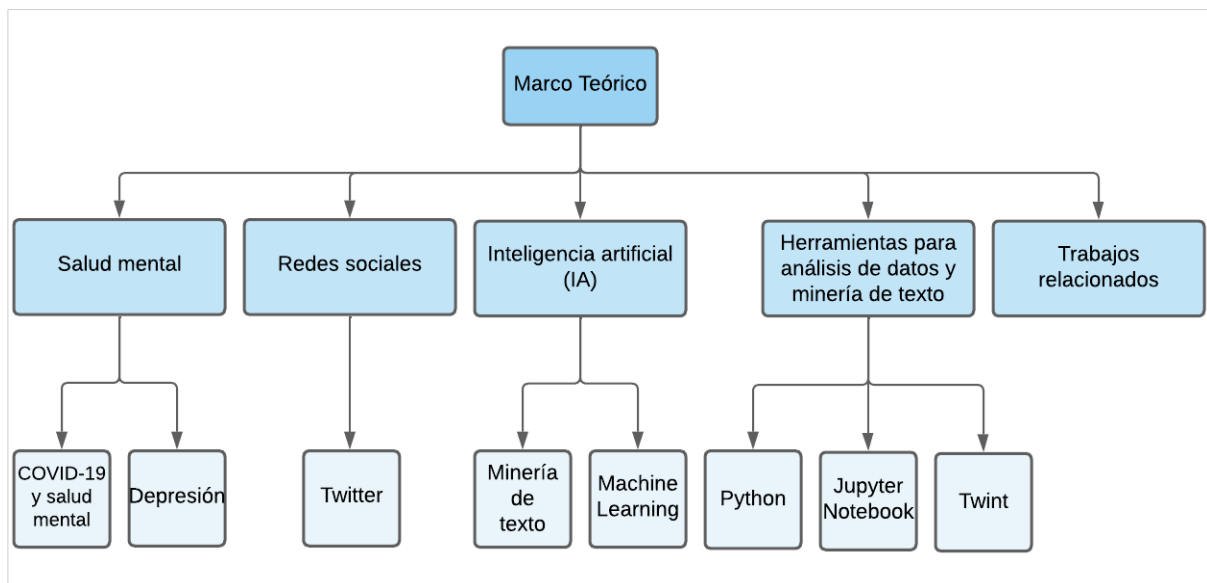
Bajo este contexto, el Trabajo de Titulación (TT) tiene como objetivo principal Identificar publicaciones con contenido depresivo en tiempos de COVID-19 en Ecuador mediante el



análisis de sentimientos en Twitter, por tal motivo, el presente informe se encuentra estructurado de la siguiente manera: en el Marco teórico se detallan los conceptos base que permitieron adquirir los conocimientos necesarios para el cumplimiento de los objetivos; la siguiente sección es la Metodología en donde se muestra el área de estudio y el procedimiento para la realización del trabajo, así como los recursos empleados y los participantes que intervinieron en el desarrollo; en los Resultados se presenta toda la evidencia de los objetivos de acuerdo con las fases de la metodología de análisis de sentimientos planteada; en la sección de Discusión se detalla la evaluación de los resultados realizada a partir de la experiencia durante todo el desarrollo de los mismos; finalmente en Conclusiones y Recomendaciones se muestran los resultados relevantes obtenidos y aspectos clave a considerarse en caso de hacer trabajos similares o trabajos futuros derivados del presente TT.

## 4. Marco teórico

En esta sección se presentan conceptos preliminares que constituyen el sustento teórico para el cumplimiento del objeto de estudio. Se inicia en una contextualización general sobre la salud mental y depresión. Se incluye una introducción a la minería de texto y procesamiento de lenguaje natural, se abarcan algunas generalidades sobre el análisis de sentimientos y sus enfoques principales, además, se describen algunas herramientas que fueron usadas para cumplir el trabajo de titulación y finalmente se presentan algunos trabajos relacionados con el objeto de estudio. En la figura 1 se muestra un mapa conceptual que brinda una visión general de los temas abordados en esta sección.



**Figura 1.** Mapa conceptual del Marco teórico, organizado en cinco categorías vinculadas con el objeto de estudio.

### 4.1. Salud mental

La salud mental es un componente integral y esencial de la salud. La Constitución de la OMS dice: “La salud es un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades”. Una importante consecuencia de esta definición es que considera la salud mental como algo más que la ausencia de trastornos o discapacidades mentales [7].

#### 4.1.1. COVID-19 y salud mental

El brote de COVID-19 fue declarado una preocupación pública internacional por la OMS el 30 de enero de 2020 [8], que llevó a la mayoría de países a tomar medidas restrictivas para mitigar la propagación de esta enfermedad. Todas estas restricciones impuestas, como limitación de movimiento (encierro), el distanciamiento social y el aumento de casos de

muerte amenazó no solo la salud física de las personas, sino que también afectó su salud mental, especialmente en términos de emociones [9] [6]. La OMS señala que la pandemia ha provocado un incremento de la demanda de servicios de salud mental [9], además otros estudios realizados sobre las emociones con el brote de COVID citaron un aumento de las emociones negativas y el pesimismo entre las personas [6].

#### **4.1.2. Depresión**

La depresión es un trastorno de salud mental común que afecta aproximadamente a 280 millones de personas en todo el mundo. Se caracteriza por una tristeza persistente y una falta de interés o placer en actividades que previamente eran gratificantes y placenteras. Además, puede alterar el sueño y el apetito, y es frecuente que concorra con cansancio y falta de concentración. La depresión es una causa importante de discapacidad en todo el mundo, e incide considerablemente en la carga de morbilidad. La depresión puede convertirse en un problema de salud serio, especialmente cuando es recurrente y de intensidad moderada a grave. Puede causar gran sufrimiento a la persona afectada y alterar sus actividades laborales, escolares y familiares. En el peor de los casos puede llevar al suicidio<sup>1</sup>.

#### **4.2. Redes sociales**

Las redes sociales son sitios web y paquetes que pueden premeditarse para que los seres humanos compartan fragmentos de contenido de forma rápida, eficaz y en tiempo real. Se puede compartir imágenes, reseñas, actividades, y ha convertido el modo al que vamos y hacemos una actividad [10].

El advenimiento de la adopción masiva de sitios de redes sociales en línea ha provocado un cambio en la forma en que las personas se comunican y comparten conocimientos. En el área de investigación, el análisis de las redes sociales casi ha reemplazado a cualquier herramienta convencional de las ciencias sociales (encuestas, entrevistas, cuestionarios) anunciando así, la ciencia social computacional. El impresionante crecimiento de las redes sociales pone a disposición una escala sin precedentes de datos personales, datos sobre eventos y relaciones sociales, sentimientos y comportamientos públicos que cuando se extraen e interpretan, tienen un valor enorme [11]. Las plataformas de medios sociales predominantes actualmente son "LinkedIn", "Facebook", "Instagram", "Twitter", "Pinterest" y "YouTube" [10].

---

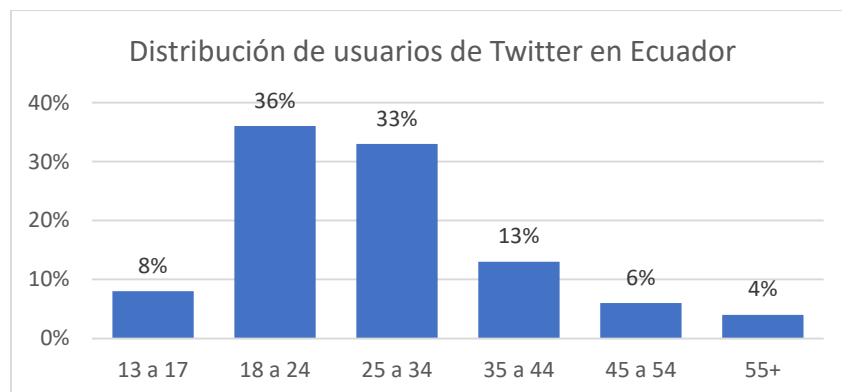
<sup>1</sup>Organización Mundial de la Salud. (2021, septiembre 13). *Depresión*. <https://www.who.int/es/news-room/fact-sheets/detail/depression>

### 4.2.1. Twitter

Twitter es una popular plataforma de microblogging<sup>2</sup> donde los miembros interactúan entre sí y crean mensajes conocidos como "tweets", posee una base de usuarios grande y en constante crecimiento, por lo que esta plataforma proporciona un rico conjunto de datos en forma de mensajes que generalmente son breves actualizaciones de estado de los usuarios de Twitter que deben expresarse en no más de 280 caracteres de longitud [12] [13].

En Twitter, cada día se generan datos que consisten en millones de mensajes cortos y actualizaciones del estado de los usuarios sobre cientos de temas diferentes. La tarea de extraer datos de estos pequeños textos se ha vuelto inmensamente útil para ordenar y clasificar la popularidad de los temas mencionados en las actualizaciones. Hoy en día, Twitter se ha convertido en una de las plataformas más populares para expresar sentimientos y pensamientos en Internet [12].

Los números publicados en los recursos publicitarios de Twitter indican que había 1.45 millones de usuarios en Ecuador a principios de 2022, que representa un 10.5% del total de la población elegible por Twitter, es decir 13.77 millones de personas mayores a 13 años<sup>3</sup>. Además, en el año 2021, la audiencia de Twitter en Ecuador estaba compuesta alrededor de un 8% de usuarios entre 13 a 17 años, un 36% de usuarios entre 18 a 24 años, un 33% de usuarios entre 25 a 34 años, un 13% a usuarios entre 35 a 44 años, y un 6% y 4 % a los usuarios entre 45 a 54 y más de 55 años respectivamente [14]. En la figura 2 se representa de forma gráfica la distribución de los usuarios de Twitter en Ecuador por rango de edad en el año 2021.



**Figura 2.** Distribución de usuarios de Twitter en Ecuador por rango de edad.

<sup>2</sup> El microblogging es una forma de bloguear a pequeña escala, generalmente compuesta de mensajes breves y concisos. [92]

<sup>3</sup> Kemp, S. (2022). Digital 2022: Ecuador. Datareportal. <https://datareportal.com/reports/digital-2022-ecuador>

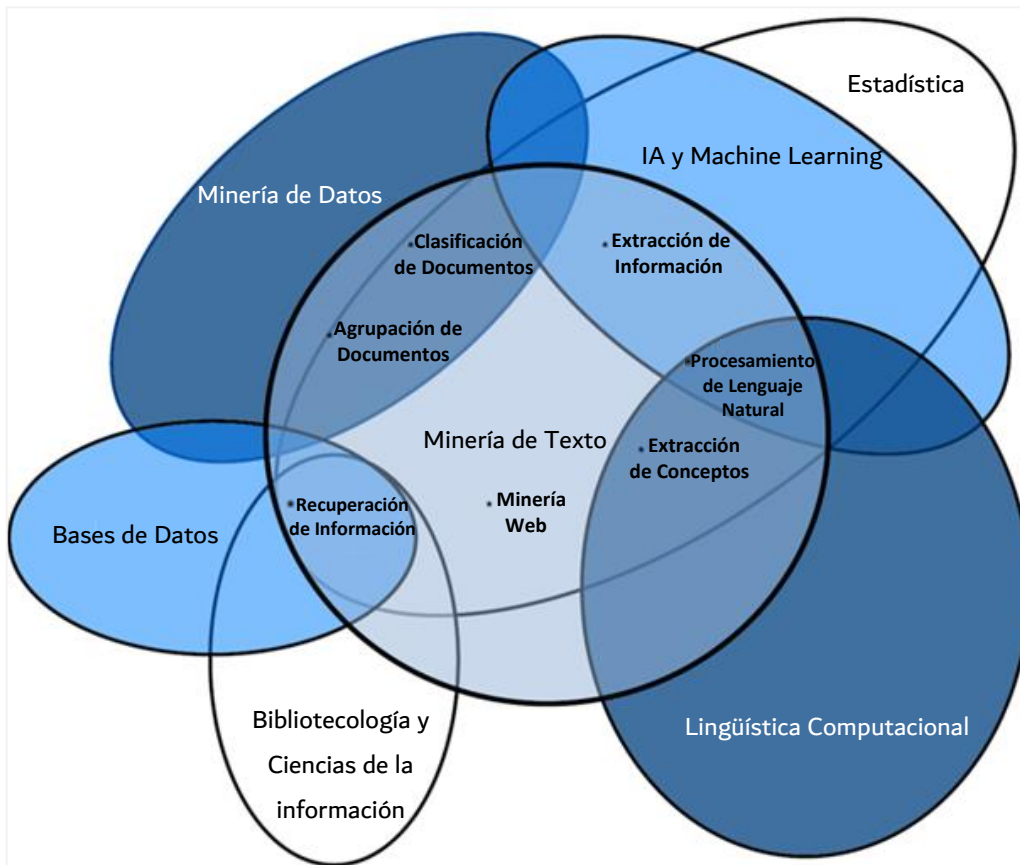
Debido al límite de longitud en los mensajes publicados en twitter, sus publicaciones son más fáciles de analizar porque los autores suelen ir directos al grano. Por lo tanto, a menudo es más fácil lograr una alta precisión del análisis de sentimientos [15]. Para recopilar información de Twitter se puede realizar utilizando su API pública, además, existen diversas aplicaciones y bibliotecas alternativas que permiten recopilar información de Twitter de manera eficiente y personalizada según las necesidades de cada usuario [16].

### **4.3. Inteligencia artificial (IA)**

La IA se refiere a los sistemas que muestran un comportamiento inteligente analizando su entorno y actuando con cierto grado de autonomía para alcanzar objetivos específicos. Esta definición no impone restricciones sobre los métodos que se utilizan para lograr la inteligencia. De hecho, IA es un término general que incluye una amplia gama de tecnologías y aplicaciones que tienen poco más en común que su aparente inteligencia, una cualidad que permanece abierta a la interpretación [17].

#### **4.3.1. Minería de texto**

La minería de textos se define como el proceso de extraer el conocimiento implícito de los datos textuales [18], es decir, describe un conjunto de técnicas lingüísticas, estadísticas y de aprendizaje automático que se refieren generalmente al proceso de extraer información y conocimiento interesante y no trivial a partir de texto no estructurado. Se trata de técnicas que ayudan a modelar y estructurar el contenido informativo de las fuentes textuales para la inteligencia empresarial, el análisis exploratorio de datos, la búsqueda o la investigación [19]. En [20] se afirma que la minería de textos se ha convertido en uno de los campos de moda que se ha incorporado en varias áreas de investigación como la lingüística computacional, la recuperación de información y la minería de datos. En la figura 3 se presenta la interrelación de la minería de texto con distintos campos relacionados [21].



**Figura 3.** Diagrama de Venn que muestra la intersección de la minería de texto con seis campos relacionados.

El objetivo principal de la minería de texto es convertir el texto en datos para su análisis, mediante aplicaciones de procesamiento del lenguaje natural y métodos analíticos, y cuando se obtienen estos datos para su análisis, se puede crear un método de procesamiento de estos datos de la manera que se necesite, o ya se han creado bibliotecas para ello [19]. Entre las aplicaciones más significativas que abordan problemas importantes de minería de texto se encuentran el análisis de sentimientos, detección de fraude, filtrado de correo no deseado, entre otros [22].

#### 4.3.1.1. Minería de texto vs Minería de datos

La minería de texto generalmente se considera como un subdominio de la minería de datos, un campo relacionado con encontrar patrones interesantes en las bases de datos, pero en la configuración del texto. De la misma manera que la minería de datos es parte del proceso global de descubrimiento de conocimiento en bases de datos, la minería de texto podría verse como parte del paradigma de descubrimiento de conocimiento en textos (KDT) [23].

La minería de datos intenta descubrir patrones interesantes a partir de bases de datos masivas. La Minería de texto es el procedimiento de extraer datos e información interesantes y significativos de un texto no estructurado. Es un campo interdisciplinario relativamente

nuevo que se interrelaciona con otros campos como la recuperación de datos, la minería de información, el aprendizaje automático, las estadísticas y la lingüística computacional [24].

En definitiva, la minería de texto es mucho más compleja que la minería de datos porque contiene patrones de datos irregulares y no estructurados, mientras que la minería de datos por lo general trata con conjuntos de datos estructurados [20].

#### **4.3.1.2. Procesamiento de Lenguaje Natural**

El procesamiento del lenguaje natural (PLN) es una rama de la inteligencia artificial que ayuda a las computadoras a comprender, interpretar y manipular el lenguaje humano. El PLN se basa en muchas disciplinas, incluidas la informática y la lingüística computacional en su búsqueda por llenar el vacío entre la comunicación humana y la comprensión informática [25]. El objetivo principal del PLN es reconocer, clasificar o extraer la información a nivel sintáctico y semántico mediante el aprovechamiento de una gran cantidad de computación y datos [26].

#### **4.3.1.3. Análisis de sentimientos**

El análisis de sentimientos o minería de opiniones se refiere al uso del procesamiento del lenguaje natural y la lingüística computacional para extraer sistemáticamente emociones, sentimientos, opiniones, es decir, la información subjetiva en una pieza de datos textuales. La minería de opiniones ha encontrado su uso principalmente dentro de la investigación de mercado, lo que permite a una empresa comprender el sentimiento con respecto a sus productos y servicios. No solo permite el seguimiento de opiniones sino también de los gustos y disgustos de las personas en general [3]. La tarea del análisis de sentimientos generalmente implica tomar un fragmento de texto, ya sea una oración, un comentario o un documento completo y devolver una "puntuación" que mide qué tan positivo o negativo es el texto [25].

##### **4.3.1.3.1. Niveles de análisis de sentimientos**

El análisis de sentimiento se clasifica principalmente en tres niveles diferentes [27],[28],[15]:

- **Nivel de documento:** Se trata de etiquetar documentos individuales con su sentimiento. Todo el documento se clasifica en clase positiva o negativa.
- **Nivel de oración:** El análisis de sentimiento a nivel de oración trata de etiquetar oraciones individuales con sus respectivas polaridades de sentimiento. Clasifica la oración en clase positiva, negativa o neutral. Los textos breves, como el contenido de las redes sociales, se analizan mejor con un análisis de sentimiento a nivel de oración, ya que generalmente consisten en una sola oración.

- **Nivel de aspecto:** Se trata de etiquetar cada palabra con su sentimiento y también identificar la entidad hacia la que se dirige el sentimiento. Se refiere a la identificación y extracción de características del producto de los datos de origen.

#### 4.3.1.3.2. Enfoques para el análisis de sentimientos

Según [27], existen principalmente dos técnicas para el análisis de sentimientos de los datos de Twitter:

- **Enfoque de aprendizaje automático**

El enfoque de aprendizaje automático (machine learning approach) abarca métodos de aprendizaje supervisados y no supervisados. Los métodos supervisados se basan en el uso de conjuntos de datos etiquetados a través de los cuales se crea un modelo para clasificar los datos de entrada no etiquetados. Los métodos no supervisados se utilizan cuando hay conjuntos de datos sin etiquetar para usar en la fase de entrenamiento, por lo que es necesario emplear algoritmos de agrupamiento para etiquetar los datos [29].

Se han formulado varias técnicas de aprendizaje automático para clasificar los tweets en clases. Las técnicas de aprendizaje automático como Naive Bayes (NB), máxima entropía (ME) y máquinas de soporte vectorial (SVM) han logrado un gran éxito en el análisis de sentimientos [27].

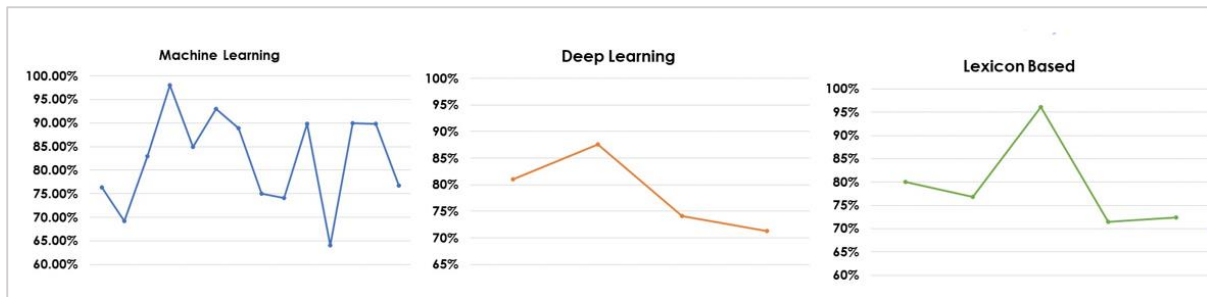
- **Enfoque basado en léxico**

El método basado en léxico (lexicon-based approach), que a menudo se considera enfoque basado en reglas (rule-based approach) [30], utiliza un diccionario de sentimientos con palabras de opinión y las compara con los datos para determinar la polaridad. Asignan puntajes de sentimiento a las palabras de opinión que describen cuán positivas, negativas y objetivas son las palabras contenidas en el diccionario [27].

Los enfoques basados en el léxico se exploran menos en análisis de sentimientos de Twitter comparado con los métodos de aprendizaje automático. La razón principal es la singularidad del texto en Twitter que no solo contiene una gran cantidad de peculiaridades textuales y expresiones coloquiales, sino que también tiene una naturaleza dinámica con nuevas expresiones y hashtags que surgen de vez en cuando [31].

En [32] se realizó una revisión sobre los enfoques de análisis de sentimiento de Twitter, y determinaron que los trabajos basados en el aprendizaje automático muestran la mayor precisión (98%), entre todos los trabajos estudiados, y los trabajos basados en Deep learning y Léxico se encuentran con un rendimiento satisfactorio, con un máximo de 87,62% y 96,11% respectivamente. Esta comparación se muestra en la figura 4 [32].





**Figura 4.** Comparación enfoques de análisis de sentimientos.

La figura 4 muestra que, aunque los trabajos realizados con machine learning presentan una mayor variación en su precisión, también producen resultados más precisos en general. Es importante destacar que estos resultados corresponden a trabajos de análisis de sentimientos generales, es decir, no se seleccionaron trabajos específicos basados en eventos o emociones particulares.

#### 4.3.2. Machine Learning

El aprendizaje automático (machine learning) es una rama en evolución de los algoritmos computacionales que están diseñados para emular la inteligencia humana aprendiendo del entorno circundante. Son considerados el caballo de batalla en la nueva era de los llamados big data. Las técnicas basadas en el aprendizaje automático se han aplicado con éxito en diversos campos que van desde el reconocimiento de patrones, la visión por computadora, la ingeniería de naves espaciales, las finanzas, el entretenimiento y la biología computacional hasta las aplicaciones biomédicas y médicas [33].

El funcionamiento del aprendizaje automático no está limitado por el tipo de sistema o hardware que se utilice, funciona de manera eficiente en todos los sistemas, además requiere menos tiempo para entrenar datos y prepararse y producir resultados en comparación con el aprendizaje profundo (deep learning), que es básicamente un subconjunto del aprendizaje automático que también se utiliza para implementar técnicas de inteligencia artificial mediante redes neuronales, y que requiere más tiempo para entrenar los datos, necesita máquinas avanzadas para funcionar de manera eficiente y necesita usar grandes cantidades de datos para brindar resultados más precisos [34].

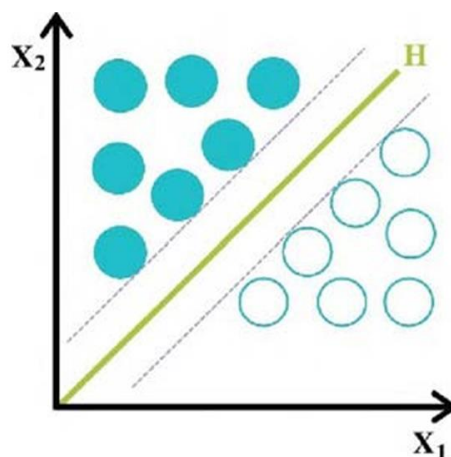
##### 4.3.2.1. Algoritmos de clasificación

###### 4.3.2.1.1. Máquinas de soporte vectorial.

Máquinas de soporte vectorial (Support Vector Machines - SVM) es un algoritmo de clasificación principalmente binaria que se usa típicamente cuando los datos a analizar son limitados. SVM analiza un conjunto de puntos de datos, encuentra un hiperplano que es

esencialmente una línea, que puede separar mejor los datos según su tipo o clase. SVM genera varios de estos hiperplanos, pero se elige al que puede separar los puntos de datos de manera óptima, es decir en el que la distancia normal de cualquiera de los puntos de datos es la mayor. Según la posición del hiperplano, los datos se separan en clases separadas [35].

Un ejemplo general es la clasificación de correos electrónicos en "spam" o "no spam". En la figura 5 se muestra cómo se representarían estas dos clases, donde el círculo celeste representa los correos electrónicos etiquetados como "spam" y el círculo blanco representa los etiquetados como "no spam". SVM puede ser entrenado con un conjunto de datos de correos electrónicos que ya estén etiquetados con estas dos clases y, una vez entrenado, puede predecir si un correo electrónico nuevo y no etiquetado es "spam" o "no spam".



**Figura 5.** Ejemplo de Máquinas de Soporte Vectorial para dos clases.

La clasificación de texto se adapta perfectamente a las SVM debido a la naturaleza escasa del texto, en el que algunas características son irrelevantes, pero tienden a estar correlacionadas entre sí y, en general, se organizan en categorías linealmente separables [36].

#### 4.3.2.1.2. Naive Bayes (NB).

De acuerdo a [37], Naive Bayes es un clasificador probabilístico que utiliza el teorema de Bayes, donde se supone que todas las características (atributos) son independientes entre sí. La ecuación 1 muestra el modelo NB:

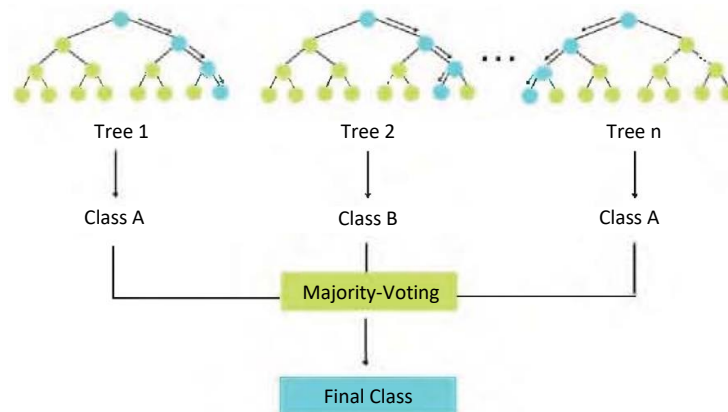
$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (1)$$

donde "P(C|X)" es la probabilidad posterior de la clase dada por el predictor, "P(X)" es la probabilidad previa del predictor, "P(C)" es la probabilidad previa de la clase y "P(X|C)" es la probabilidad de obtener X dado que C es verdadera.

Aunque el clasificador Naive Bayes es simple, es efectivo debido a su robustez frente a características irrelevantes, además funciona bien en dominios con muchas características importantes, y se considera más confiable para la clasificación de texto y el análisis de sentimientos [38].

#### 4.3.2.1.3. Bosque aleatorio.

Un clasificador de bosque aleatorio (Random Forest) es esencialmente un conjunto de árboles de decisión. Cada árbol de decisión arroja una "decisión", es decir, una etiqueta que predice la clase de los datos proporcionados. La clase que aparece con mayor frecuencia se elige como etiqueta de los datos. A medida que aumentamos el número de árboles, también aumenta la precisión de la predicción. El clasificador de bosque aleatorio se usa comúnmente para regresión, clasificación y otras tareas. La figura 6 muestra el funcionamiento básico de un algoritmo de árbol de decisión [35].



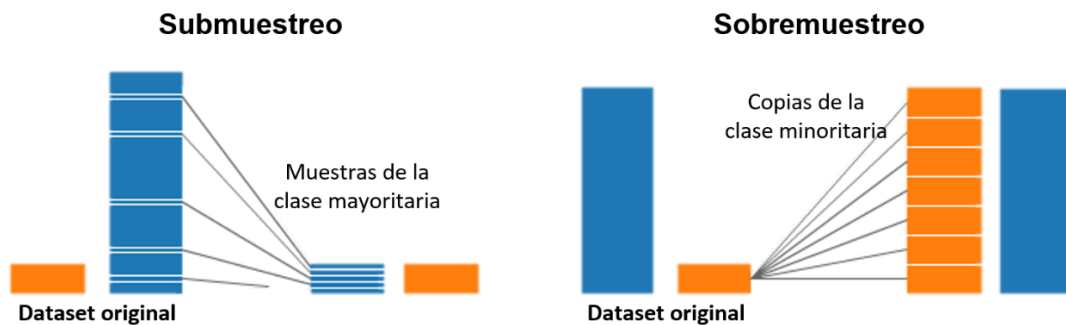
**Figura 6.** Funcionamiento del algoritmo Bosque aleatorio.

Un ejemplo de Random Forest es la clasificación de imágenes. En donde extrae las características de las imágenes y luego las clasifica en distintas categorías. El proceso de clasificación consiste en la creación de varios árboles de decisión independientes como se muestra en la figura 6, cada uno entrenado con una muestra aleatoria de las características de las imágenes, para finalmente realizar la clasificación mediante una votación de los resultados de todos los árboles.

#### 4.3.2.2. Desequilibrio de clases

Al enfrentarse a la situación de crear un modelo de clasificación es habitual que las clases no se encuentran balanceadas, esto puede dar como resultado una precisión bastante alta simplemente prediciendo la clase mayoritaria, pero no logra capturar la clase minoritaria, que suele ser el objetivo de crear el modelo en primer lugar.

Las técnicas de sobremuestreo y submuestreo (ver figura 7) son muy útiles para resolver la distribución de etiquetas de clase desequilibrada para la clasificación multiclase [39].

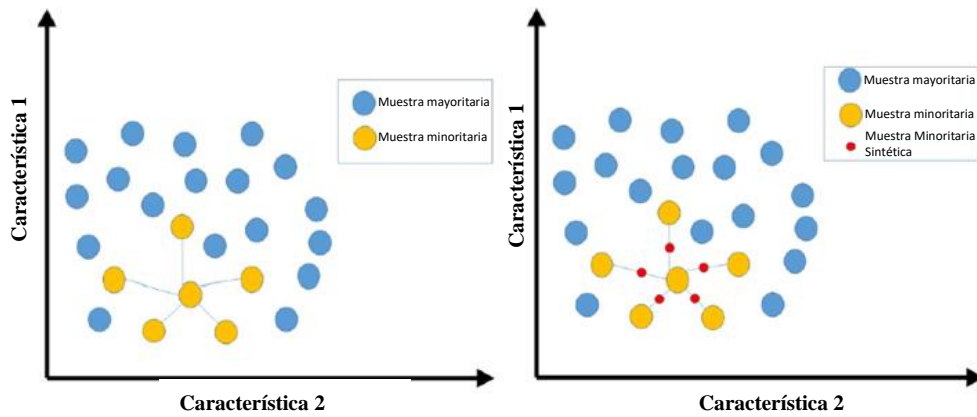


**Figura 7.** Submuestreo y sobremuestreo en clases desequilibradas.

El **sobremuestreo aleatorio** es un método no heurístico que tiene como objetivo equilibrar la distribución de clases a través de la replicación aleatoria de ejemplos de clases minoritarias [40]. Por ejemplo, en un conjunto de datos con dos clases, una clase azul con 1000 puntos y una clase naranja con 100 puntos, se realizarían copias aleatorias de los 100 puntos de la clase naranja hasta tener un total de 1000 puntos para cada clase, lo que resultaría en un conjunto de datos equilibrado con 2000 puntos, tal como se representa en la parte derecha de la figura 7.

El **submuestreo aleatorio** también es un método no heurístico que tiene como objetivo equilibrar la distribución de clases mediante la eliminación aleatoria de ejemplos de clases mayoritarias [40]. Por ejemplo, se podría tomar una muestra aleatoria de 100 puntos de la clase azul y combinarla con los 100 puntos de la clase naranja para obtener un conjunto de datos equilibrado con 200 puntos, tal como se representa en la parte izquierda de la figura 7.

**SMOTE** (técnica de sobremuestreo de minorías sintéticas) es otro método popular para realizar el sobremuestreo. En SMOTE, se crean nuevas instancias basadas en la interpolación entre varias instancias de clase minoritaria que se encuentran juntas. Esto permite que SMOTE opere en el espacio de características en lugar de operar en el espacio de datos [41]. Básicamente, la clase minoritaria se puede sobremuestrear creando casos sintéticos en el espacio de características formado por la instancia y sus K vecinos más cercanos, como se muestra en la figura 8 [42].



**Figura 8.** Representación de sobremuestreo usando SMOTE.

Basado en la figura 8, suponiendo que se tiene un conjunto de datos que contiene 1000 ejemplos de la clase A (círculos celestes) y solo 100 ejemplos de la clase B (círculos naranjas). Dado que la clase B es la minoritaria, SMOTE genera una nueva muestra sintética (círculos rojos) para la clase B interpolando aleatoriamente entre el ejemplo original y uno de sus vecinos. Este proceso se repite hasta que se ha generado un número suficiente de nuevas muestras sintéticas para la clase B, de modo que el conjunto de datos ahora tiene un equilibrio entre ambas clases.

El sobremuestreo y submuestreo aleatorio presentan algunos inconvenientes, el sobremuestreo aleatorio puede aumentar la probabilidad de que se produzca un sobreajuste (overfitting), ya que duplica registros aleatorios de la clase minoritaria. De esta manera un clasificador, por ejemplo, podría construir reglas que aparentemente son precisas, pero que en realidad cubren un ejemplo replicado. Por otro lado, el principal inconveniente del submuestreo aleatorio es que este método puede descartar datos potencialmente importantes de las muestras de la clase mayoritaria [40].

En Python se puede usar la librería “imblearn”, dado que está diseñada específicamente para manejar conjuntos de datos desequilibrados. Proporciona varios métodos como submuestreo (undersampling), sobremuestreo (oversampling) y SMOTE para manejar y eliminar el desequilibrio de un conjunto de datos [41].

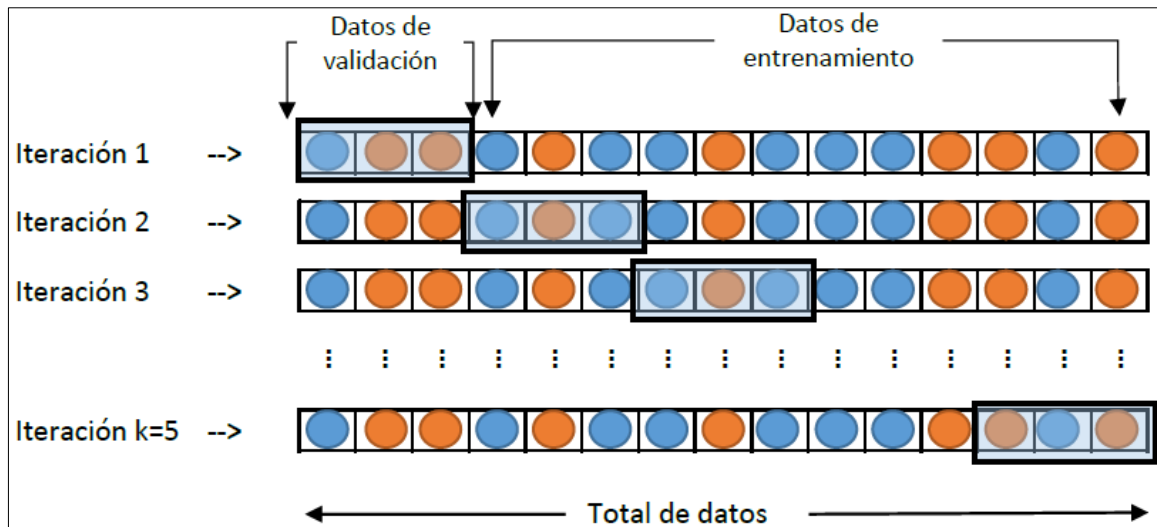
#### **4.3.2.3. Validación cruzada**

La estrategia más típica en el aprendizaje automático es dividir un conjunto de datos en conjuntos de entrenamiento y prueba, el problema con esta estrategia es que no sabemos si una alta precisión de validación indica un buen modelo. Por ejemplo, si realizamos entrenamiento en el 80% del conjunto de datos y el 20 % restante para fines de prueba, es posible que el 20% de los datos tenga información importante que estamos dejando fuera del entrenamiento en el modelo. Ahí es donde entra la validación cruzada.

- **Validación cruzada de K-Fold**

En este método, se divide los datos de entrenamiento en  $k$  subconjuntos (pliegues), se realiza el entrenamiento en todos los subconjuntos dejando un subconjunto ( $k-1$ ) para la evaluación del modelo entrenado. Luego se va iterando  $k$  veces con una división única de entrenamiento:validación [43].

En la figura 9 se observa un ejemplo de cómo sería el proceso de validación cruzada de K-fold con 5 iteraciones.



**Figura 9.** Validación cruzada con 5 iteraciones

Como se muestra en la figura 9, el conjunto de datos se divide en 5 partes iguales y se utiliza cada parte una vez para evaluar el modelo mientras se entrena con las 4 partes restantes, así sucesivamente hasta terminar las 5 iteraciones. Esto ayuda a evaluar el rendimiento del modelo de forma más robusta y evitar el sobreajuste.

En definitiva, la validación cruzada permite garantizar que todas las observaciones del conjunto de datos original tengan la oportunidad de aparecer en la serie de entrenamiento y en la serie de prueba.

#### **4.3.2.4. Métricas de evaluación**

##### **4.3.2.4.1. Matriz de confusión**

La matriz de confusión es una medida muy popular utilizada para resolver problemas de clasificación. Se puede aplicar tanto a la clasificación binaria como a problemas de clasificación multiclase [41]. En la tabla 1 se muestra la estructura de la matriz de confusión para una clasificación binaria.

**Tabla 1.** Estructura de la matriz de confusión

	Clasificado como Positivo	Clasificado como Negativo
Son positivos	<b>TP</b>	<b>FN</b>
Son negativos	<b>FP</b>	<b>TN</b>

Las matrices de confusión representan recuentos de valores predichos y reales. La salida "TN" significa True Negative, que muestra el número de ejemplos negativos clasificados con precisión. De manera similar, "TP" significa True Positive, que indica la cantidad de ejemplos positivos clasificados con precisión. El término "FP" muestra el valor de falso positivo, es decir, el número de ejemplos negativos reales clasificados como positivos; y "FN" significa un valor de falso negativo que es el número de ejemplos positivos reales clasificados como negativos [41].

Según [31], las métricas de evaluación utilizadas con más frecuencia son la exactitud, la precisión, la exhaustividad y la puntuación F, adoptadas de los problemas de clasificación tradicionales, estas métricas se muestran a continuación en base a la matriz de confusión (ver tabla 1).

- **Exactitud (Accuracy):** accuracy es la métrica de evaluación más utilizada y mide la frecuencia con la que el método que se evalúa hizo la predicción correcta. Se calcula como la suma de las predicciones verdaderas dividida por el número total de predicciones.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

- **Precisión (Precision):** la precisión representa la exactitud del método y se calcula como la proporción de instancias que se pronosticaron como positivas y que lo fueron, dividida por el número total de instancias que se pronosticaron como positivas.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

- **Exhaustividad (Recall):** que también se conoce como sensibilidad, denota la fracción de instancias positivas que se predijo que eran positivas y se calcula como:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

- **Puntuación F (F-score):** por lo general, calcular la exhaustividad y la precisión no es suficiente. Una combinación de los dos es más apropiada para evaluar el desempeño de los métodos. La puntuación F es la métrica que combina exhaustividad y precisión. Esta métrica también se conoce como puntuación F armónica, puntuación F1 o precisión de medida F y se calcula como:

$$F - score = 2 * \frac{precision * recall}{precision + recall} \quad (5)$$

#### **4.4. Herramientas para análisis de datos y minería de texto.**

Existen una gran variedad de herramientas que son ampliamente utilizadas en proyectos de análisis de datos y minería de texto. En este TT en particular, se utilizaron Python, Jupyter Notebook y Twint como herramientas principales para su desarrollo.

##### **4.4.1. Python**

Python se ha convertido en el lenguaje de programación de facto para el análisis de datos y el aprendizaje automático. Es un lenguaje de secuencias de comandos que se puede usar de forma interactiva y no requiere la compilación del código fuente en un ejecutable para ejecutarse, lo que facilita la transferencia de un programa Python entre computadoras y sistemas operativos [44]. Una ventaja de Python es que hay muchas bibliotecas de código abierto disponibles gratuitamente para su uso [45], esto facilita la creación de una solución propia de análisis de sentimientos.

##### **4.4.2. Jupyter Notebook**

Jupyter Notebook ([jupyter.org](http://jupyter.org)) es una potente herramienta de código abierto basada en navegador para el desarrollo interactivo y la presentación de proyectos de ciencia de datos. Cada cuaderno consta de una colección de celdas ejecutables y cada celda contiene texto formateado usando el lenguaje Markdown o código ejecutable (generalmente Python o R) [46].

##### **4.4.3. Twint**

Twint es una herramienta avanzada de raspado de Twitter (Twitter scraping) escrita en Python que permite raspar tweets de los perfiles de Twitter sin usar la API de Twitter. Twint utiliza los operadores de búsqueda de Twitter que permite recopilar tweets de usuarios específicos, recopilar tweets relacionados con ciertos temas, hashtags y tendencias, u ordenar información confidencial de tweets como correo electrónico y números de teléfono [47]. Twint también realiza consultas especiales a Twitter, lo que le permite rastrear los seguidores de un usuario de Twitter, los Tweets que le han gustado a un usuario y a quién sigue sin ninguna autenticación o API.

Hay ciertas características de Twint que lo hacen más útil y único de otras API de raspado de Twitter, esto es:



- La API de Twitter tiene restricciones para raspar solo los últimos 3200 Tweets [48]. Twint puede recuperar casi todos los Tweets.
- La configuración es realmente rápida ya que no hay problemas para configurar la API de Twitter.
- Se puede usar de forma anónima sin registrarse en Twitter.
- Es gratis, sin limitaciones de precios.
- Brinda opciones fáciles de usar para almacenar tweets raspados en diferentes formatos: CSV, JSON, SQLite y Elasticsearch.

#### 4.5. Trabajos relacionados

En esta sección se encuentran los trabajos relacionados con el objeto de estudio y que fueron obtenidos durante la revisión de literatura.

- En el trabajo llamado *“Identificación de depresión mediante el análisis de sentimientos”* utilizan el análisis de sentimientos para detectar posibles síntomas de depresión por medio de los dispositivos móviles de personas que corran riesgo de sufrir este trastorno mental, para esto obtuvieron sus conversaciones de WhatsApp, que se analizan para detectar posibles emociones y en función del resultado obtenido poder determinar la probabilidad de sufrir depresión [49].
- En *“Sentiment Analysis of the COVID-related r/Depression Posts”* se presenta un análisis de sentimientos de los mensajes relacionados con COVID-19 publicados en r/Depression en la plataforma social Reddit. Para esto clasificaron los datos de texto sin etiquetar en categorías relacionadas con el contenido del mensaje y dado que las publicaciones no están etiquetadas y es imposible leer una gran cantidad de datos de texto y asignarles un tema relevante de forma manual, utilizaron un modelo estadístico de temas no supervisado (topic model) para asignar un sentimiento relevante a cada publicación. Posteriormente, utilizaron modelos de aprendizaje automático para realizar clasificaciones de texto para examinar y evaluar la exactitud de las etiquetas de los temas y poder realizar un análisis de opinión sobre los datos recién etiquetados [50].
- En el estudio llamado *“Depression Detection Using Sentiment Analysis of Tweets”* usan análisis de sentimientos para detectar depresión, para lo cual se clasificaron los tweets con rasgos depresivos entre los tweets extraídos de Twitter. Mediante un modelo intentan maximizar la utilización de todas las características lingüísticas disponibles presentes en los tweets y hace uso de técnicas adecuadas de limpieza y

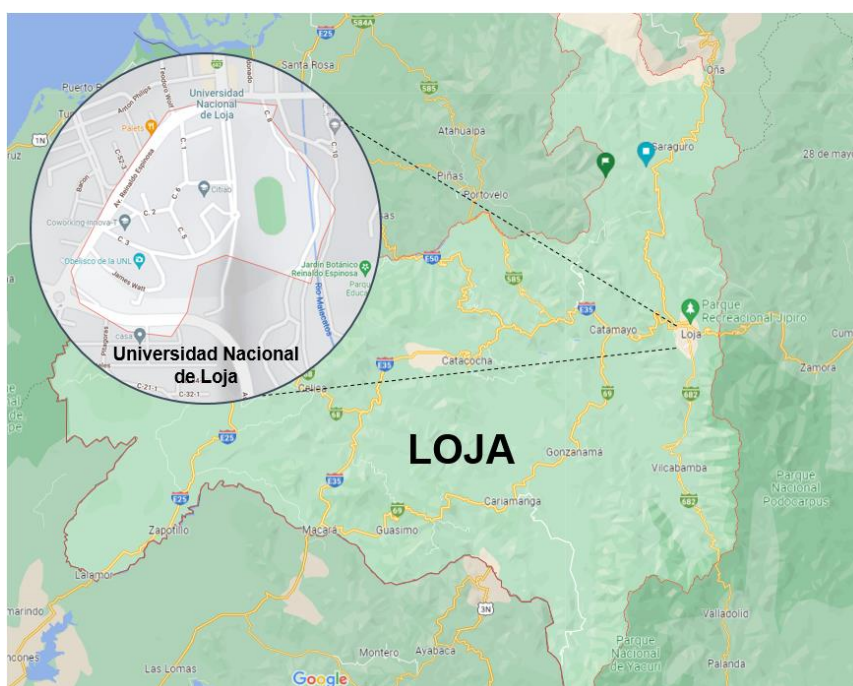
preprocesamiento para un diagnóstico más preciso de la depresión. Utilizaron Vader Analyzer basado en reglas y un modelo híbrido de CNN-LSTM [51].

- “An Initiative to Identify Depression using Sentiment Analysis: A Machine Learning Approach”, es un proyecto en el cual se propone un algoritmo mediante el cual se extraen los tweets de Twitter usando R studio y luego se analizan sus sentimientos, es decir, se otorgan puntuaciones a cada sentimiento mediante el cual se identifica si la persona está deprimida o no. Y mediante los conjuntos de datos de Twitter se evalúan a través del algoritmo propuesto. Si el sentimiento ha puntuado como positivo negativo y natural para las emociones positivas, se obtiene una puntuación, pero si la emoción es negativa como ira, disgusto, etc., no obtiene puntuaciones [52].

## 5. Metodología

### 5.1. Área de estudio

El presente Trabajo de Titulación (TT) se desarrolló en el cantón Loja, en la Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables de la Universidad Nacional de Loja (UNL), en la Carrera de Ingeniería en Sistemas/Computación (ver figura 10).



**Figura 10.** Área de estudio (Universidad Nacional de Loja)

### 5.2. Procedimiento

Para alcanzar el objetivo general del presente TT se usó el siguiente proceso para cada uno de los objetivos específicos:

1. Construir un conjunto de datos a partir de las publicaciones de Twitter, ver sección 6.1.
  - a. Revisión de literatura sobre análisis de sentimientos en Twitter, ver sección 6.1.1.
  - b. Definir intervalo de tiempo para la extracción de datos, ver sección 6.1.2.
  - c. Recopilación de los tweets mediante la herramienta de scraping Twint, ver sección 6.1.3.
2. Aplicar el análisis de sentimientos mediante una técnica basada en Machine Learning, ver sección 6.2.
  - a. Realizar el Preprocesamiento de los datos obtenidos en la fase anterior usando la herramienta de PLN para Python como NLTK, ver sección 6.2.1.

- b. Extracción de características, ver sección 6.2.2.
  - c. Detección de sentimiento, ver sección 6.2.3.
3. Interpretar los resultados obtenidos en el análisis de sentimientos, ver sección 6.3.
- a. Evaluar el desempeño de los algoritmos mediante métricas de precisión, accuracy, recall y F1 Score, ver sección 6.3.1.
  - b. Predecir contenido depresivo con tweets prepandemia, ver sección 6.3.2.
  - c. Realizar un análisis univariado y bivariado de los datos para representar y comparar la cantidad de tweets depresivos, ver sección 6.3.3.

### **5.3. Recursos**

#### **5.3.1. Recursos científicos**

Se tomó en cuenta varios métodos y metodologías para la elaboración del presente TT; los cuales se presentan a continuación:

##### **5.3.1.1. Método científico**

Es un método de investigación usado principalmente en la producción de conocimiento en las ciencias, mediante el cual podemos alcanzar un conocimiento objetivo de la realidad [53]. Las etapas que integran el método científico son: 1) definición del problema, 2) formulación de hipótesis, 3) recopilación y análisis de datos, 4) confirmación o rechazo de hipótesis, 5) resultados, 6) conclusiones.

Este método se empleó desde la elaboración del marco teórico y durante todo el transcurso de los objetivos del proyecto. En lo que respecta a la definición del problema, en el marco teórico se presenta la relación entre la salud mental y covid-19 y como puede desencadenar en un problema de salud como la depresión. En cuanto a la formulación de hipótesis, esto se hizo acorde a las fases de la metodología KDT, específicamente en la fase Minería de texto y construcción de hipótesis en el desarrollo del objetivo 2, en donde en base a la cantidad de tweets recopilados y preprocesados se planteó la hipótesis que a su vez tiene relación con la pregunta de investigación planteada en este proyecto.

La recopilación y análisis de datos se lo realizó en el objetivo 1 mediante la recopilación de dos clases distintas de tweets, una clase para tweets que pueden tener indicios de depresión usando palabras clave y que fue depurado con la ayuda de un especialista en el tema de la salud mental, la otra clase de tweets recopilados fueron tweets aleatorios que no contienen características depresivas; en cuanto al objetivo 3 se realizó el análisis de los mejores modelos para identificar depresión y así determinar cuál de ellos ofrece un mejor rendimiento.

Además, para realizar la confirmación o rechazo de la hipótesis planteada, en el objetivo 3 se extrajeron tweets de época prepandemia, específicamente del año 2019 para realizar una clasificación de los tweets depresivos de ese año utilizando el modelo que tuvo mejor rendimiento en el entrenamiento, obteniendo con esto una comparación con los datos ya obtenidos en temporada de covid-19 y consiguiendo de esta manera los resultados de variabilidad en las publicaciones depresivas en prepandemia (2019) y pandemia (2020 y 2021). Finalmente, las conclusiones se utilizaron para contrastar y presentar los resultados obtenidos en el desarrollo del proyecto.

#### **5.3.1.2. Método analítico**

Según [54], el método analítico consiste en distinguir, conocer y clasificar los distintos elementos que conforman un conocimiento general, y a partir del conocimiento general de una realidad realiza la distinción, conocimiento y clasificación de los distintos elementos esenciales que forman parte de ella y de las interrelaciones que sostienen entre sí.

Este método se empleó para conocer las características de los tweets, ya que mediante la recolección de los tweets relacionados con depresión se pudo tener un conocimiento general sobre las publicaciones que se realizan en twitter relacionado al tema, y mediante la investigación se pudo conocer y filtrar las publicaciones que no estaban relacionadas al objeto de estudio para asegurar de que solo hayan tweets que manifiesten depresión mediante la ayuda de un especialista en el área de salud mental, y en base a esto realizar una adecuada clasificación del sentimiento mediante los algoritmos de machine learning utilizados.

#### **5.3.2. Recursos técnicos**

##### **5.3.2.1. Entrevista**

La entrevista es una técnica que posibilita obtener información acerca de las características de un problema de un informante clave. Dicha información puede ser novedosa o complementaria y ayuda a cuantificar características y la naturaleza del objeto de estudio [55].

Esta técnica se utilizó para realizar una entrevista dirigida a la Dra. Ximena Amaya Valarezo, psicóloga clínica, mediante la cual se obtuvo información relevante que sirvió como sustento y justificación del presente TT, y ya que se realizó a un profesional en cuanto a salud mental se refiere, esta información también contribuyó a resaltar la importancia de los resultados del presente TT. La constancia de la entrevista realizada se encuentra en el Anexo 1.

### **5.3.2.2. Encuesta**

Según [56], las encuestas son investigaciones que proporcionan una visión general, mediante la recogida de información estandarizada de una población específica o una muestra representativa de la misma (sujetos del estudio), por medio de un cuestionario o entrevista. Esta técnica se usó como complemento a la entrevista realizada a la Dra. Ximena Amaya, en este caso se realizó una encuesta dirigida al Dr. Jorge Fernando Jiménez Sánchez, Psicólogo de la Unidad de Bienestar Universitario en la Universidad Nacional de Loja, mediante esta encuesta se obtuvo información para sustentar y justificar el presente TT, y dar solidez a la información obtenida en la entrevista mencionada. La evidencia de la encuesta se presenta en el Anexo 2.

### **5.3.2.3. Metodología KDT**

Para el presente TT se tomó como referencia las fases de la metodología KDT, adaptando sus fases a las metodologías propuestas en el análisis de sentimientos, esta metodología se utilizó durante los 3 objetivos para el desarrollo del análisis de sentimientos de acuerdo a cada una de sus fases (ver sección 6), las cuales son: Selección de datos, preprocesamiento de datos, transformación de datos, minería de texto y construcción de hipótesis, y finalmente la interpretación/evaluación. Estas fases se describen a continuación:

#### **Selección de datos.**

Esta primera fase se la realizó en el objetivo 1, donde se extrajeron tweets relacionados a depresión y tweets aleatorios de todo el territorio de Ecuador, utilizando palabras clave para recolectar los tweets que pueden ser considerados como depresivos, estos datos fueron el objeto de estudio para las fases restantes de la metodología KDT, el proceso de recolección se realizó de forma automática mediante el uso de la herramienta de scraping Twint.

#### **Preprocesamiento de datos.**

Debido a que las publicaciones de Twitter contienen mucho texto informal e incluyen usos idiosincrásicos, es difícil que un algoritmo pueda entender el contexto real de un tweet, es por esto que se realizó una serie de pasos para preprocesar las publicaciones, realizando en primer lugar una limpieza manual para eliminar tweets que no tengan relación al objeto de estudio, para luego eliminar símbolos y caracteres especiales, menciones, conversión a minúsculas, tokenización, eliminación de palabras vacías y lematización. Todo este proceso se realizó al inicio del objetivo 2.

### **Transformación de datos.**

En esta fase se representó todo el texto en números o vectores de números para que puedan ser entrenados con los algoritmos de machine learning, para realizar esto se usó la técnica de tf-idf para convertir el texto en una matriz o vector de características en base a los tweets preprocesados. Todo este proceso se lo realizó para unigramas, bigramas y trigramas, a fin de comparar los resultados con cada uno de ellos en el entrenamiento de los datos.

### **Minería de texto.**

En esta fase los datos se entrenaron mediante los algoritmos Máquinas de Soporte Vectorial, Random Forest y Naive Bayes, cada uno de ellos con unigramas, bigramas y trigramas mediante los vectores de características obtenidos en la fase anterior. El rendimiento de cada uno de los modelos se validó utilizando la técnica de validación cruzada para asegurarnos que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba. Finalmente se comparó el rendimiento de cada uno de los modelos y se guardaron los modelos con mejor rendimiento para usarlos y analizarlos en la fase posterior.

### **Interpretación/evaluación.**

Finalmente, en base a los mejores modelos se generaron gráficas de comparación para determinar cuál de ellos ofrece el mejor rendimiento para identificar contenido depresivo, y en base al mejor modelo se predijeron publicaciones en tiempo prepandemia, específicamente del año 2019. Los resultados de la predicción se compararon con los datos en tiempos de pandemia (2020 y 2021) para determinar la variabilidad en las publicaciones depresivas entre los distintos años.

## **5.4. Participantes**

El presente TT enfocado en la línea de investigación de Sistemas inteligentes, se contó con los siguientes participantes.

- Byron Stalin Montaña Beltran, como estudiante investigador y autor del presente TT, iniciando sus actividades desde el planteamiento del tema del PTT, hasta el desarrollo y finalización de los diferentes objetivos establecidos en el presente TT.
- El Ing. Luis Chamba-Eras, como director del TT, quien supervisó los avances académicos y técnicos desarrollados por el autor del presente TT.
- La Ing. María Del Cisne Ruilova Sánchez, como tutor académico, quien supervisó los avances académicos desarrollados por el autor del presente TT.
- La Dra. Ximena Dennise Amaya Valarezo y el Dr. Jorge Fernando Jiménez Sánchez, que aportaron con su conocimiento como psicólogos para dar sustento en la problematización del presente TT.

- La Dra. Sandra Otero, como psicóloga clínica, ayudó a la identificación y filtración de los tweets depresivos recolectados, lo que permitió el correcto desarrollo y culminación del primer objetivo del presente TT.



## 6. Resultados

En esta sección se detallan los resultados de los objetivos específicos del presente Trabajo de Titulación (TT), por cada objetivo se presentan las actividades y tareas que se realizaron para dar con el cumplimiento de los mismos. En el objetivo 1 se construyó el conjunto de datos utilizando publicaciones que fueron recopiladas en Twitter en tiempos de COVID-19, tomando como referencia una revisión de literatura, que permitió conocer y comprender de mejor manera los aspectos necesarios que fueron usados en cada una de las fases del análisis de sentimientos. En el objetivo 2, se realizó la clasificación y detección del sentimiento usando modelos de machine learning, utilizando las fases de la metodología KDT: selección, preprocesamiento, transformación, minería de texto y construcción de hipótesis, interpretación/evaluación. Por último, en el objetivo 3 se analizaron los mejores modelos mediante métricas de rendimiento, y con el mejor modelo se realizó la predicción con nuevos datos recolectados en prepandemia, esto se complementó con un análisis univariado y bivariado a los conjuntos de datos recopilados y filtrados.

### 6.1. Objetivo 1: Construir un conjunto de datos a partir de las publicaciones de Twitter.

Para la construcción del conjunto de datos, fue necesario en primer lugar realizar una revisión de literatura sobre el análisis de sentimientos en Twitter, lo que permitió conocer de mejor manera la plataforma de recolección de datos, metodología, herramientas, lenguajes de programación y software en general necesario para el análisis de sentimientos. A continuación, se muestra la revisión de literatura y el proceso que se llevó a cabo para obtener los datos de Twitter y construir el conjunto de datos para su posterior uso.

#### 6.1.1. Tarea: Realizar una revisión de literatura sobre análisis de sentimientos en Twitter.

##### 6.1.1.1. Análisis de sentimientos en Twitter

Analizar el sentimiento en Twitter supone asignar a cada mensaje publicado un valor relacionado con la carga emocional que transmite. En relación a esta carga emocional se pueden distinguir algunos tipos de variables [57]:

- **Polaridad:** indica si el mensaje tiene un sentimiento positivo o negativo. En algunos análisis se introduce una tercera categoría para clasificar los mensajes neutros.
- **Intensidad:** proporciona un valor numérico en relación con la intensidad del sentimiento. Se puede distinguir entre una intensidad positiva y una intensidad negativa.

- **Emoción:** clasifica el texto según los distintos tipos de emociones, como puede ser la alegría, la tristeza o la ira.

#### 6.1.1.2. Metodología de análisis de sentimientos

Varios autores plantean metodologías de análisis de sentimientos que varían en 1 o 2 fases. Sin embargo, basado en los estudios [58] [59] [32] [60] [61], una metodología general de análisis de sentimientos en Twitter se puede realizar de acuerdo a las siguientes fases:

- Recolección de datos
- Preprocesamiento
- Extracción de características
- Clasificación de sentimientos
- Detección de sentimiento (polaridad)
- Evaluación y análisis

Estas fases se pueden contrastar con una metodología de minería de texto como la metodología KDT que se muestra a continuación.

#### Descubrimiento de conocimiento en texto (KDT)

Descubrimiento de conocimiento en texto (Knowledge Discovery in Text, KDT) es el proceso que explora grandes conjuntos de datos para identificar patrones útiles y relevantes dentro de ellos. Este proceso también se conoce como minería de datos de texto (Text data mining, TDM) porque puede verse como un proceso de minería de datos que explora datos de texto [22].

El proceso de descubrimiento de conocimiento en texto, implica dominio en diferentes áreas de conocimiento, métodos de recuperación de información, extracción de la información, procesamiento del lenguaje natural y minería de datos. El dominio de estos conocimientos ayuda al investigador a desarrollar cada una de las etapas necesarias previas al descubrimiento de información [62].

De acuerdo a [63] la metodología KDT consta de 5 etapas, estas son:

- **Selección de datos:** consiste en seleccionar los datos adecuados que se procesarán y analizarán en los siguientes pasos.
- **Preprocesamiento de datos:** la tarea es filtrar la información ruidosa y realizar algunos procesos preliminares para facilitar los siguientes pasos, por ejemplo, extraer entidades de nombre de los datos de texto o etiquetar parte del discurso de los datos de texto.

- **Transformación de datos:** los datos de texto se convierten al formato que es fácil de procesar mediante algoritmos de minería, por ejemplo, el formato de vectores, secuencias o tablas de índice invertidas.
- **Minería de texto:** consiste en aplicar algoritmos de minería para encontrar patrones candidatos (palabras o grupos de palabras que se utilizan para identificar conceptos en el texto). Otro autor [64] también incluye en esta fase a la construcción de hipótesis.
- **Interpretación/evaluación:** los patrones candidatos producidos en el paso anterior se evalúan y los interesantes se emiten como conocimiento final.

Estas fases se resumen de forma gráfica en la figura 11.



**Figura 11.** Fases de la metodología KDT

Como se observa en la figura 11, el proceso KDT tiene 5 etapas, sin embargo, algunos autores consideran que la metodología KDT se puede agrupar en 3 fases principales, de acuerdo a [65] [22] estas fases son:

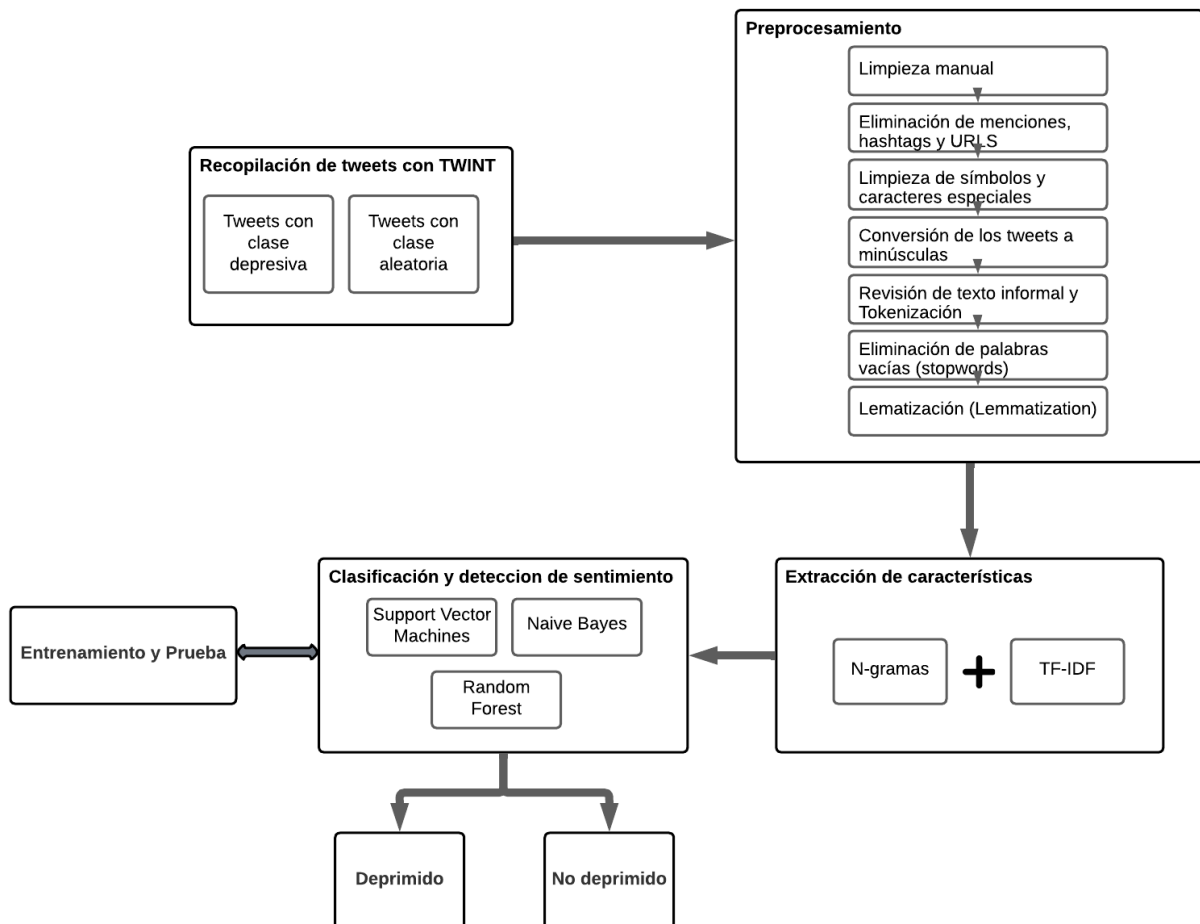
- Preparación de texto o datos
- Minería de texto
- Visualización

En consecuencia, los pasos de la metodología pueden variar dependiendo del propósito de la aplicación, no obstante, las 5 fases de análisis de sentimientos propuestas por los autores ya mencionados tienen concordancia con la metodología KDT (Descubrimiento De Conocimiento En Texto), con lo cual se consideró factible aplicar esta metodología para el análisis de sentimientos en Twitter. Para tener una mejor perspectiva de la relación entre las fases mencionadas anteriormente, se presenta una comparación en la tabla 2.

**Tabla 2.** Comparación fases de Análisis de sentimientos y KDT.

Fases de análisis de sentimientos	Etapas de la metodología KDT
<ul style="list-style-type: none"> <li>• Recolección de datos</li> </ul>	<ul style="list-style-type: none"> <li>• Selección de datos</li> </ul>
<ul style="list-style-type: none"> <li>• Preprocesamiento</li> </ul>	<ul style="list-style-type: none"> <li>• Preprocesamiento de datos</li> </ul>
<ul style="list-style-type: none"> <li>• Extracción de características</li> </ul>	<ul style="list-style-type: none"> <li>• Transformación de datos</li> </ul>
<ul style="list-style-type: none"> <li>• Clasificación de sentimientos</li> <li>• Detección de sentimiento (polaridad)</li> </ul>	<ul style="list-style-type: none"> <li>• Minería de texto (y construcción de hipótesis)</li> </ul>
<ul style="list-style-type: none"> <li>• Evaluación y análisis</li> </ul>	<ul style="list-style-type: none"> <li>• Interpretación/evaluación</li> </ul>

Utilizando la metodología KDT como base para realizar el análisis de sentimientos, se realizó el flujo de trabajo representado en detalle siguiendo las fases especificadas mediante el transcurso de los objetivos del presente proyecto, estos detalles se representan en la figura 12.



**Figura 12.** Flujo detallado del análisis de sentimientos planteado.

### 6.1.1.3. Técnicas de extracción de características

En procesamiento del lenguaje natural (PLN), cualquier problema basado en texto debe convertirse en una forma que pueda modelarse. El principal problema al trabajar con el procesamiento del lenguaje es que los algoritmos de aprendizaje automático no pueden funcionar directamente en el texto sin formato. Por lo tanto, se necesita alguna técnica de extracción de características para convertir el texto en una matriz (o vector) de características [66].

Algunas de las formas de extracción de características más importantes son:

- Bag-of-Words
- TF-IDF

#### 6.1.1.3.1. Bag of Words (BoW)

El modelo BoW se usa en la clasificación de documentos, donde la frecuencia de ocurrencia de cada palabra se usa como una característica para entrenar un clasificador [66]. En este modelo, un texto (como una oración o un documento) se representa como la bolsa (bag) de sus palabras, sin tener en cuenta la gramática e incluso el orden de las palabras, pero manteniendo la multiplicidad (número de veces que aparece una palabra) [67].

#### 6.1.1.3.2. TF-IDF

Es una técnica de recuperación y extracción de información que tiene como objetivo expresar la importancia de una palabra a un documento que forma parte de una colección de documentos que generalmente llamamos corpus [69].

El TF-IDF se basa en el método de frecuencia y tiene en cuenta la aparición de una palabra en todos los documentos. Penaliza las palabras comunes al asignarles pesos más bajos al tiempo que da importancia a las palabras que aparecen menos en todo el corpus pero que aparecen con frecuencia en pocos documentos individuales. Consta de 2 fórmulas combinadas, la frecuencia de término (TF), que especifica la frecuencia con la que un término aparece en todo el documento y frecuencia inversa de documentos (IDF), que mide si un término es raro o frecuente en los documentos de todo el corpus [70].

Y está formulado como:

$$TF - IDF(t, d) = TF(t, d) * IDF(t) \quad (6)$$

donde "TF(t, d)" es el número de veces que aparece la palabra "t" en el documento "d" y,

$$IDF(t) = \log\left(\frac{n}{DF(t)}\right) + 1 \quad (7)$$

donde “n” es el número total de documentos, y “DF(t)” es el número de documentos que contienen la palabra “t” [37].

Esta técnica se usó en el presente TT para llevar a cabo la extracción de características. A diferencia del modelo Bag-of-Words (BoW), TF-IDF tiene en cuenta cuales son las palabras relevantes e irrelevantes presentes en el conjunto de datos. Además, para capturar el orden de las palabras, se empleó en conjunto con otra técnica llamada N-gramas.

El modelo **N-grama** es un método de comprobación de “n” palabras continuas de una secuencia dada de texto o discurso. Este modelo ayuda a predecir el siguiente elemento de una secuencia. En el análisis del sentimiento, el modelo de n-gramas ayuda a analizar el sentimiento del texto o documento. Unigrama se refiere al n-grama de tamaño 1, bigrama se refiere al n-grama de tamaño 2, trigramas se refiere al n-grama de tamaño 3, y así sucesivamente [68].

#### **6.1.1.4. Herramientas usadas para el análisis de sentimientos**

##### **6.1.1.4.1. Pandas**

Pandas es una librería de Python especializada en el manejo y análisis de estructuras de datos de alto nivel diseñadas para que al trabajar con datos estructurados o tabulares sea rápido, fácil y expresivo. Desde su aparición en 2010, ha ayudado a que Python sea un entorno de análisis de datos poderoso y productivo. Los objetos principales en pandas que se utilizan son DataFrame, una estructura de datos tabular orientada a columnas con etiquetas de fila y columna, y Series, un objeto de matriz etiquetado unidimensional [71].

Esta librería fue utilizada para representar los conjuntos de datos en una estructura tabular, leer y escribir en formatos Excel, y para preparar los datos antes de aplicar los algoritmos de aprendizaje automático.

##### **6.1.1.4.2. H3**

H3 es una solución geoespacial para la partición jerárquica y la indexación espacial en la esfera. Desarrollado por Uber, H3 se ha utilizado activamente como una de las herramientas para las propias necesidades operativas de Uber, que incluye la optimización dinámica de los precios de los viajes y el análisis cuantitativo de los datos geográficos para la toma de decisiones, así como para la visualización. Escrita de forma nativa en C, la librería H3 también tiene una gran selección de enlaces disponibles para otros lenguajes de programación. Estos

incluyen, C#, JavaScript, Python y R. Una de las características únicas de H3 es la integración de particiones de rejilla de apertura en hexágonos [72].

Esta herramienta se usó para cubrir todo el territorio del Ecuador con una estructura de malla hexagonal para recopilar la mayor cantidad de publicaciones de Twitter en el área geográfica y periodo de tiempo establecido.

#### **6.1.1.4.3. NLTK (kit de herramientas de lenguaje natural)**

NLTK (Natural Language Toolkit) es una plataforma líder para crear programas de Python y permite trabajar con datos de lenguaje humano. Proporciona interfaces fáciles de usar para más de 50 corpus y recursos léxicos como WordNet, junto con un conjunto de bibliotecas de procesamiento de texto para clasificación, tokenización, lematización, etiquetado, análisis y razonamiento semántico, contenedores para bibliotecas PLN de potencia industrial, y un foro de discusión activo [73]. NLTK fue empleado en el Preprocesamiento de datos para la tokenización de cada una de las publicaciones y para la eliminación de palabras vacías a través del corpus de palabras en español que ofrece esta librería.

#### **6.1.1.4.4. Stanza**

Es una colección de herramientas de procesamiento de lenguaje natural en Python de código abierto que admite 66 idiomas humanos, que introduce una adaptación de su librería CoreNLP a Python.

A pesar de que actualmente existe una considerable variedad de herramientas de PLN disponibles, Stanza fue creada para superar algunas de las limitaciones. En primer lugar, los conjuntos de herramientas existentes suelen ser compatibles con unos pocos idiomas principales. Esto ha limitado significativamente la capacidad de la comunidad para procesar textos multilingües. En segundo lugar, las herramientas de uso generalizado a veces están poco optimizadas en cuanto a precisión, ya sea porque se centran en la eficiencia (por ejemplo, spaCy) o por el uso de modelos menos potentes (por ejemplo, CoreNLP).

Debidos a estos motivos, se seleccionó a Stanza y a NLTK para realizar las tareas de PLN en este proyecto, ya que están basados en Python y sobre todo son multilingüe, con lo cual permite trabajar con texto en español [74].

En la tabla 3, se muestran un resumen de las principales características de algunas bibliotecas comparadas con Stanza y NLTK.

**Tabla 3.** Bibliotecas de software de PLN y sus características

<b>Librería</b>	<b>Lenguaje</b>	<b>Multilingüe</b>	<b>Licencia</b>
NLTK	Python	Si	Apache 2.0
SpaCy	Python	Si	MIT license
Transformers	Python	Parcialmente	Apache 2.0
CoreNLP	Java	Si	GNU GPLv3
Stanza	Python	Si	Apache 2.0

#### **6.1.1.4.5. Openrefine**

Openrefine es una poderosa herramienta para trabajar con datos desordenados; limpiarlos, transformándolo de un formato a otro, y ampliándolo con servicios web y datos externos. Permite la manipulación directa de datos en Wikidata a través de un servicio de reconciliación y una extensión de edición, todo disponible dentro de una interfaz gráfica de usuario y que no requiere habilidades de codificación. Además, los lanzamientos oficiales de OpenRefine se pueden descargar directamente desde el sitio web de la aplicación [75].

Esta herramienta se utilizó para corregir principalmente errores tipográficos presentes en la mayoría de publicaciones a través de la visualización del vocabulario que contiene el conjunto de datos, todo esto como complemento al proceso de preprocesamiento.

#### **6.1.1.4.6. Librería Emoji**

Es una librería de Python que permite imprimir un emoticón en la pantalla usando la función “emojize()” que toma como parámetro el nombre del emoticón encerrado entre dos puntos. En cambio si se desea conocer el texto de un emoticón en particular, se puede usar la función “demojize()” y pasar el emoticón como parámetro <sup>4</sup>. Esta herramienta se empleó para convertir los emoticones de las publicaciones en texto que pueda ser analizado por los algoritmos de clasificación.

#### **6.1.1.4.7. Scikit-learn**

Es una biblioteca de aprendizaje automático de código abierto escrita en Python. Permite la integración fácil y rápida de métodos de aprendizaje automático en código Python. La biblioteca scikit-learn comprende un amplio ancho de banda de métodos para clasificación, regresión, estimación de matriz de covarianza, entre otros [76].

Además, proporciona utilidades para las formas más comunes de extraer características numéricas del contenido de texto. Utilizando la configuración predeterminada de “TfidfTransformer”, podemos determinar la frecuencia de termino y la frecuencia inversa de

---

<sup>4</sup> Python, “emoji · PyPI.” <https://pypi.org/project/emoji/> (accessed Jun. 25, 2022).



documento (ver sección 6.1.1.3.2) [77]. De la misma forma se puede usar el módulo “Tfidfvectorizer” para calcular los valores tf-idf de forma más directa.

En resumen, la principal diferencia entre los dos módulos es la siguiente [78]:

- Con Tfidftransformer, calculará sistemáticamente el conteo de palabras usando CountVectorizer y luego calculará los valores IDF y solo luego calculará las puntuaciones Tf-idf.
- Con Tfidfvectorizer se hace los tres pasos a la vez. De forma interna, calcula el recuento de palabras, los valores IDF y las puntuaciones Tf-idf, todo usando el mismo conjunto de datos.

La librería scikit-learn fue empleada para implementar la técnica tf-idf en la extracción de características, así como los algoritmos de Máquinas de soporte vectorial, Random forest y Naive bayes en código Python.

#### **6.1.1.4.8. Joblib**

Joblib es parte del ecosistema SciPy y proporciona utilidades para canalizar trabajos de Python. La API de Joblib2 proporciona utilidades para guardar y cargar objetos de Python que hacen uso de las estructuras de datos NumPy de manera eficiente [79]. En Python podemos hacer uso de la función joblib para guardar y conservar modelos sklearn. Una vez que el modelo se guarda en el disco o en cualquier otra ubicación, podemos volver a cargarlo o restaurarlo para hacer predicciones sobre nuevos datos [80]. Esta herramienta se usó para guardar los mejores modelos de clasificación entrenados y que puedan ser utilizados en futuras predicciones.

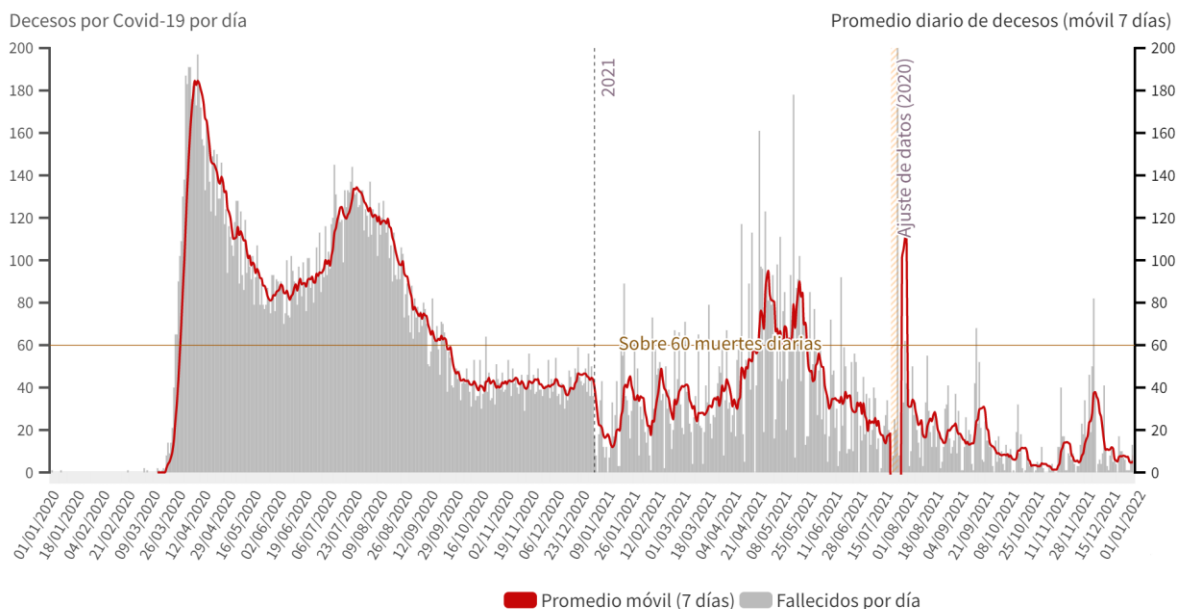
#### **6.1.1.4.9. Pickle**

El módulo pickle de Python se usa para serializar y deserializar una estructura de objeto de Python. Lo que hace pickle primero es serializar el objeto antes de escribirlo en el archivo. El decapado (pickling) es una forma de convertir un objeto python (list, array) en un flujo de caracteres. La idea es que este flujo de caracteres contenga toda la información necesaria para reconstruir el objeto en otro script de Python [81].

Pickle se usó en el presente TT para guardar el vocabulario de unigramas, bigramas y trigramas generado en la extracción de características en un formato reducido y eficiente para utilizarlo en futuras predicciones con nuevos datos.

### 6.1.2. Tarea: Definir intervalo de tiempo para la extracción de datos.

Desde el surgimiento del COVID-19, llegó a tener impacto en el estilo de vida y bienestar de las personas, en Ecuador tuvo mucho impacto sobre todo en el número de muertes provocadas por esta enfermedad. En la figura 13 se observa el número de muertes en promedio a nivel de Ecuador<sup>5</sup> desde el año 2020 hasta inicios del 2022 en donde se puede notar que las muertes por COVID-19 persisten desde el año 2020 hasta fines del año 2021, pero que han ido disminuyendo paulatinamente. Por lo tanto, para la recolección de los tweets se tomó como rango de tiempo desde la fecha en que el COVID-19 fue declarado una preocupación pública internacional por la OMS [8], es decir desde el 30 de enero de 2020, y se estableció como fecha límite hasta el 31 de diciembre del 2021, con lo cual se recopilaron publicaciones en Twitter de aproximadamente 2 años.



**Figura 13.** Promedio de muertes por Covid-19 en Ecuador

Además de lo mostrado en la figura 13, se seleccionaron los dos primeros años de la pandemia COVID-19 para recopilar publicaciones en Twitter debido a que fue un periodo crítico para las personas, caracterizado por miedo, incertidumbre, aislamiento social y cambios drásticos en la vida cotidiana, lo que puede verse reflejado en las publicaciones de Twitter. Además, analizar la información subjetiva a lo largo de estos dos años podría proporcionar una visión más clara del impacto en las publicaciones depresivas durante el inicio de la pandemia.

<sup>5</sup> Observatorio Social del Ecuador, "Personas fallecidas por coronavirus en Ecuador." <https://www.covid19ecuador.org/fallecidos>

### 6.1.3. Tarea: Recopilar tweets mediante una herramienta de scraping en Twitter.

Para realizar la recopilación de los tweets, se empieza a ejecutar las fases de la metodología KDT (ver sección 6.1.1.2), por lo tanto, de aquí en adelante se muestran las fases que se fueron aplicando de acuerdo a las tareas de cada objetivo del presente TT.

#### Fase 1. Selección de datos

Para la extracción de información en la plataforma de Twitter se utilizó la herramienta de Twitter scraping llamada Twint, ya que brinda libre acceso, no tiene limitaciones en la fecha de extracción de los tweets y no se necesita usar la API de Twitter (ver sección 4.4.3). Para la extracción de tweets se utilizaron palabras clave para obtener solo la información que es de interés para el proyecto. Para obtener un mejor resultado en la información extraída, y para saber qué palabras clave son las adecuadas, se tomó en cuenta lo mencionado en [82], en este estudio identificaron con la ayuda de un psicólogo a las palabras más usadas por personas depresivas en publicaciones de Twitter. También de acuerdo a [83] [84], coinciden en que la depresión es con mayor frecuencia autoinformado y al menos inicialmente autoevaluado, por lo tanto, si una persona dice que está deprimida es muy probable que este deprimida.

En base a los puntos mencionados anteriormente, la tabla 4 muestra las palabras clave que se usaron para obtener los tweets que puedan tener indicativos de depresión.

**Tabla 4.** Palabras clave usadas para la extracción de tweets.

<b>PALABRAS CLAVE</b>
agobiado/a
agotado/a
angustiado/a
ansiedad
decaído
depresión
depresivo/a
deprimido/a
desanimado/a
desesperado/a
desmotivado/a
nervioso
antidepresivo/s
deseperanzado
suicidio

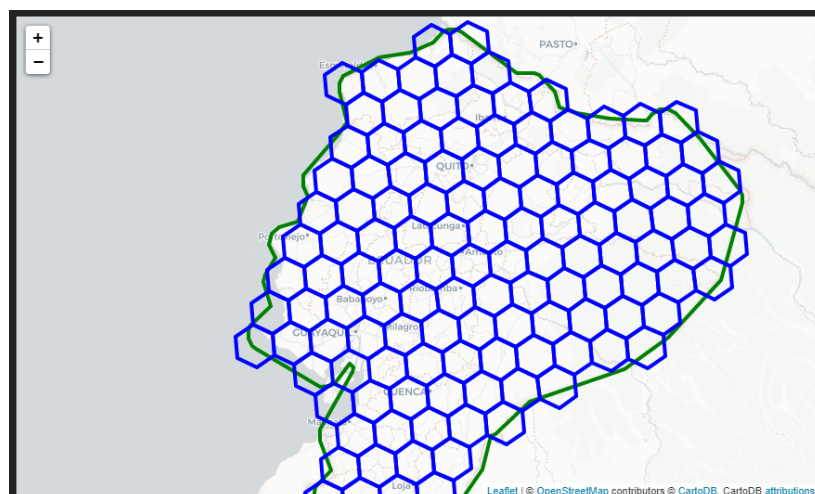
Aparte de la recolección de tweets depresivos, se recolectaron publicaciones aleatorias de twitter que no tengan indicativos de depresión, usando las palabras clave “soy” y “estoy”, ya que representan publicaciones más personales relacionadas al autor del tweet.

Además, para realizar todo el proceso de recopilación de tweets se realizaron las siguientes actividades:

#### a) **Determinar ubicación geográfica de los tweets**

El conjunto de datos relacionado con depresión y el conjunto de datos aleatorio se recopiló de usuarios que se encuentran en Ecuador. Para recopilar las publicaciones de las zonas de Ecuador, se utilizó la geolocalización mediante cuadrículas hexagonales de puntos centrales usando la biblioteca H3 de Uber (ver sección 6.1.1.4.2), la cual permitió generar mosaicos hexagonales de aproximadamente el mismo tamaño que cubran el territorio de Ecuador para lograr obtener la mayor cantidad de Tweets disminuyendo el sesgo en la recopilación.

En la figura 14 se observa el territorio de Ecuador cubierto mediante los mosaicos hexagonales que se generaron.



**Figura 14.** Mapa de mosaicos H3 cubriendo territorio de Ecuador (azul), delimitado por la línea verde.

#### b) **Obtener los Geodatos de Ecuador**

Para obtener las coordenadas que se usaron en la generación del mapa de mosaicos (ver figura 14), se usó la herramienta Geojson.io, la cual facilita un entorno interactivo para generar y visualizar datos geográficos (ver anexo 3).

#### c) **Ejecutar el scraping de tweets con la librería Twint**

A través de la librería Twint, en donde se usaron como parámetros las palabras clave (ver tabla 4) y las distintas ubicaciones geográficas dentro de Ecuador (ver figura 14), se obtuvo un total de 31978 tweets relacionados con depresión. Mediante la recolección de los tweets

se formó un dataset que sirvió como fuente principal de datos para el desarrollo de las siguientes etapas.

En cuanto a la recolección de los tweets aleatorios, se utilizó como parámetros una ubicación geográfica central del Ecuador, ya que no se necesita tanta precisión en recolectar la mayor cantidad de datos posibles como es el caso de los tweets depresivos. La obtención se limitó a 10000 tweets aleatorios extraídos. En el anexo 4 se presenta una previa de los dataset iniciales recopilados.

Cabe recalcar que ambos datasets iniciales tienen datos sin tratar, es decir aún no es factible aplicar algoritmos para un análisis de sentimientos ya que, al ser conjuntos de datos iniciales, contienen mucha información y atributos que son irrelevantes para el objeto de estudio. En la tabla 5 se presentan todos los atributos de cada tweet obtenidos en la recopilación de los dataset iniciales.

**Tabla 5.** Atributos de los tweets recopilados en el dataset inicial.

<b>Atributos de los tweets recopilados</b>	
id	hashtags
conversation_id	cashtags
created_at	link
date	retweet
time	quote_url
timezone	video
user_id	thumbnail
username	near
name	geo
place	source
tweet	user_rt_id
lenguaje	user_rt
mentions	retweet_id
urls	reply_to
photos	retweet_date
replies_count	translate
retweets_count	trans_src
likes_count	trans_dest

De todos los atributos de los dataset iniciales mostrados en la tabla 5, la característica “tweet” es la más relevante para el desarrollo del proyecto ya que contiene el texto de la publicación del tweet. Para que estos datos en bruto sean más fáciles de utilizar y aplicar el análisis de

sentimiento, se realizó un preprocesamiento de datos que se puede comprobar durante el desarrollo del objetivo 2.

Además, para garantizar la reproducibilidad, trazabilidad y transparencia del TT, todo el código utilizado para la extracción de tweets, tanto para obtener las coordenadas geográficas de Ecuador como para ejecutar el scraping con Twint, se llevó a cabo en Jupyter Notebook y está disponible para su revisión detallada en el repositorio GitHub <sup>6</sup>.

## **6.2. Objetivo 2: Aplicar el análisis de sentimientos mediante una técnica basada en Machine Learning.**

Después de recopilar los datos textuales de Twitter, se desarrolló varias tareas cumpliendo con las siguientes fases de la metodología KDT. En primer lugar, se realizó el preprocesamiento de los datos para eliminar el ruido presente en el texto de los tweets, además se aplicó la extracción de características para convertir el texto en vectores numéricos, y finalmente se realizó el entrenamiento y prueba con los datos usando 3 distintos algoritmos de machine learning que son máquinas de soporte vectorial, random forest y naive bayes para conocer el comportamiento y rendimiento de cada uno de ellos con los datos existentes.

### **6.2.1. Tarea: Realizar el Preprocesamiento de los datos obtenidos en la fase anterior.**

#### **Fase 2. Preprocesamiento de datos**

Las publicaciones de Twitter son diferentes de cualquier otro texto en libros o artículos, ya que incluyen usos idiosincrásicos como menciones de usuarios, retweets, texto informal, entre otros; por lo que se requiere cierto procesamiento a los datos recolectados, por ejemplo en la publicación siguiente: *"#NoMasPresxsPorPlantar #ReglamentosParaTodxs #NoMasPresxsPorFumar #GMM20 #MMMEC20 @Lenin @ottosonnenh Y todo esto hablando del THC, recreacional y espiritual, pero es imposible no mencionar sus cualidades curativas, el CBD es un gran analgésico, Antidepresivo, ansiolítico!! "* es difícil que los algoritmos puedan entender cuál es el contenido real del tweet.

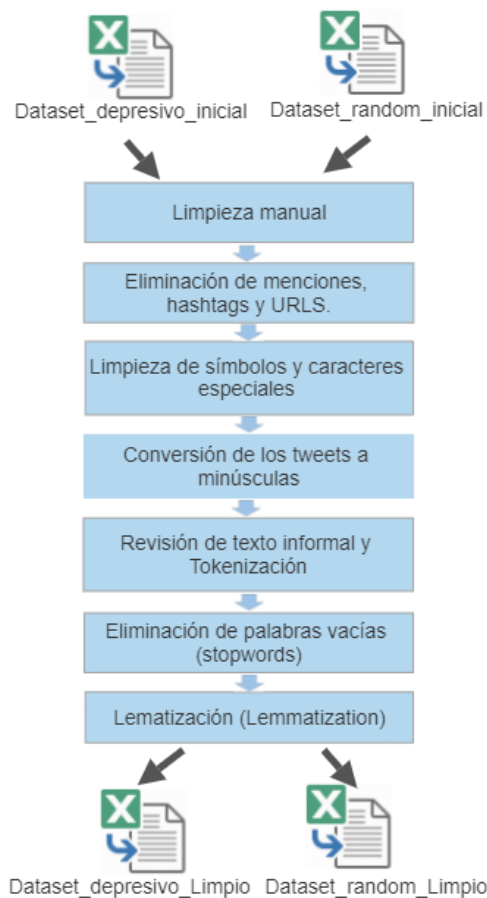
Cabe mencionar que ciertas características para el preprocesamiento ya fueron tomadas en cuenta al momento de la extracción de los tweets usando la herramienta Twint; por ejemplo, no se extrajeron los tweets que contenían links, ya que estos suelen ser usados en su mayoría con fines promocionales, además no se extrajeron los retweets porque por lo general no suelen representar la opinión del autor de esa publicación.

---

<sup>6</sup> Código en repositorio de la extracción de tweets:

[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/tree/main/1.Extraccion\\_Tweets](https://github.com/byronmb/Identificacion_Depresion_Ecuador/tree/main/1.Extraccion_Tweets)

Se llevaron a cabo varios pasos para el preprocesamiento de los datos, en la figura 15 se muestra el flujo de trabajo.



**Figura 15.** Flujo de trabajo del preprocesamiento de datos.

### a. Limpieza manual

En primer lugar, se realizó una limpieza manual del dataset que contiene tweets depresivos (“Dataset\_depresivo\_inicial”) para eliminar tweets que hacen referencia a publicidad, tweets informativos, menciones a otras personas, entre otros. Además, los tweets depurados de forma manual se validaron con la ayuda de la Dra. Sandra Otero, Magister en Psicología Clínica Infantojuvenil<sup>7</sup>, para garantizar la calidad de los datos manteniendo solo los tweets que indicaban depresión emocional. En cuanto al dataset de tweets aleatorios (“Dataset\_random\_inicial”), también se realizó la limpieza manual para eliminar datos que hagan referencia a publicidad, temas políticos, referente a otras personas, entre otros. Aparte de esto, se añadieron 3000 tweets que fueron descartados del dataset depresivo para que el dataset aleatorio contenga palabras relacionadas a depresión, ansiedad, entre otros, pero que no están considerados como depresivos.

<sup>7</sup> Evidencia del proceso de validación con la Dra. Sandra Otero:  
[https://drive.google.com/drive/folders/1WKKCqOgM\\_PI1aYCbVBTIHoH6-4Zu8vhW?usp=sharing](https://drive.google.com/drive/folders/1WKKCqOgM_PI1aYCbVBTIHoH6-4Zu8vhW?usp=sharing)

Asimismo, antes de realizar la filtración de las características propias de los tweets, se eliminaron registros duplicados y atributos que no son de utilidad (ver tabla 5).

### b. Eliminación de menciones, hashtags y URLs.

Mediante el uso de la sustitución basada en expresiones regulares, se eliminaron los símbolos hashtags (solo símbolo #), y menciones (@nombreusuario) de los tweets. También se procesó las palabras individuales de los tweets; es decir, se eliminaron vocales repetidas presentes en muchas palabras (por ejemplo: “hoolaaaa” a “hola”), tomando en cuenta algunas excepciones como “facebook”, “mood”, “boomerang”, “desee”, entre otras palabras que se descubrieron que estaban presentes en los tweets al realizar la limpieza manual.

Algunos ejemplos de tweets preprocesados con estas características se muestran en la tabla 6.

**Tabla 6.** Ejemplos de eliminación de menciones, hashtags y urls en los tweets

	<b>Tweet Original</b>	<b>Tweet después del preprocesamiento</b>
<b>Eliminación de menciones</b>	@paolaopioide @CarlosSacoto7 @PoliciaEcuador @teleamazonasec Solo los que sufrimos de depresion sabemos lo terrible de eso.	Solo los que sufrimos de depresion sabemos lo terrible de eso.
<b>Eliminación del símbolo “#”</b>	Soy fuerte, pero a veces eso de serlo cansa. #cansado #sad #auxilio #depresion	Soy fuerte, pero a veces eso de serlo cansa. <b>cansado sad auxilio depresion</b>
<b>Eliminación de vocales repetidas</b>	Es bueno escuchar achaque tras achaque, que te señalen <b>tooooooos</b> tus errores que te digan recién que todo el ejercicio que estabas haciendo esta en parte mal.? Cuándo te ven casi destruida y deprimida porque hoy viste tus verdaderos resultados en números.	Es bueno escuchar achaque tras achaque, que te señalen <b>todos</b> tus errores que te digan recién que todo el ejercicio que estabas haciendo esta en parte mal.? Cuándo te ven casi destruida y deprimida porque hoy viste tus verdaderos resultados en números.

### c. Limpieza de símbolos y caracteres especiales.

Los tweets extraídos están llenos de caracteres especiales como “\$” y “-”, que son importantes de eliminar antes de cualquier tipo de análisis de datos, por lo tanto se eliminaron todos los signos de puntuación, símbolos y números que generalmente no suelen contener información que ayude al análisis de sentimientos y que podrían generar resultados impredecibles si se ignoran. También se eliminaron espacios extra presentes en el texto de forma que solo haya un espacio entre palabras, todo esto se realizó mediante el uso de expresiones regulares.



Además, tomando en cuenta que muchos tweets tienen la característica de contener uno o más emoticonos que representan distintos sentimientos de las personas que los escriben, se convirtieron todos los emoticonos a texto en español mediante el uso del paquete de Python “emoji” (ver sección 6.1.1.4.6).

En la tabla 7, se presentan ejemplos de tweets a los cuales se aplicaron las distintas características mencionadas en el preprocesamiento.

**Tabla 7.** Ejemplos de limpieza de símbolos, espacios repetidos y conversión de emoticonos en los tweets

	<b>Tweet Original</b>	<b>Tweet después del preprocesamiento</b>
<b>Eliminación de signos de puntuación y símbolos.</b>	Es normal, hay días que entre la depresión.?... bueno esta noche es mi caso 😞😞 ganas de gritar, salir corriendo y cerrar mis ojos y que ya haya terminado todo este mal	Es normal hay días que entre la depresión bueno esta noche es mi caso 😞😞 ganas de gritar salir corriendo y cerrar mis ojos y que ya haya terminado todo este mal
<b>Eliminación de números</b>	Llevo <b>13</b> años con ataques de pánico y ansiedad que suelen darme de manera espontánea y rara vez....es algo que me hace sentir triste por que se que no soy normal como cualquier otro ser humano. En cambio yo vivo con el miedo de cuando me volverá a dar 😞 ataquedeapanico	Llevo años con ataques de pánico y ansiedad que suelen darme de manera espontánea y rara vez es algo que me hace sentir triste por que se que no soy normal como cualquier otro ser humano En cambio yo vivo con el miedo de cuando me volverá a dar 😞 ataquedeapanico
<b>Eliminación de espacios extra</b>	Estoy desesperada, sin poder gritar ni quejarme con nadie, porque todos estamos igual. Sumidos en desesperación silenciosa. FuckCovid	Estoy desesperada sin poder gritar ni quejarme con nadie porque todos estamos igual Sumidos en desesperación silenciosa FuckCovid
<b>Conversión de emoticonos a texto</b>	La depresión es una cosa horrible 😞 y he escuchado hasta a psicólogos tomarlo a la ligera Un abrazo 🤗 piensa que ahora debe estar mejor	La depresión es una cosa horrible :cara_llorando: y he escuchado hasta a psicólogos tomarlo a la ligera Un abrazo :cara_con_manos_abrazando: piensa que ahora debe estar mejor

En relación a los emoticonos presentes en los tweets, se encontró que muchas personas usan varios emoticonos repetidos, por este motivo se realizó un preprocesamiento del texto convertido de los emoticonos (ver tabla 7). En primer lugar se eliminó el texto de los emoticonos repetidos, también se agregaron espacios entre los emoticonos utilizando como puntos de referencia la estructura del texto convertido (:texto\_emote:), y por último se eliminaron los dos puntos generados en el texto al momento de la conversión automática realizada en Python.

Los pasos mencionados para depurar los emoticonos convertidos a texto se representan mediante ejemplos en la tabla 8.

**Tabla 8.** Ejemplos de depuración de los textos que fueron convertidos a partir de los emoticones

	<b>Tweet Original</b>	<b>Tweet después del preprocesamiento</b>
<b>Eliminar repetidos</b>	Ya me dio una ansiedad maldita :cara_llorando_fuerte::cara_llorando_fuerte: :cara_llorando_fuerte:	Ya me dio una ansiedad maldita :cara_llorando_fuerte:
<b>Agregar espacios</b>	Y es así como la ansiedad me envuelve otra vez :cara_decepcionada::cara_desanimada:	Y es así como la ansiedad me envuelve otra vez :cara_decepcionada: :cara_desanimada:
<b>Quitar 2 puntos</b>	No crees en depresión y ansiedad hasta que te pasa :cara_desanimada:	No crees en depresión y ansiedad hasta que te pasa cara_desanimada

Así mismo, al ser Twitter de naturaleza informal, las personas no acostumbran a hacer uso de las tildes en muchas publicaciones, por esta razón las palabras “relación” y “relacion” son consideradas por los algoritmos como distintas. Para evitar esta pérdida de relación semántica, se eliminaron todas las tildes de las vocales presentes en los tweets y se reemplazaron por su equivalente sin tilde.

En la tabla 9 se presenta un ejemplo de la eliminación de tildes en un tweet.

**Tabla 9.** Ejemplo de eliminación de tildes en un Tweet.

	<b>Tweet Original</b>	<b>Tweet después del preprocesamiento</b>
<b>Eliminación de tildes</b>	Hoy no tengo ganas de nada. Maldita depresión	Hoy no tengo ganas de nada Maldita depresion

#### **d. Conversión de los tweets a minúsculas**

Para los algoritmos de aprendizaje no es lo mismo la palabra “depresion” que “DEPRESION”. Estas palabras son tratadas como dos totalmente distintas, sin ningún tipo de relación entre ellas. Para evitar que esto suceda y mantener el significado de las palabras sin tener en cuenta la forma de sus caracteres, todos los tweets se convirtieron a su equivalente en letras minúsculas. En la tabla 10 se muestra un ejemplo de un tweet convertido a minúsculas.

**Tabla 10.** Ejemplo de conversión de tweet a minúsculas.

	<b>Tweet Original</b>	<b>Tweet después del preprocesamiento</b>
<b>Conversión del texto a minúsculas</b>	siento que voy a caer de nuevo en depresión! <b>AYUDA</b>	siento que voy a caer de nuevo en depresion <b>ayuda</b>

## e. Revisión de texto informal y Tokenización

### • Revisión del texto

Debido al texto informal presente en Twitter, existen muchas palabras que están escritas incorrectamente, ya sea porque están unidas unas con otras, están incompletas o simplemente están mal escritas. Para resolver esto y otros errores que se pasaron por alto en las fases anteriores del preprocesamiento, se exportaron los datasets con el nombre “Dataset\_depresivo\_semiLimpio” y “Dataset\_random\_semiLimpio” para procesarlos mediante el uso de la herramienta Openrefine (ver sección 6.1.1.4.5), ya que brinda una interfaz intuitiva para encontrar inconsistencias en las palabras de ambos datasets. Como recurso de apoyo se preparó un video<sup>8</sup> que presenta una breve demostración de la utilización que se le dio a Openrefine con el objetivo de brindar una visión general de cómo se logró la resolución de los errores en el texto.

### • Tokenización

Luego de la depuración del texto informal en Openrefine se obtuvieron los dataset “Dataset\_depresivo\_Openrefine” y “Dataset\_random\_Openrefine”, los cuales fueron utilizados a partir de este paso para continuar con algunas fases más hasta completar el preprocesamiento de los tweets.

En este paso de la Tokenización, se dividió cada tweet en palabras o tokens, es decir se delimitaron las palabras del texto y se convirtieron esas palabras en elementos de una lista para su procesamiento en las siguientes fases. Este proceso se lo realizó mediante el uso de la librería NLTK (ver sección 6.1.1.4.3).

En la tabla 11, se muestra un ejemplo de un tweet que pasó por la fase de Tokenización.

**Tabla 11.** Ejemplo de un tweet tokenizado.

	<b>Tweet Original</b>	<b>Tweet después del preprocesamiento</b>
<b>Tokenización del texto</b>	estoy deprimido todos los dias amanezco con muchas ganas y pienso en que hacer llega el siguiente dia y paso durmiendo todo el dia estoy durmiendo aproximadamente horas al dia parece que estuviera normal pero no cara_sonriendo_ligeramente y no tengo ganas de nada nada nada	[ 'estoy', 'deprimido', 'todos', 'los', 'dias', 'amanezco', 'con', 'muchas', 'ganas', 'y', 'pienso', 'en', 'que', 'hacer', 'llega', 'el', 'siguiente', 'dia', 'y', 'paso', 'durmiendo', 'todo', 'el', 'dia', 'estoy', 'durmiendo', 'aproximadamente', 'horas', 'al', 'dia', 'parece', 'que', 'estuviera', 'normal', 'pero', 'no', 'cara_sonriendo_ligeramente', 'y', 'no', 'tengo', 'ganas', 'de', 'nada', 'nada', 'nada' ]

<sup>8</sup> Openrefine para la resolución de errores en el texto: <https://youtu.be/M3Lq21P4J74>

#### f. Eliminación de palabras vacías (stopwords)

Los tweets por lo general contienen muchas palabras que, aunque son necesarias para construir oraciones con sentido, carecen de información que pueden afectar la eficiencia del modelo de manera adversa. Estas palabras son proposiciones, pronombres, conjunciones, entre otras; por lo tanto, en esta etapa se aplicó la filtración de palabras vacías o stopwords con el propósito de reducir el número de palabras almacenadas en la lista de tokens procesados en la fase anterior.

La filtración de las palabras vacías se realizó utilizando el corpus de palabras vacías proporcionado por la librería NLTK (ver sección 6.1.1.4.3), excepto algunos pronombres en primera persona y palabras de negación; ya que, dependiendo del tipo de publicación pueden cambiar el significado del texto.

La tabla 12 muestra un ejemplo de filtración de las palabras vacías en un tweet.

**Tabla 12.** Ejemplo de eliminación de stopwords de un tweet

	<b>Tweet Tokenizado</b>	<b>Tweet después del preprocesamiento</b>
<b>Eliminación de stopwords en el texto</b>	['estoy', 'deprimido', 'todos', 'los', 'dias', 'amanezco', 'con', 'muchas', 'ganas', 'y', 'pienso', 'en', 'que', 'hacer', 'llega', 'el', 'siguiente', 'dia', 'y', 'paso', 'durmiendo', 'todo', 'el', 'dia', 'estoy', 'durmiendo', 'aproximadamente', 'horas', 'al', 'dia', 'parece', 'que', 'estuviera', 'normal', 'pero', 'no', 'cara_sonriendo_ligeramente', 'y', 'no', 'tengo', 'ganas', 'de', 'nada', 'nada', 'nada']	['deprimido', 'dias', 'amanezco', 'muchas', 'ganas', 'pienso', 'hacer', 'llega', 'siguiente', 'dia', 'paso', 'durmiendo', 'dia', 'durmiendo', 'aproximadamente', 'horas', 'dia', 'parece', 'normal', 'pero', 'no', 'cara_sonriendo_ligeramente', 'no', 'ganas', 'nada', 'nada', 'nada']

#### g. Lematización (Lemmatization)

Existen muchas palabras diferentes que representan a una misma palabra. Por ejemplo, las palabras “canto”, “cantas”, “canta” son distintas conjugaciones de un mismo verbo (cantar). Por lo tanto, se aplicó la lematización a las palabras procesadas en la fase anterior para reducirlas a su forma canónica o lema.

Al ser el texto en español, se utilizó la librería Stanza para realizar la lematización de las palabras procesadas porque da mejores resultados para tratar morfología<sup>9</sup> en español (ver sección 6.1.1.4.4). En la tabla 13, se presenta un ejemplo de un tweet preprocesado al que se le aplicó la lematización.

<sup>9</sup> La morfología estudia la relación entre la forma de las palabras y la información gramatical y semántica que contienen las distintas formas que puede adoptar una misma palabra [93].

Tabla 13. Ejemplo de texto aplicado la lematización

	Tweet Tokenizado	Tweet después del preprocesamiento
Lematización del texto	['a', 'veces', 'es', 'valido', 'sentirse', 'agobiada', 'desesperada', 'y', 'con', 'una', 'incertidumbre', 'tenaz']	['vez', 'valer', 'sentirse', 'agobiada', 'desesperado', 'incertidumbre', 'tenaz']

Los datasets que fueron procesados mediante todas las fases mencionadas anteriormente se guardaron con el nombre de “Dataset\_depresivo\_Limpio” para los tweets depresivos y “Dataset\_random\_Limpio” para los tweets aleatorios.

Finalmente, para fines de visualización de datos, se realizó un análisis de nube de palabras (wordcloud) de los tweets preprocesados en ambos conjuntos de datos.



Figura 16. Nube de palabras de los tweets aleatorios.



Figura 17. Nube de palabras de los tweets depresivos

La figura 16 muestra la presencia de palabras encontradas en el conjunto de datos formado a partir de tweets aleatorios que no contienen ningún rasgo que indique depresión y se observa una gran diversidad en las palabras que están presentes con más frecuencia, se distingue palabras como ansiedad y depresión que resaltan debido a la inclusión de tweets descartados del dataset con tweets depresivos que contienen estas palabras pero que no expresan tendencias depresivas. La figura 17 en cambio muestra claramente la presencia de ciertas palabras que representan una emoción negativa y la presencia de tales palabras en el lenguaje de una persona es un indicador de tendencia depresiva.

Para garantizar la reproducibilidad, trazabilidad y transparencia del TT, en el repositorio<sup>10</sup> se presenta con mayor detalle todo el proceso de preprocesamiento a los datos que fueron realizados en Openrefine y en Jupyter Notebook.

### **6.2.2. Tarea: Extracción de características**

#### **Fase 3. Transformación de datos**

Los algoritmos de aprendizaje automático funcionan con valores numéricos, por lo tanto, primero se tiene que representar el texto en números o vectores de números. Para realizar esto se usó la técnica de tf-idf (ver sección 6.1.1.3.2) para convertir el texto en una matriz o vector de características en base a los tweets preprocesados.

En primer lugar, se cargaron los datasets (“Dataset\_depresivo\_Limpio”, “Dataset\_random\_Limpio”) que fueron preprocesados en la sección anterior (sección 6.2.1), y se etiquetaron de acuerdo a la característica de los tweets, estas son:

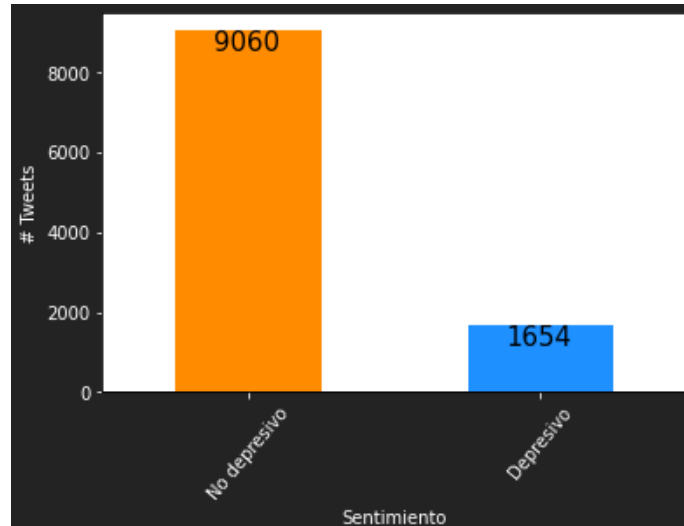
- Clase 0 para tweets con sentimiento aleatorio (no depresivo).
- Clase 1 para tweets con sentimiento depresivo.

Luego se unificaron y se mezclaron ambos datasets clasificados para generar un solo conjunto de datos que contiene tanto tweets depresivos como tweets no depresivos. El conjunto de datos unido y etiquetado se almacenó en un archivo llamado “Dataset\_unido” para uso posterior, quedando un total de 10714 tweets en el conjunto de datos, en la figura 18 se muestra una representación de la cantidad de tweets clasificados con su respectiva clase.

---

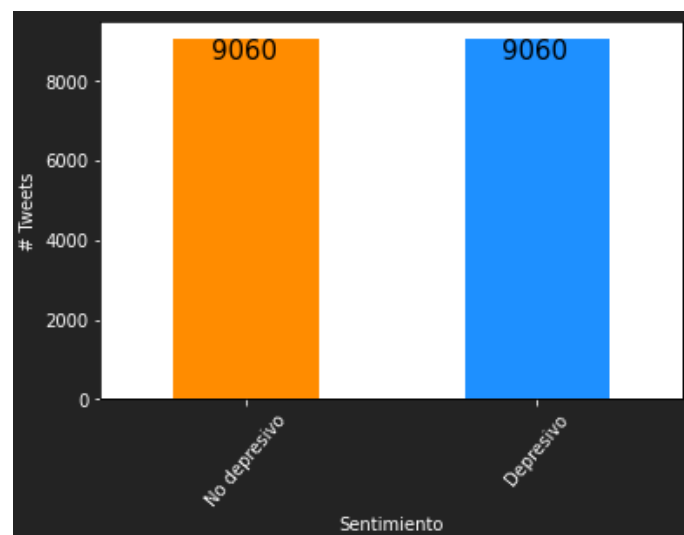
<sup>10</sup> Código en repositorio del preprocesamiento:

[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/tree/main/2.Preprocesamiento](https://github.com/byronmb/Identificacion_Depresion_Ecuador/tree/main/2.Preprocesamiento)



**Figura 18.** Tweets clasificados como aleatorios y depresivos.

Como se muestra en la figura 18, los datos que fueron recolectados y preprocesados muestran un desequilibrio en sus clases; ya que, en la limpieza de los tweets depresivos recolectados, la cantidad de datos se redujeron significativamente; por lo tanto, para garantizar que los algoritmos de aprendizaje no estén sesgados hacia la clase mayoritaria, se aplicó la técnica de sobremuestreo de minorías sintéticas SMOTE (ver sección 4.3.2.2) utilizando “over\_sampling” de la librería imblearn para equilibrar la distribución de las clases en el conjunto de datos, como se observa en la figura 19.



**Figura 19.** Conjunto de datos con las clases equilibradas

Se realizó la extracción de características utilizando “TfidfTransformer” de la librería sklearn (sección 6.1.1.4.7) para conocer de forma más clara y detallada este proceso. Se realizó el cálculo de la frecuencia de términos (TF), la frecuencia inversa de documentos (IDF) para finalmente obtener los valores tf-idf del conjunto de datos. Todo este proceso para obtener el

TF-IDF vectorizado se lo realizó para unigramas, bigramas y trigramas, a fin de comparar los resultados con cada uno de ellos en la fase de entrenamiento. Los pasos en detalle se muestran a continuación.

### 6.2.2.1. TF-IDF para unigramas

- **Cálculo de la frecuencia de términos (TF)**

Mediante el módulo CountVectorizer proporcionado por la biblioteca Scikit-learn se calculó el número de veces que aparece cada palabra en cada uno de los documentos (tweets), y se guardó el vocabulario de todos los unigramas posibles con el nombre de "vocabulary\_Unigrama.pkl" para uso posterior en posibles predicciones con nuevos datos, esto se realizó mediante el uso de la librería Pickle (ver sección 6.1.1.4.9). En la tabla 14 se muestra una vista generada del recuento de todas palabras del conjunto de datos.

**Tabla 14.** Recuento de palabras (TF) del conjunto de datos.

	ab	abajo	abandonar	abandono	....	zurda	zurdo	ñaña	ñaño
<b>0</b>	0	0	0	0	....	0	0	0	0
<b>1</b>	0	0	0	0	....	0	0	0	0
<b>2</b>	0	0	0	0	....	0	0	0	0
<b>3</b>	0	0	0	0	....	0	0	0	0
<b>4</b>	0	0	0	0	....	0	0	0	0
....	....	....	....	....	....	....	....	....	....
<b>10709</b>	0	0	0	0	....	0	0	0	0
<b>10710</b>	0	0	0	0	....	0	0	0	0
<b>10711</b>	0	0	0	0	....	0	0	0	0
<b>10712</b>	0	0	0	0	....	0	0	0	0
<b>10713</b>	0	0	0	0	....	0	0	0	0

El recuento de palabras (unigramas) obtuvo como resultado 10714 filas, que es el total de documentos o tweets y 7152 columnas, es decir el total de palabras únicas (unigramas) que están presentes en todo el dataset y que están presentes en al menos 2 documentos (min\_df=2).

Como se observa en la tabla 14, la mayoría de los elementos en el vector de características son 0, por lo que se genera una matriz dispersa. Esto se debe a que el número de palabras que aparecen en cada documento (tweet) es solo una pequeña parte de todo el conjunto de documentos; por lo tanto, hay muchas palabras que no aparecen y se marcarán como 0.



- **Cálculo de los valores IDF**

Se calculó los valores IDF en base a los recuentos de palabras que se obtuvieron anteriormente. En la tabla 15 se muestra algunos valores IDF que se obtuvieron mediante un marco de datos de Python, estos valores se presentan ordenados de forma ascendente.

**Tabla 15.** Previa de los valores idf para unigramas del conjunto de datos

	valores_idf
no	1.989609
yo	2.209745
ansiedad	2.612710
hacer	3.196851
depresión	3.330503
...	...
piloto	9.180788
pilsener	9.180788
pincel	9.180788
escobo	9.180788
descontrolado	9.180788

Como se muestra en la tabla 15, las palabras “no” y “yo” tienen los valores IDF más bajos. Esto quiere decir que estas palabras aparecen en la mayoría de los documentos de la colección, ya que cuanto menor sea el valor IDF de una palabra, menos única será para cualquier documento en particular.

Además, hay que tener en cuenta que la fórmula TF-IDF de la notación estándar de los libros de texto (ver sección 6.1.1.3.2) es diferente a la fórmula de cálculo de sklearn, en donde se agrega la constante “1” al numerador y denominador (ecuación 8) lo cual evita las divisiones por cero. Por lo tanto, la fórmula TF-IDF queda como se muestra en la ecuación 9.

$$idf(t) = \log \frac{n_{d+1}}{1 + df(t)} + 1 \quad (8)$$

$$tf - idf(t, d) = tf(t, d) * idf(t) \quad (9)$$

- **Cálculo de las puntuaciones TF-IDF**

Una vez que se calculó los valores IDF, ya se puede calcular las puntuaciones tf-idf para cualquier documento (tweet) o conjunto de documentos.

Se realizó el cálculo de las puntuaciones tf-idf para todos los documentos del conjunto de datos. En la tabla 16 se presenta una previa de la representación de valores tf-idf de un documento (tweet) que se colocó en un marco de datos para fines de visualización y se ordenaron sus puntajes en orden descendente.

**Tabla 16.** Valores tf-idf de un tweet del conjunto de datos

	<b>tfidf</b>
<b>gym</b>	0.597944
<b>ayudar</b>	0.490538
<b>matar</b>	0.458744
<b>deber</b>	0.381294
<b>ansiedad</b>	0.214487
...	...
<b>divertido</b>	0.000000
<b>diversión</b>	0.000000
<b>distraído</b>	0.000000
<b>distraída</b>	0.000000
<b>ñaño</b>	0.000000

En la tabla 16 se observa que solo ciertas palabras tienen puntaje tf-idf, ya que son solo las que aparecen en este documento en particular. Todas las palabras de este documento tienen una puntuación tf-idf y todo lo demás aparece como ceros, ya que cuanto más común sea la palabra en todos los documentos, menor será su puntuación; por ejemplo, en este caso la palabra “ansiedad” tiene una puntuación menor en comparación a las demás palabras, esto quiere decir que esta palabra está presente en muchos más documentos. Además, cuanto más exclusiva sea una palabra en el actual documento (por ejemplo, "gym" y "ayudar"), mayor será la puntuación.

Se obtuvo las variables independientes y dependientes aplicando el sobremuestreo con SMOTE, para la variable independiente se usó el vector de características tf-idf obtenido y para la variable dependiente se usó la etiqueta sentimiento (0 para no depresivo y 1 para depresivo).

Finalmente, las variables independientes y dependiente obtenidas se almacenaron con el nombre de “x\_tfidf\_Unigrama” y “y\_tfidf\_Unigrama” respectivamente para uso posterior en el entrenamiento y prueba de los modelos.

#### **6.2.2.2. TF-IDF para bigramas**

- **Cálculo de la frecuencia de términos (TF)**

Mediante el módulo CountVectorizer proporcionado por la biblioteca Scikit-learn se calculó el número de veces que aparece cada bigrama (2 palabras consecutivas) en cada uno de los documentos (tweets), y se guardó el vocabulario de todos los bigramas posibles con el nombre de “vocabulary\_Bigrama.pkl” para uso posterior en posibles predicciones con nuevos datos, esto se realizó mediante el uso de la librería Pickle (ver sección 6.1.1.4.9). En la tabla 17 se muestra una vista generada del recuento de todos los bigramas del conjunto de datos.

**Tabla 17.** Recuento de bigramas (TF) del conjunto de datos.

	abajo no	abandonar serie	abogado contraparte	aborto asesinato	...	yo volvi	yo yo	you can	zapatitos rojos
0	0	0	0	0	...	0	0	0	0
1	0	0	0	0	...	0	0	0	0
2	0	0	0	0	...	0	0	0	0
3	0	0	0	0	...	0	0	0	0
4	0	0	0	0	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...
10709	0	0	0	0	...	0	0	0	0
10710	0	0	0	0	...	0	0	0	0
10711	0	0	0	0	...	0	0	0	0
10712	0	0	0	0	...	0	0	0	0
10713	0	0	0	0	...	0	0	0	0

El recuento de bigramas obtuvo como resultado 10714 filas, que es el total de documentos o tweets y 9043 columnas, es decir el total de bigramas únicos que están presentes en todo el dataset y que están presentes en al menos 2 documentos ( $\text{min\_df}=2$ ).

Como se muestra en la tabla 17, al igual que con los unigramas, la mayoría de los elementos en el vector de características son 0, por lo que se genera una matriz dispersa. Esto se debe a que el número de bigramas que aparecen en cada documento (tweet) es solo una pequeña parte de todo el conjunto de documentos; por lo tanto, hay muchos bigramas que no aparecen y se marcan como 0.

- **Cálculo de los valores IDF**

Se calculó los valores IDF en base a los bigramas que se obtuvieron anteriormente. En la tabla 18 se muestra algunos valores IDF que se obtuvieron mediante un marco de datos de Python, estos valores se presentan ordenados de forma ascendente.

**Tabla 18.** Previa de los valores idf para bigramas del conjunto de datos

	valores_idf
no yo	4.194900
yo dar	4.323563
no poder	4.468259
yo decir	5.069914
ansiedad yo	5.337757
...	...
hecho bolita	9.180788
harto mascarilla	9.180788
harto leer	9.180788
hombre parecer	9.180788
zapatitos rojos	9.180788

Como se muestra en la tabla 18, los bigramas “no yo” y “yo dar” tienen los valores IDF más bajos. Esto quiere decir que estos bigramas aparecen en la mayoría de los documentos de la colección, ya que cuanto menor sea el valor IDF del bigrama, menos única será para cualquier documento en particular.

- **Cálculo de las puntuaciones TF-IDF**

Una vez que se calculó los valores IDF para bigramas, se calculan las puntuaciones tf-idf para cualquier documento (tweet) o conjunto de documentos. Se realizó el cálculo de las puntuaciones tf-idf para todos los documentos del conjunto de datos. En la tabla 19 se presenta una previa de la representación de valores tf-idf para bigramas de un documento (tweet) que se colocó en un marco de datos para fines de visualización y se ordenaron sus puntajes en orden descendente.

**Tabla 19.** Valores tf-idf (bigramas) de un tweet del conjunto de datos

	<b>tfidf</b>
<b>hacer ejercicio</b>	0.596923
<b>ejercicio deber</b>	0.384000
<b>peso no</b>	0.384000
<b>desesperado hacer</b>	0.344435
<b>bajar peso</b>	0.319055
...	...
<b>etapa vida</b>	0.000000
<b>etapa super</b>	0.000000
<b>etapa depresión</b>	0.000000
<b>estupida vez</b>	0.000000
<b>zapatitos rojo</b>	0.000000

En la tabla 19 se observa que solo ciertos bigramas tienen puntaje tf-idf, ya que son solo los que aparecen en este documento en particular. Todos los bigramas de este documento tienen una puntuación tf-idf y todo lo demás aparece como ceros, ya que cuanto más común sea el bigrama en todos los documentos, menor será su puntuación; por ejemplo, en este caso los bigramas “bajar peso” y “desesperado hacer” tienen una puntuación menor en comparación a los demás bigramas, esto quiere decir que estos bigramas están presentes en muchos más documentos. Además, cuanto más exclusivo sea un bigrama en el actual documento (por ejemplo, "hacer ejercicio" y "ejercicio deber"), mayor será la puntuación.

Se obtuvo las variables independientes y dependientes aplicando el sobremuestreo con SMOTE, para la variable independiente se usó el vector de características tf-idf obtenido y para la variable dependiente se usó la etiqueta sentimiento (0 para no depresivo y 1 para depresivo).

Finalmente, las variables independientes y dependiente obtenidas se almacenaron con el nombre de “x\_tfidf\_Bigrama” y “y\_tfidf\_Bigrama” respectivamente para uso posterior en el entrenamiento y prueba de los modelos.

### 6.2.2.3. TF-IDF para trigramas

- **Cálculo de la frecuencia de términos (TF)**

Mediante el módulo CountVectorizer proporcionado por la biblioteca Scikit-learn se calculó el número de veces que aparece cada trigrama (3 palabras consecutivas) en cada uno de los documentos (tweets), y se guardó el vocabulario de todos los trigramas posibles con el nombre de “vocabulary\_Trigrama.pkl” para uso posterior en posibles predicciones con nuevos datos, esto se realizó mediante el uso de la librería Pickle (ver sección 6.1.1.4.9). En la tabla 20 se muestra una vista generada del recuento de todos los trigramas del conjunto de datos.

**Tabla 20.** Recuento de trigramas (TF) del conjunto de datos.

	abandonar serie duro	abrace yo decir	abrazo depresion ansiedad	abri texto biblico	...	yo volver adicta	yo volviendo adicta	yo volviendo loco	you can do
<b>0</b>	0	0	0	0	...	0	0	0	0
<b>1</b>	0	0	0	0	...	0	0	0	0
<b>2</b>	0	0	0	0	...	0	0	0	0
<b>3</b>	0	0	0	0	...	0	0	0	0
<b>4</b>	0	0	0	0	...	0	0	0	0
...	...	...	...	...	...	...	...	...	...
<b>10709</b>	0	0	0	0	...	0	0	0	0
<b>10710</b>	0	0	0	0	...	0	0	0	0
<b>10711</b>	0	0	0	0	...	0	0	0	0
<b>10712</b>	0	0	0	0	...	0	0	0	0
<b>10713</b>	0	0	0	0	...	0	0	0	0

El recuento de trigramas obtuvo como resultado 10714 filas, que es el total de documentos o tweets y 2321 columnas, es decir el total de trigramas únicos que están presentes en todo el dataset y que están presentes en al menos 2 documentos (min\_df=2).

Como se muestra en la tabla 20, al igual que con los unigramas y bigramas, la mayoría de los elementos en el vector de características son 0, por lo que se genera una matriz dispersa. Esto se debe a que el número de trigramas que aparecen en cada documento (tweet) es solo una pequeña parte de todo el conjunto de documentos; por lo tanto, hay muchos trigramas que no aparecen y se marcaran como 0.

- **Cálculo de los valores IDF**

Se calculó los valores IDF en base a los trigramas que se obtuvieron anteriormente. En la tabla 21 se muestra algunos valores IDF que se obtuvieron mediante un marco de datos de Python, estos valores se presentan ordenados de forma ascendente.

**Tabla 21.** Previa de los valores idf para trigramas del conjunto de datos

	<b>valores_idf</b>
<b>yo dar ansiedad</b>	5.768540
<b>yo dar cuenta</b>	6.408199
<b>no yo gustar</b>	6.782892
<b>no poder dormir</b>	6.813664
<b>yo dar depresion</b>	6.845413
...	...
<b>huele bicho cabron</b>	9.180788
<b>hpta no poder</b>	9.180788
<b>hpta fumarme tabaco</b>	9.180788
<b>hoy yo llamar</b>	9.180788
<b>you can do</b>	9.180788

Como se observa en la tabla 21, los trigramas “yo dar depresion” y “no poder dormir” tienen valores IDF más bajos en comparación con otros trigramas. Esto quiere decir que estos trigramas aparecen en más documentos de la colección, ya que cuanto menor sea el valor IDF del trigramas, menos única será para cualquier documento en particular.

- **Cálculo de las puntuaciones TF-IDF**

Una vez que se calculó los valores IDF para trigramas, se calculan las puntuaciones tf-idf para cualquier documento (tweet) o conjunto de documentos.

Se realizó el cálculo de las puntuaciones tf-idf para todos los documentos del conjunto de datos. En la tabla 22 se presenta una previa de la representación de valores tf-idf para trigramas de un documento (tweet) que se colocó en un marco de datos para fines de visualización y se ordenaron sus puntajes en orden descendente.

**Tabla 22.** Valores tf-idf (trigramas) de un tweet del conjunto de datos

	<b>tfidf</b>
<b>no hacer ejercicio</b>	0.529491
<b>hacer ejercicio deber</b>	0.529491
<b>peso no yo</b>	0.529491
<b>no yo dar</b>	0.398647
<b>persona sufrir ansiedad</b>	0.000000
...	...
<b>grave ver influencers</b>	0.000000
<b>grito desesperado no</b>	0.000000
<b>guayabo moral mañana</b>	0.000000
<b>gustar no yo</b>	0.000000
<b>you can do</b>	0.000000

En la tabla 22 se observa que solo ciertos trigramas tienen puntaje tf-idf, ya que son solo los que aparecen en este documento en particular. Todos los trigramas de este documento tienen una puntuación tf-idf y todo lo demás aparece como ceros, ya que cuanto más común sea el trigramma en todos los documentos, menor será su puntuación; por ejemplo, en este caso el trigramma “no yo dar” tiene una puntuación menor en comparación a los demás trigramas, esto quiere decir que este trigramma está presente en más documentos. Además, cuanto más exclusivo sea un trigramma en el actual documento (por ejemplo, "no hacer ejercicio" y "hacer ejercicio deber"), mayor será la puntuación.

Se obtuvo las variables independientes y dependientes aplicando el sobremuestreo con SMOTE, para la variable independiente se usó el vector de características tf-idf obtenido y para la variable dependiente se usó la etiqueta sentimiento (0 para no depresivo y 1 para depresivo) del dataset.

Finalmente, las variables independientes y dependiente obtenidas se almacenaron con el nombre de “x\_tfidf\_Trigrama” y “y\_tfidf\_Trigrama” respectivamente para uso posterior en el entrenamiento y prueba de los modelos.

En el repositorio Github<sup>11</sup> se presenta todo el proceso de extracción de características para unigramas, bigramas y trigramas realizado en Jupyter notebook, para garantizar la reproducibilidad, trazabilidad y transparencia del TT.

### **6.2.3. Tarea: Detección de sentimiento**

#### **Fase 4. Minería de texto y construcción de hipótesis**

Los vectores de características con N-gramas obtenidos en la tarea de extracción de características se usaron para realizar el entrenamiento con los 3 modelos de Machine learning, es decir, se usaron los vectores con unigramas, bigramas y trigramas para entrenar el modelo de Máquinas de Soporte Vectorial, Random Forest y Naive Bayes, y así determinar cuál de los 3 ofrece el mejor rendimiento en cada uno de los modelos.

Además, debido a la gran cantidad de tweets depresivos identificados en el preprocesamiento, se plantea la siguiente hipótesis: En tiempos de covid-19 existe un aumento en la cantidad de publicaciones depresivas en Twitter.

---

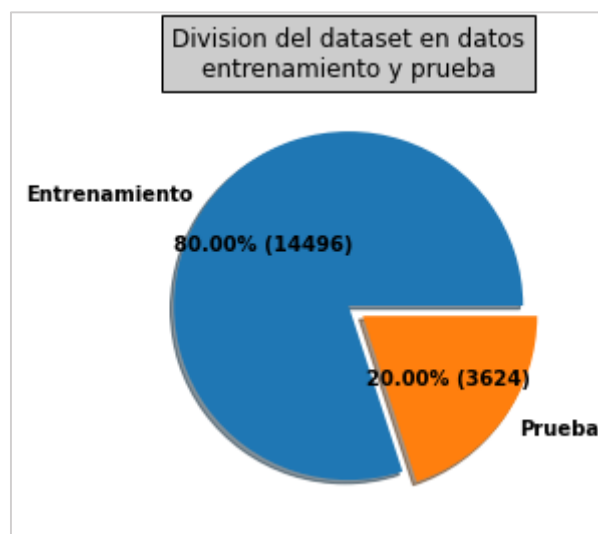
<sup>11</sup> Código en repositorio de la extracción de características:  
[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/blob/main/3.Extraccion\\_caracteristicas/ExtraccionCaracteristicas.ipynb](https://github.com/byronmb/Identificacion_Depresion_Ecuador/blob/main/3.Extraccion_caracteristicas/ExtraccionCaracteristicas.ipynb)

### 6.2.3.1. Modelo de Máquinas de Soporte Vectorial (SVM)

- **Máquinas de Soporte Vectorial con Unigramas**

Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Unigrama”, y el archivo “y\_tfidf\_Unigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 20. Dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 20.** División de los datos en entrenamiento y prueba para el modelo SVM con Unigramas

A través de la librería Scikit-Learn se empleó el algoritmo Máquinas de Soporte Vectorial para la creación del modelo usando kernel = “linear” como hiperparámetro, ya que las características en clasificación de texto en general se organizan en categorías linealmente separables (ver sección 4.3.2.1.1).

Así mismo, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (ver sección 4.3.2.3) a los datos de entrenamiento con 5 pliegues (K=5), cuyos resultados se muestran en la tabla 23.



**Tabla 23.** Validación cruzada del modelo SVM con unigramas.

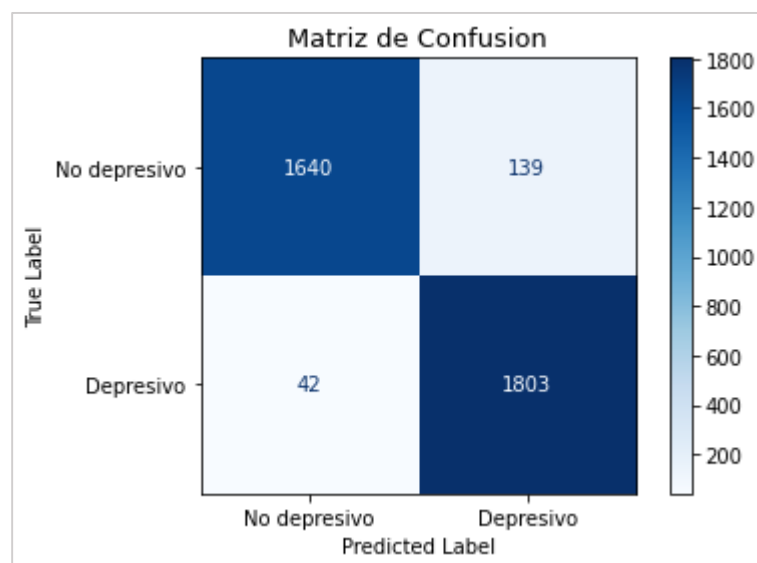
	<b>Exactitud (Accuracy)</b>
<b>K=5</b>	0.94896552
	0.94377372
	0.95584684
	0.94756813
	0.94963781
<b>Exactitud media:</b>	<b>0.949158</b>
<b>Desviación estándar:</b>	<b>0.003912</b>

El resultado de la validación cruzada da una precisión media de 95%, dicho resultado debería ser similar para el conjunto de prueba en caso de que los datos estén particionados adecuadamente, esto se detalla a continuación.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 24 y la figura 21 respectivamente.

**Tabla 24.** Reporte de clasificación para el modelo SVM con unigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.98	0.92	0.95	1779
<b>1 (Contenido depresivo)</b>	0.93	0.98	0.95	1845
<b>Accuracy (exactitud)</b>			0.95	3624
<b>Macro avg</b>	0.95	0.95	0.95	3624
<b>Weighted avg</b>	0.95	0.95	0.95	3624



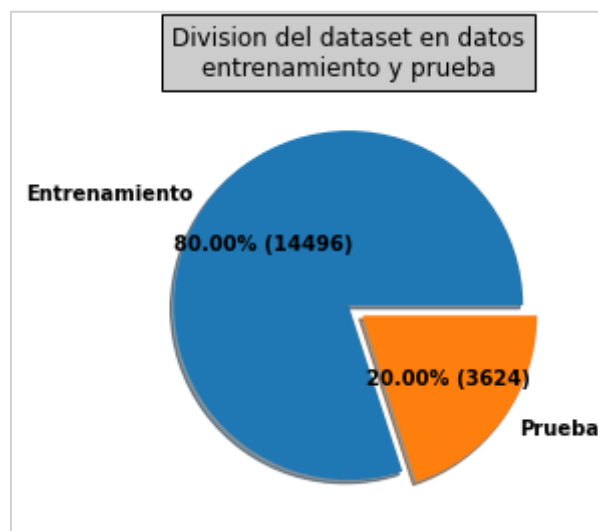
**Figura 21.** Matriz de confusión del modelo SVM con unigramas.

En la tabla 24 se observa que la exactitud en la predicción con el conjunto de datos de prueba es de 95%, y se comprueba que los resultados están muy apegados a la exactitud media de la validación cruzada que es igualmente 95%; por lo tanto, el modelo actual es confiable para identificar contenido depresivo y no depresivo. En cuanto a la matriz de confusión reflejada en la figura 21, se observa el número de publicaciones que fueron predichos de forma correcta (diagonal principal), y se distingue que la cantidad de valores respecto a los que no fueron clasificados de forma correcta por el modelo está acorde al porcentaje reflejado en el reporte de clasificación (ver tabla 24).

- **Máquinas de Soporte Vectorial con Bigramas**

Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Bigrama”, y el archivo “y\_tfidf\_Bigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 22. Dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 22.** División de los datos en entrenamiento y prueba para el modelo SVM con Bigramas.

De la misma forma que se hizo para crear el modelo SVM con Unigramas, se usó la librería Scikit-Learn para emplear el algoritmo Máquinas de Soporte Vectorial a fin de crear el modelo usando kernel = “linear” como hiperparámetro. Así mismo, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (ver sección 4.3.2.3) a los datos de entrenamiento con 5 pliegues (K=5), cuyos resultados se muestran en la tabla 25.

**Tabla 25.** Validación cruzada del modelo SVM con bigramas.

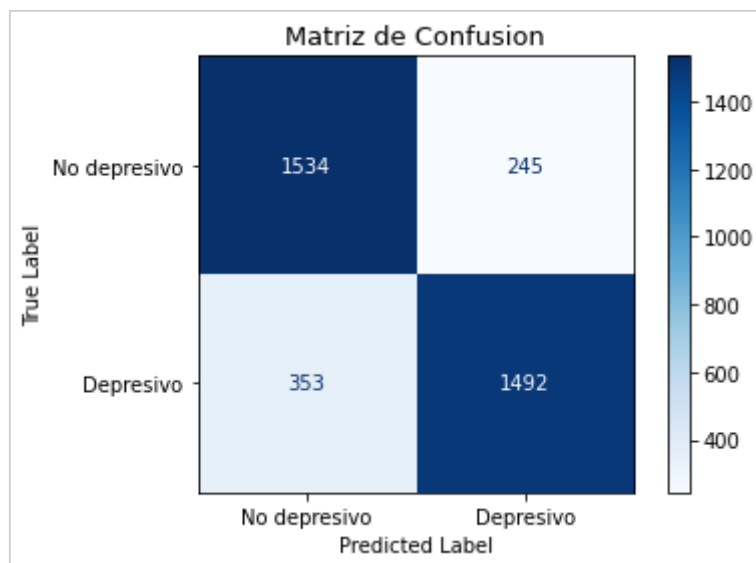
	<b>Exactitud (Accuracy)</b>
<b>K=5</b>	0.81965517
	0.82166264
	0.81993791
	0.82476716
	0.8361504
<b>Exactitud media:</b>	<b>0.824435</b>
<b>Desviación estándar:</b>	<b>0.006134</b>

El resultado de la validación cruzada da una precisión media de 82%, donde se aprecia que este modelo tuvo un rendimiento bastante menor respecto a los resultados del modelo con unigramas presentado anteriormente; sin embargo, se continúa ejecutando las siguientes pruebas para validar los resultados obtenidos.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 26 y la figura 23 respectivamente.

**Tabla 26.** Reporte de clasificación para el modelo SVM con bigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.81	0.86	0.84	1779
<b>1 (Contenido depresivo)</b>	0.86	0.81	0.83	1845
<b>Accuracy (exactitud)</b>			0.83	3624
<b>Macro avg</b>	0.84	0.84	0.83	3624
<b>Weighted avg</b>	0.84	0.83	0.83	3624



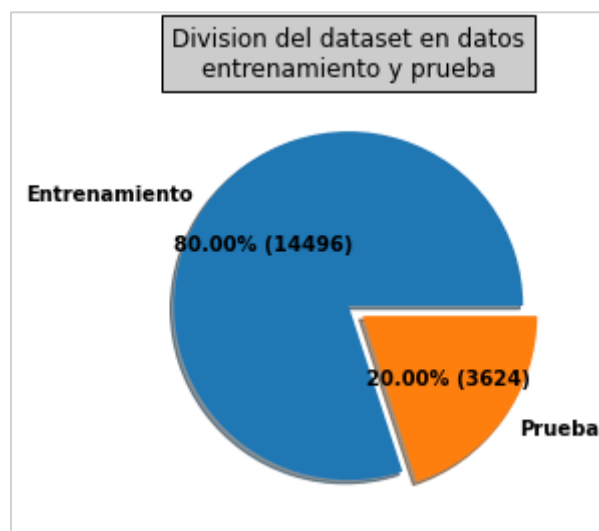
**Figura 23.** Matriz de confusión del modelo SVM con bigramas.

En la tabla 26 se ve que la exactitud en la predicción con el conjunto de datos de prueba es 83%, y se comprueba que los resultados están muy apegados a la exactitud media de la validación cruzada que es 82%, por lo tanto, los resultados son confiables; sin embargo, los porcentajes obtenidos son bajos para ser aceptables en comparación con los resultados obtenidos con unigramas, esto se ve más claramente en la matriz de confusión reflejada en la figura 23, en donde se observa que la cantidad de valores predichos de forma correcta (diagonal principal) es más baja, y la cantidad de valores que no fueron clasificados correctamente (diagonal secundaria) es más alta en comparación con los valores de la matriz de confusión con unigramas (ver figura 21).

- **Máquinas de Soporte Vectorial con Trigramas**

Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Trigrama”, y el archivo “y\_tfidf\_Trigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 24. Dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 24.** División de los datos en entrenamiento y prueba para el modelo SVM con Trigramas

De la misma forma que se hizo para crear el modelo SVM con Unigramas y Bigramas, se usó la librería Scikit-Learn para emplear el algoritmo Máquinas de Soporte Vectorial a fin de crear del modelo usando kernel = “linear” como hiperparámetro.

Así mismo, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (ver sección

4.3.2.3) a los datos de entrenamiento con 5 pliegues (K=5), cuyos resultados se presentan en la tabla 27.

**Tabla 27.** Validación cruzada del modelo SVM con trigramas.

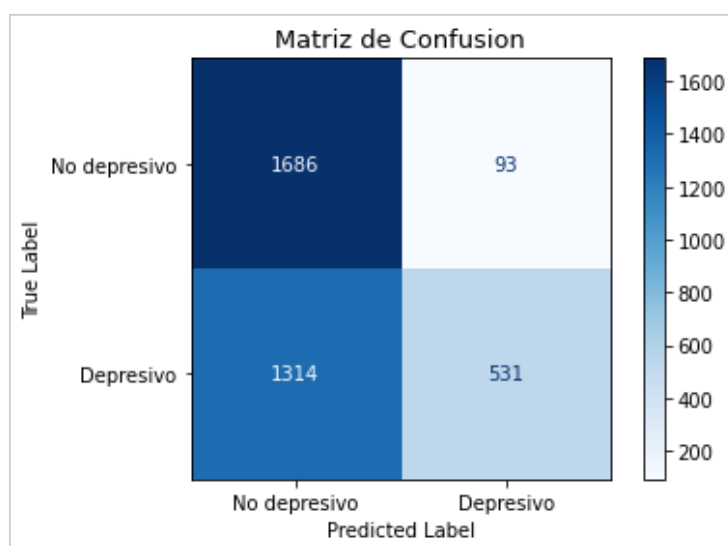
	<b>Exactitud (Accuracy)</b>
<b>K=5</b>	0.63
	0.61952397
	0.61710935
	0.6115902
	0.6264229
<b>Exactitud media:</b>	<b>0.620929</b>
<b>Desviación estándar:</b>	<b>0.006576</b>

El resultado de la validación cruzada da una precisión media de 62%, donde se aprecia que este modelo tuvo un rendimiento bastante menor respecto a los resultados del modelo con unigramas y bigramas presentado anteriormente; sin embargo, se continúa ejecutando las siguientes pruebas para validar los resultados obtenidos.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 28 y la figura 25 respectivamente.

**Tabla 28.** Reporte de clasificación para el modelo SVM con trigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.56	0.95	0.71	1779
<b>1 (Contenido depresivo)</b>	0.85	0.29	0.43	1845
<b>Accuracy (exactitud)</b>			0.61	3624
<b>Macro avg</b>	0.71	0.62	0.57	3624
<b>Weighted avg</b>	0.71	0.61	0.57	3624

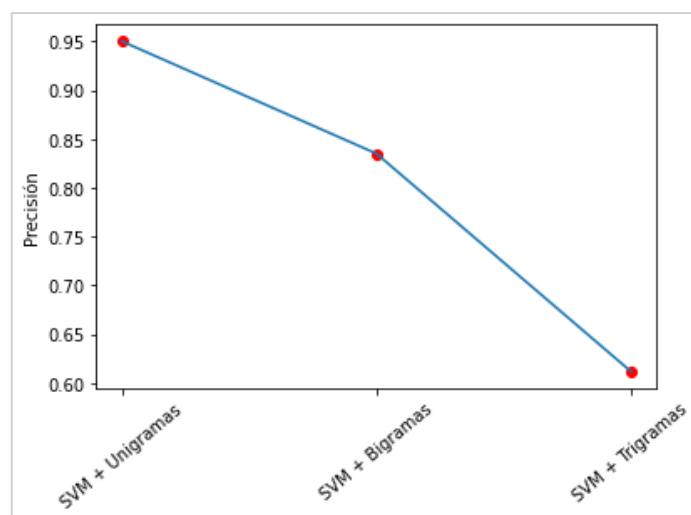


**Figura 25.** Matriz de confusión del modelo SVM con trigramas.

En la tabla 28 se ve que la exactitud en la predicción con el conjunto de datos de prueba es 61%, y se comprueba que los resultados están muy apegados a la exactitud media de la validación cruzada que es 62%, por lo tanto, los resultados son confiables; sin embargo, los porcentajes obtenidos son bastante bajos para ser aceptables, incluso más que el resultado obtenido en la predicción con los bigramas mostrado anteriormente. Los resultados se ven más claramente en la matriz de confusión reflejada en la figura 25, en donde se observa que la cantidad de valores predichos de forma correcta (diagonal principal) es incluso menor en comparación con las que no fueron clasificadas correctamente (diagonal secundaria), sobre todo en el caso del contenido clasificado como depresivo.

- **Rendimiento de los 3 modelos de Máquinas de Soporte Vectorial**

Para una mejor comprensión, se presenta la comparación en la figura 26 del rendimiento de los 3 modelos SVM con unigramas, bigramas y con trigramas a través del porcentaje obtenido en cada uno de ellos.



**Figura 26.** Comparación del rendimiento de los 3 modelos SVM.

Como se observa en la figura 26, el rendimiento de los 3 modelos de Máquinas de Soporte Vectorial (SVM) va decreciendo en su exactitud de forma casi lineal de acuerdo a su uso con unigramas, bigramas y trigramas respectivamente, siendo el modelo SVM con unigramas el que mejor porcentaje de rendimiento ofrece; por lo tanto, a este modelo con mejor rendimiento se lo exportó para uso posterior con el nombre “modelo\_SVM\_Unigram.pkl” mediante el uso de la librería Joblib (ver sección 6.1.1.4.8). Para garantizar la transparencia del trabajo, todo el proceso de entrenamiento y prueba del modelo Máquinas de Soporte Vectorial con unigramas, bigramas y trigramas se muestra en mayor detalle en el repositorio<sup>12</sup>.

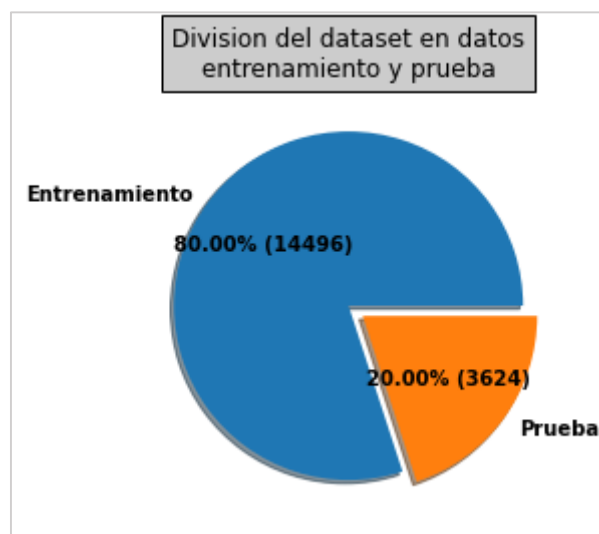
<sup>12</sup> Código en repositorio del entrenamiento y prueba con Support Vector Machines: [https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/blob/main/Support\\_Vector\\_Machines.ipynb](https://github.com/byronmb/Identificacion_Depresion_Ecuador/blob/main/Support_Vector_Machines.ipynb)

### 6.2.3.2. Modelo de Random Forest

- **Random Forest con Unigramas**

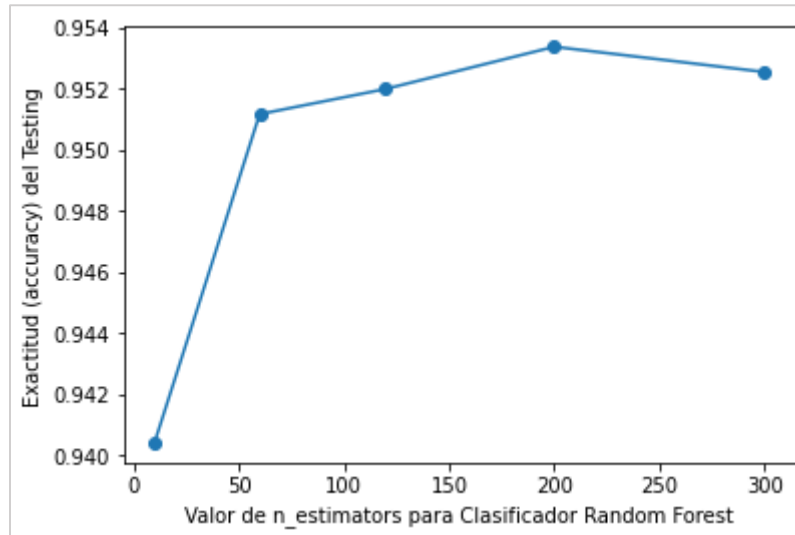
Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Unigrama”, y el archivo “y\_tfidf\_Unigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 27. Dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 27.** División de los datos en entrenamiento y prueba para el modelo RF con Unigramas.

Para entrenar el modelo Random Forest, fue necesario en primer lugar establecer la cantidad óptima de árboles de decisión para el modelo. De acuerdo a [85], una mayor cantidad de árboles solo aumenta su costo computacional y no tiene una ganancia significativa en el rendimiento y sugiere que un rango entre 64 y 128 árboles es suficiente para obtener resultados satisfactorios; no obstante, se ejecutó el modelo con 5 valores distintos para el número de árboles que están alrededor de la cantidad recomendada para comprobar el rendimiento con el conjunto de datos que se está usando, tal como se muestra en la figura 28.



**Figura 28.** Rendimiento del modelo RF con unigramas en distintas cantidades de árboles.

Como se observa en la figura 28, el modelo Random Forest con unigramas logra un mejor rendimiento en el conjunto de prueba cuando el número de árboles está alrededor de 200; por lo tanto, a través de la librería Scikit-Learn se empleó el algoritmo Random Forest para la creación del modelo usando el hiperparámetro “n\_estimators=200”.

Además, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (sección 4.4.3) a los datos de entrenamiento con 5 pliegues (K=5), cuyos resultados se muestran en la tabla 29.

**Tabla 29.** Validación cruzada del modelo RF con unigramas.

	<b>Exactitud (Accuracy)</b>
<b>K=5</b>	0.95068966
	0.94687823
	0.95343222
	0.95239738
	0.954812
<b>Exactitud media:</b>	<b>0.951642</b>
<b>Desviación estándar:</b>	<b>0.002736</b>

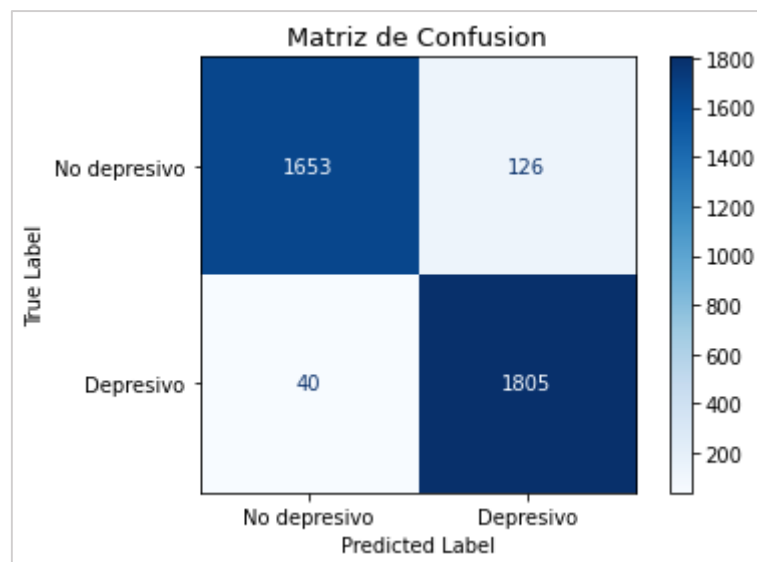
El resultado de la validación cruzada da una precisión media de 95%, dicho resultado debería ser similar para el conjunto de prueba en caso de que los datos estén particionados adecuadamente, esto se detalla a continuación.



Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se muestran en la tabla 30 y la figura 29 respectivamente.

**Tabla 30.** Reporte de clasificación para el modelo RF con unigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.98	0.93	0.95	1779
<b>1 (Contenido depresivo)</b>	0.93	0.98	0.96	1845
<b>Accuracy (exactitud)</b>			0.95	3624
<b>Macro avg</b>	0.96	0.95	0.95	3624
<b>Weighted avg</b>	0.96	0.95	0.95	3624



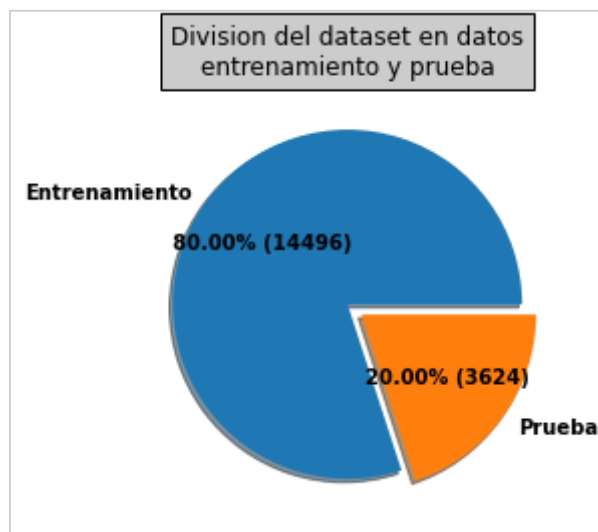
**Figura 29.** Matriz de confusión del modelo RF con unigramas.

En la tabla 30 se observa que la exactitud en la predicción con el conjunto de datos de prueba es de 95%, y se comprueba que los resultados están muy apegados a la exactitud media de la validación cruzada que es igualmente 95%, por lo tanto, el modelo es confiable para identificar contenido depresivo y no depresivo. En cuanto a la matriz de confusión reflejada en la figura 29, se observa el número de publicaciones que fueron predichos de forma correcta (diagonal principal), y se distingue que la cantidad de valores respecto a los que no fueron clasificados de forma correcta por el modelo está acorde al porcentaje reflejado en el reporte de clasificación (tabla 30).

- **Random Forest con Bigramas**

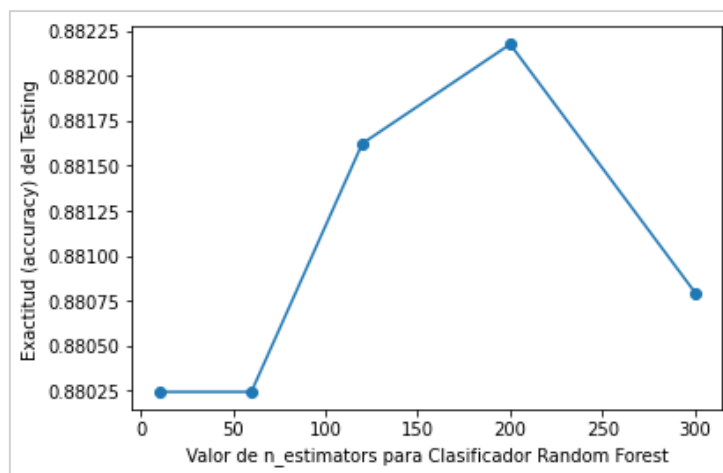
Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Bigrama”, y el archivo “y\_tfidf\_Bigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 30. Dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 30.** División de los datos en entrenamiento y prueba para el modelo RF con Bigramas.

Para entrenar el modelo Random Forest, al igual que en el caso del modelo con unigramas, primero se buscó determinar la cantidad óptima de árboles de decisión para el modelo, por ende, se ejecutó el modelo con 5 valores distintos para el número de árboles para comprobar el rendimiento con el conjunto de datos que se está usando, tal como se muestra en la figura 31.



**Figura 31.** Rendimiento del modelo RF con bigramas en distintas cantidades de árboles.

Como se observa en la figura 31, el modelo Random Forest con bigramas logra un mejor rendimiento en el conjunto de prueba cuando el número de árboles está alrededor de 200; por lo tanto, a través de la librería Scikit-Learn se empleó el algoritmo Random Forest para la creación del modelo usando el hiperparámetro “n\_estimators=200”.

Además, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (sección 4.4.3) a los datos de entrenamiento con 5 pliegues (K=5), cuyos resultados se muestran en la tabla 31.

**Tabla 31.** Validación cruzada del modelo RF con bigramas.

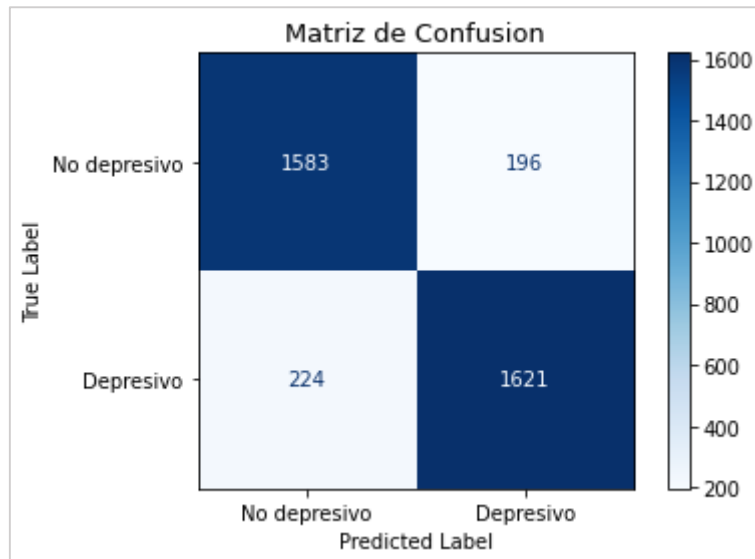
	<b>Exactitud (Accuracy)</b>
<b>K=5</b>	0.88517241
	0.87685409
	0.88168334
	0.88202829
	0.88237323
<b>Exactitud media:</b>	<b>0.881622</b>
<b>Desviación estándar:</b>	<b>0.002686</b>

El resultado de la validación cruzada da una precisión media de 88%, donde se aprecia que este modelo tuvo un rendimiento bastante menor respecto a los resultados del modelo con unigramas presentado anteriormente; sin embargo, se continúa ejecutando las siguientes pruebas para validar los resultados obtenidos.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 32 y la figura 32 respectivamente.

**Tabla 32.** Reporte de clasificación para el modelo RF con bigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.88	0.89	0.88	1779
<b>1 (Contenido depresivo)</b>	0.89	0.88	0.89	1845
<b>Accuracy (exactitud)</b>			0.88	3624
<b>Macro avg</b>	0.88	0.88	0.88	3624
<b>Weighted avg</b>	0.88	0.88	0.88	3624



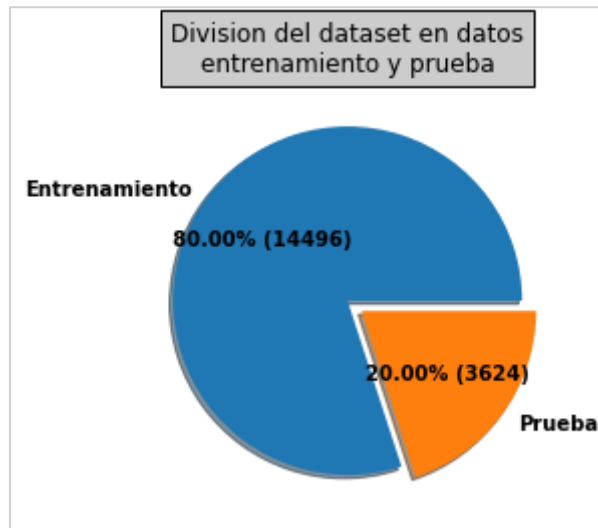
**Figura 32.** Matriz de confusión del modelo RF con bigramas.

En la tabla 32 se ve que la exactitud en la predicción con el conjunto de datos de prueba es 88%, y se comprueba que los resultados están muy apegados a la exactitud media de la validación cruzada que es igualmente 88%; por lo tanto, los resultados son confiables; sin embargo, los porcentajes obtenidos son bajos para ser aceptables en comparación con los resultados obtenidos con unigramas, esto se ve más claramente en la matriz de confusión reflejada en la figura 32, en donde se observa que la cantidad de valores predichos de forma correcta (diagonal principal) es más baja, y la cantidad de valores que no fueron clasificados correctamente (diagonal secundaria) es más alta en comparación con los valores de la matriz de confusión con unigramas (ver figura 29).

- **Random Forest con Trigramas**

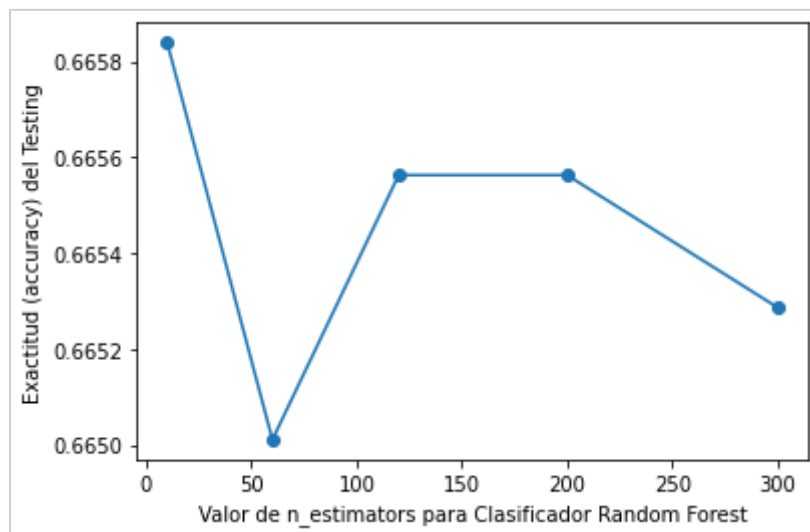
Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Trigrama”, y el archivo “y\_tfidf\_Trigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 33, dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 33.** División de los datos en entrenamiento y prueba para el modelo RF con Trigramas.

De la misma forma que se hizo para crear el modelo RF con Unigramas y Bigramas, primero se buscó determinar la cantidad óptima de árboles de decisión para el modelo, por ende, se ejecutó el modelo con 5 valores distintos para el número de árboles para comprobar el rendimiento con el conjunto de datos que se está usando, tal como se muestra en la figura 34.



**Figura 34.** Rendimiento del modelo RF con trigramas en distintas cantidades de árboles.

Como se observa en la figura 34, el modelo Random Forest con trigramas tiene una variación casi insignificante con las distintas cantidades de árboles en las que se ejecutó, sin embargo, logra un rendimiento un poco más alto cuando el número de árboles está alrededor de 10; por lo tanto, a través de la librería Scikit-Learn, se empleó el algoritmo Random Forest para la creación del modelo usando el hiperparámetro "n\_estimators=10".

Así mismo, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (ver sección

4.3.2.3) a los datos de entrenamiento con 5 pliegues (K=5), cuyos resultados se muestran en la tabla 33.

**Tabla 33.** Validación cruzada del modelo RF con trigramas.

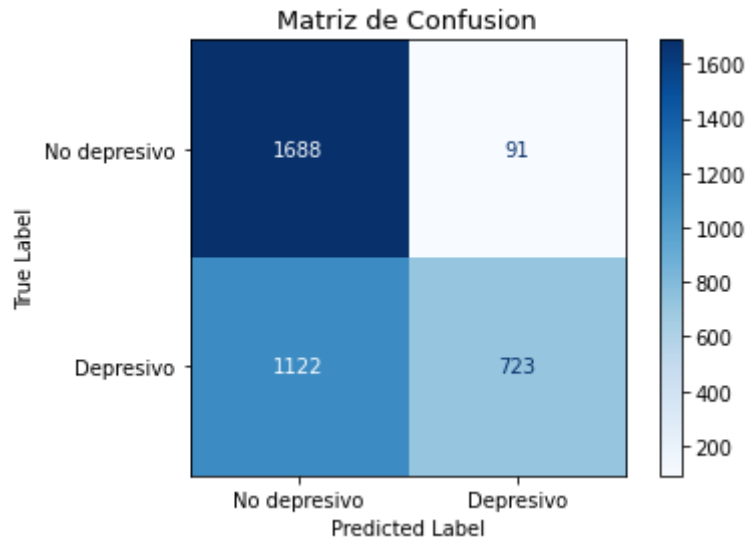
	<b>Exactitud (Accuracy)</b>
<b>K=5</b>	0.66965517
	0.66367713
	0.66747154
	0.65643325
	0.67747499
<b>Exactitud media:</b>	<b>0.666942</b>
<b>Desviación estándar:</b>	<b>0.006924</b>

El resultado de la validación cruzada da una precisión media de 66%, donde se aprecia que este modelo tuvo un rendimiento bastante menor respecto a los resultados del modelo con unigramas y bigramas presentado anteriormente; sin embargo, se continúa ejecutando las siguientes pruebas para validar los resultados obtenidos.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 34 y la figura 35 respectivamente.

**Tabla 34.** Reporte de clasificación para el modelo RF con trigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.60	0.95	0.74	1779
<b>1 (Contenido depresivo)</b>	0.89	0.39	0.54	1845
<b>Accuracy (exactitud)</b>			0.67	3624
<b>Macro avg</b>	0.74	0.67	0.64	3624
<b>Weighted avg</b>	0.75	0.67	0.64	3624

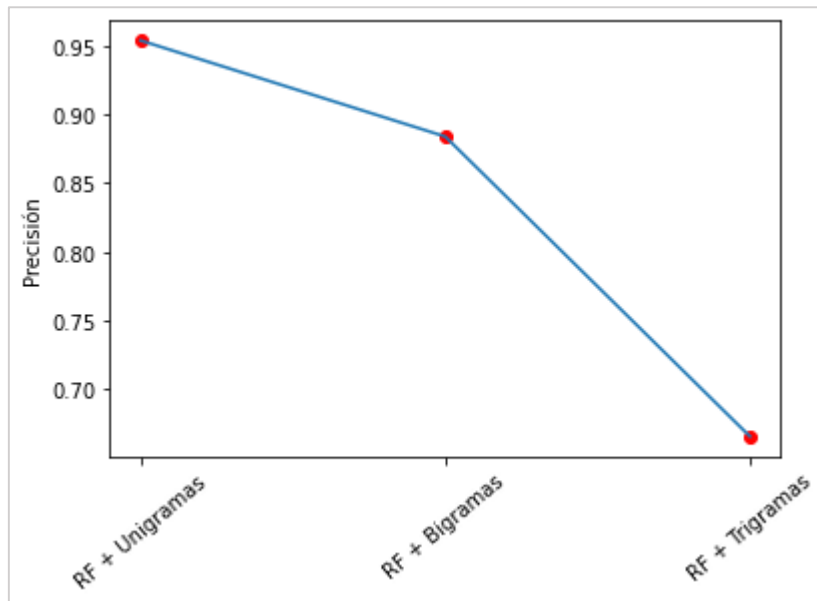


**Figura 35.** Matriz de confusión del modelo RF con trigramas.

En la tabla 34 se ve que la exactitud en la predicción con el conjunto de datos de prueba es 67%, y se comprueba que los resultados están muy apegados a la exactitud media de la validación cruzada que es 66%, por lo tanto, los resultados son confiables; sin embargo, los porcentajes obtenidos son bastante bajos para ser aceptables, incluso siendo más bajos que el resultado obtenido en la predicción en el modelo con bigramas mostrado anteriormente. Los resultados se ven más claramente en la matriz de confusión reflejada en la figura 35, en donde se observa que la cantidad de valores predichos de forma correcta (diagonal principal) es menor en comparación con las que no fueron clasificados correctamente (diagonal secundaria), sobre todo en el caso del contenido clasificado como depresivo.

- **Rendimiento de los 3 modelos de Random Forest.**

Para una mejor comprensión, se presenta la comparación en la figura 36 del rendimiento de los 3 modelos Random Forest con unigramas, bigramas y trigramas a través del porcentaje obtenido en cada uno de ellos.



**Figura 36.** Comparación del rendimiento de los 3 modelos Random Forest.

Como se muestra en la figura 36, el rendimiento de los 3 modelos de Random Forest (RF) va decreciendo en su exactitud de acuerdo a su uso con unigramas, bigramas y trigramas respectivamente, siendo el modelo RF con unigramas el que mejor porcentaje de rendimiento ofrece; por lo tanto, a este modelo con mejor rendimiento se lo exportó para uso posterior con el nombre “modelo\_RF\_Unigram.pkl” mediante el uso de la librería Joblib. (ver sección 6.1.1.4.8).

Para garantizar la transparencia y reproducibilidad del trabajo, todo el proceso de entrenamiento y prueba del modelo Random Forest con unigramas, bigramas y trigramas se presenta con mayor detalle en el repositorio<sup>13</sup>.

### 6.2.3.3. Modelo de Naive Bayes

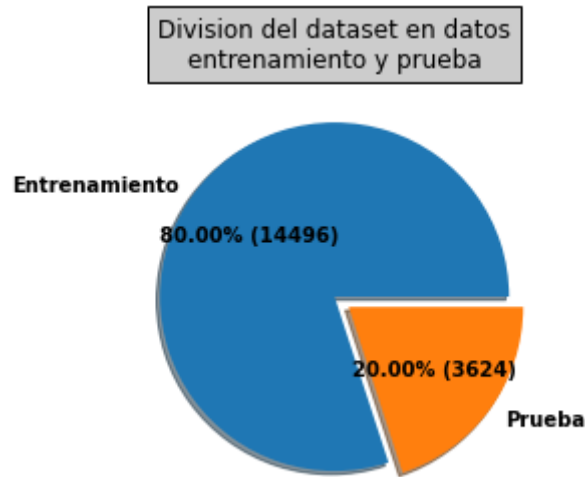
- **Naive Bayes con Unigramas**

Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Unigrama”, y el archivo “y\_tfidf\_Unigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 37, dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).

<sup>13</sup> Código en repositorio del entrenamiento y prueba con Random Forest:  
[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/blob/main/Random\\_Forest.ipynb](https://github.com/byronmb/Identificacion_Depresion_Ecuador/blob/main/Random_Forest.ipynb)





**Figura 37.** División de los datos en entrenamiento y prueba para el modelo NB con Unigramas.

A través de la librería Scikit-Learn, se empleó el algoritmo Naive Bayes para la creación del modelo. Además, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (ver sección 4.3.2.3) a los datos de entrenamiento con 5 pliegues ( $K=5$ ), cuyos resultados se muestran en la tabla 35.

**Tabla 35.** Validación cruzada del modelo NB con unigramas.

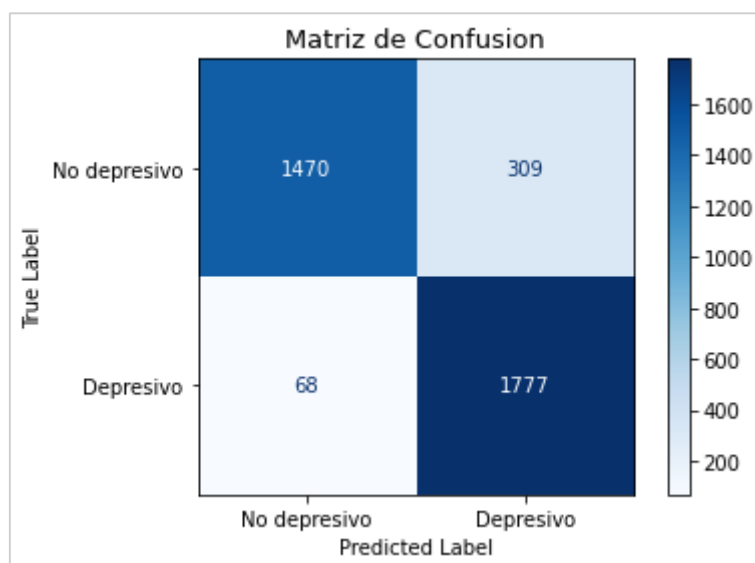
	Exactitud (Accuracy)
<b>K=5</b>	0.89310345
	0.87478441
	0.89582615
	0.88030355
	0.8851328
<b>Exactitud media:</b>	<b>0.885830</b>
<b>Desviación estándar:</b>	<b>0.007821</b>

El resultado de la validación cruzada da una precisión media de 88%, dicho resultado debería ser similar para el conjunto de prueba en caso de que los datos estén particionados adecuadamente, esto se detalla a continuación.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 36 y la figura 38 respectivamente.

**Tabla 36.** Reporte de clasificación para el modelo NB con unigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.96	0.83	0.89	1779
<b>1 (Contenido depresivo)</b>	0.85	0.96	0.90	1845
<b>Accuracy (exactitud)</b>			0.90	3624
<b>Macro avg</b>	0.90	0.89	0.90	3624
<b>Weighted avg</b>	0.90	0.90	0.90	3624



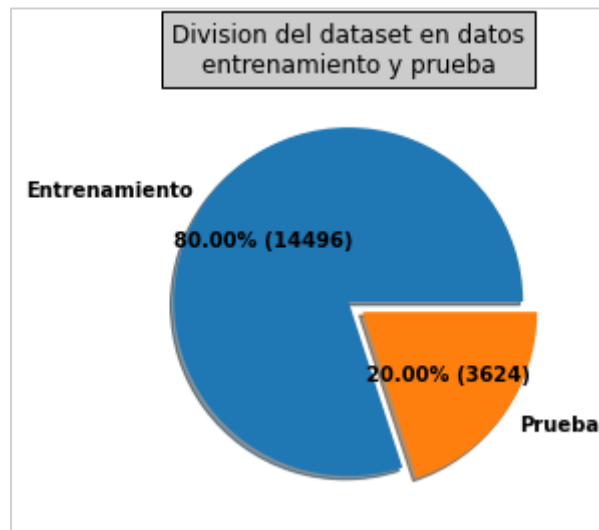
**Figura 38.** Matriz de confusión del modelo NB con unigramas.

En la tabla 36 se observa que la exactitud en la predicción con el conjunto de datos de prueba es de 90%, y se comprueba que los resultados están muy apegados a la exactitud media de la validación cruzada que es 88%; por lo tanto, el modelo es confiable para identificar contenido depresivo y no depresivo. En cuanto a la matriz de confusión reflejada en la figura 38, se observa el número de publicaciones que fueron predichos de forma correcta (diagonal principal), y se distingue que la cantidad de valores respecto a los que no fueron clasificados de forma correcta por el modelo está acorde al porcentaje reflejado en el reporte de clasificación (tabla 36).

- **Naive Bayes con Bigramas**

Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Bigrama”, y el archivo “y\_tfidf\_Bigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 39, dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 39.** División de los datos en entrenamiento y prueba para el modelo NB con Bigramas.

De igual manera que se hizo para crear el modelo NB con Unigramas, se usó la librería Scikit-Learn para emplear el algoritmo Naive Bayes a fin de crear el modelo. Así mismo, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (ver sección 4.3.2.3) a los datos de entrenamiento con 5 pliegues ( $K=5$ ), cuyos resultados se muestran en la tabla 37.

**Tabla 37.** Validación cruzada del modelo NB con bigramas.

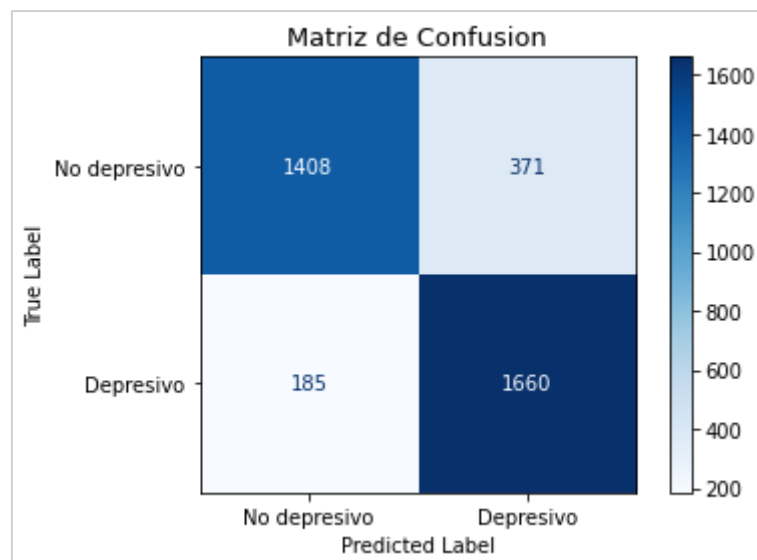
	Exactitud (Accuracy)
<b>K=5</b>	0.82
	0.81407382
	0.82373232
	0.82511211
	0.82718179
<b>Exactitud media:</b>	<b>0.822020</b>
<b>Desviación estándar:</b>	<b>0.004612</b>

El resultado de la validación cruzada da una precisión media de 82%, donde se aprecia que este modelo tuvo un rendimiento menor respecto a los resultados del modelo con unigramas presentado anteriormente, sin embargo, se continúa ejecutando las siguientes pruebas para validar los resultados obtenidos.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 38 y la figura 40 respectivamente.

**Tabla 38.** Reporte de clasificación para el modelo NB con bigramas.

	<b>Precisión</b>	<b>Recall</b>	<b>F1-Score</b>	<b>Support</b>
<b>0 (Contenido no depresivo)</b>	0.88	0.79	0.84	1779
<b>1 (Contenido depresivo)</b>	0.82	0.90	0.86	1845
<b>Accuracy (exactitud)</b>			0.85	3624
<b>Macro avg</b>	0.85	0.85	0.85	3624
<b>Weighted avg</b>	0.85	0.85	0.85	3624



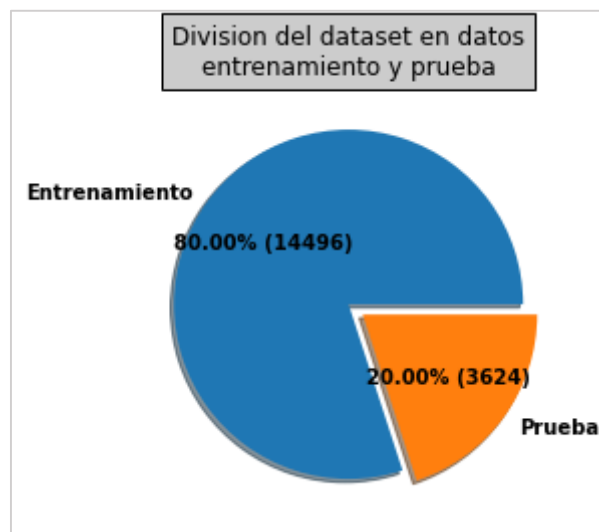
**Figura 40.** Matriz de confusión del modelo NB con bigramas.

En la tabla 38 se ve que la exactitud en la predicción con el conjunto de datos de prueba es 85%, y se comprueba que los resultados están cercanos a la exactitud media de la validación cruzada que es 82%, por lo tanto, los resultados son confiables; sin embargo, los porcentajes obtenidos son bajos para ser aceptables en comparación con los resultados obtenidos con unigramas, esto se ve más claramente en la matriz de confusión reflejada en la figura 40, en donde se observa que la cantidad de valores predichos de forma correcta (diagonal principal) es más baja, y la cantidad de valores que no fueron clasificados correctamente (diagonal secundaria) es más alta en comparación con los valores de la matriz de confusión con unigramas (ver figura 38).

- **Naive Bayes con Trigramas**

Para la ejecución del presente algoritmo, se importó el vector de características “x\_tfidf\_Trigrama”, y el archivo “y\_tfidf\_Trigrama” para obtener las etiquetas (sentimiento) del conjunto de datos. Estos archivos representan a las variables independientes y dependientes respectivamente.

Los datos se dividieron en entrenamiento y prueba en proporción de 80% para entrenamiento y 20% para prueba como se muestra en la figura 41. Dichas muestras se tomaron de forma aleatoria mediante la librería Scikit-Learn (ver sección 6.1.1.4.7).



**Figura 41.** División de los datos en entrenamiento y prueba para el modelo NB con Trigramas.

De la misma forma que se hizo para crear el modelo NB con Unigramas y Bigramas, se usó la librería Scikit-Learn para emplear el algoritmo Naive Bayes a fin de crear el modelo. Así mismo, para garantizar que el rendimiento del modelo sea independiente de la partición entre datos de entrenamiento y prueba, se utilizó la técnica de validación cruzada (ver sección 4.3.2.3) a los datos de entrenamiento con 5 pliegues (K=5), cuyos resultados se muestran en la tabla 39.

**Tabla 39.** Validación cruzada del modelo NB con trigramas.

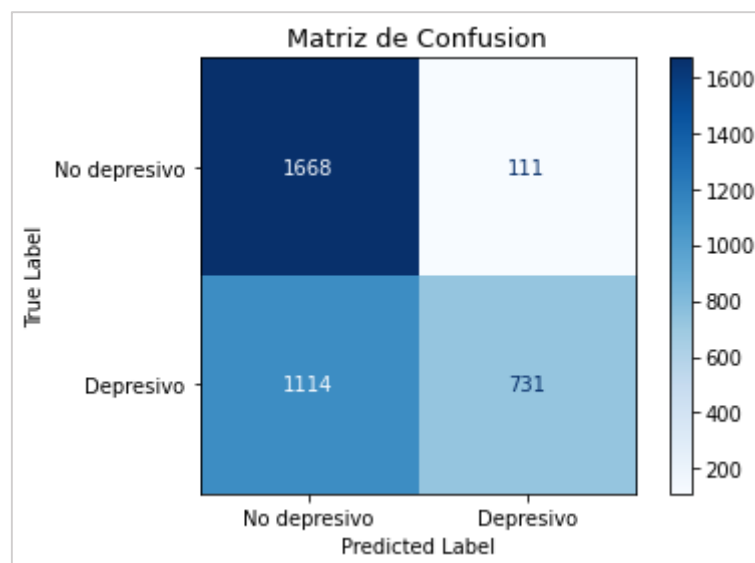
	<b>Exactitud (Accuracy)</b>
<b>K=5</b>	0.67137931
	0.6664367
	0.66367713
	0.65746809
	0.67506037
<b>Exactitud media:</b>	<b>0.666804</b>
<b>Desviación estándar:</b>	<b>0.006102</b>

El resultado de la validación cruzada da una precisión media de 66%, donde se aprecia que este modelo tuvo un rendimiento bastante menor respecto a los resultados del modelo con unigramas y bigramas presentado anteriormente, sin embargo, se continúa ejecutando las siguientes pruebas para validar los resultados obtenidos.

Se generó el reporte de clasificación para medir la calidad de las predicciones con el conjunto de datos de prueba, además de la respectiva matriz de confusión, éstos se observan en la tabla 40 y la figura 42 respectivamente.

**Tabla 40.** Reporte de clasificación para el modelo NB con trigramas.

	Precisión	Recall	F1-Score	Support
<b>0 (Contenido no depresivo)</b>	0.60	0.94	0.73	1779
<b>1 (Contenido depresivo)</b>	0.87	0.40	0.54	1845
<b>Accuracy (exactitud)</b>			0.66	3624
<b>Macro avg</b>	0.73	0.67	0.64	3624
<b>Weighted avg</b>	0.74	0.66	0.64	3624



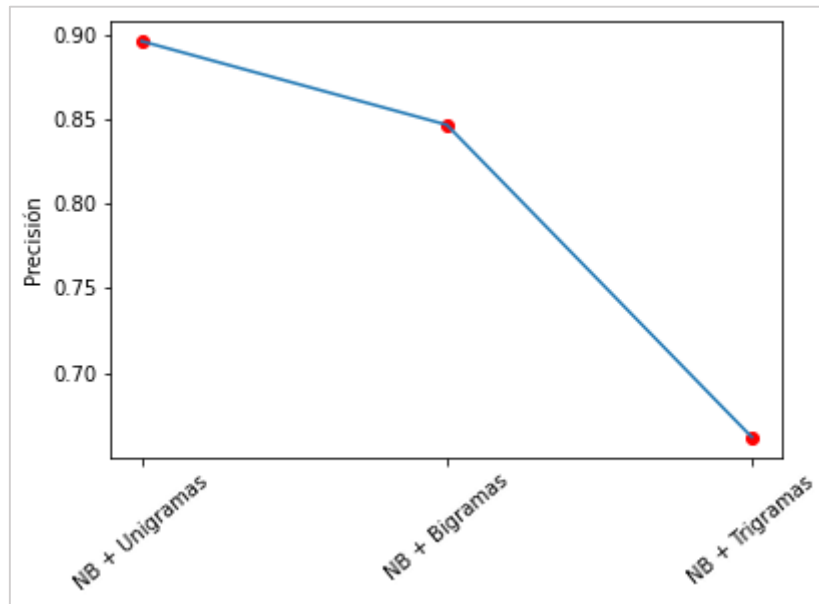
**Figura 42.** Matriz de confusión del modelo NB con trigramas.

En la tabla 40 se ve que la exactitud en la predicción con el conjunto de datos de prueba es 66%, y se comprueba que los resultados están cercanos a la exactitud media de la validación cruzada que es igualmente 66%, por lo tanto, los resultados son confiables; sin embargo, los porcentajes obtenidos son bastante bajos para ser aceptables, incluso más que el resultado obtenido en la predicción con los bigramas mostrado anteriormente. Los resultados se ven más claramente en la matriz de confusión reflejada en la figura 42, en donde se observa que la cantidad de valores predichos de forma correcta (diagonal principal) es incluso menor en

comparación con las que no fueron clasificados correctamente (diagonal secundaria), sobre todo en el caso del contenido clasificado como depresivo.

- **Rendimiento de los 3 modelos de Naive Bayes.**

Para una mejor comprensión, se presenta la comparación en la figura 43 del rendimiento de los 3 modelos Naive Bayes con unigramas, bigramas y trigramas a través del porcentaje obtenido en cada uno de ellos.



**Figura 43.** Comparación del rendimiento de los 3 modelos Naive Bayes.

Como se observa en la figura 43, el rendimiento de los 3 modelos Naive Bayes (NB) va decreciendo en su exactitud de forma significativa de acuerdo a su uso con unigramas, bigramas y trigramas respectivamente, siendo el modelo Naive Bayes con unigramas el que mejor porcentaje de rendimiento ofrece; por lo tanto, a este modelo con mejor rendimiento se lo exportó para uso posterior con el nombre “modelo\_NB\_Unigram.pkl” mediante el uso de la librería Joblib. (ver sección 6.1.1.4.8).

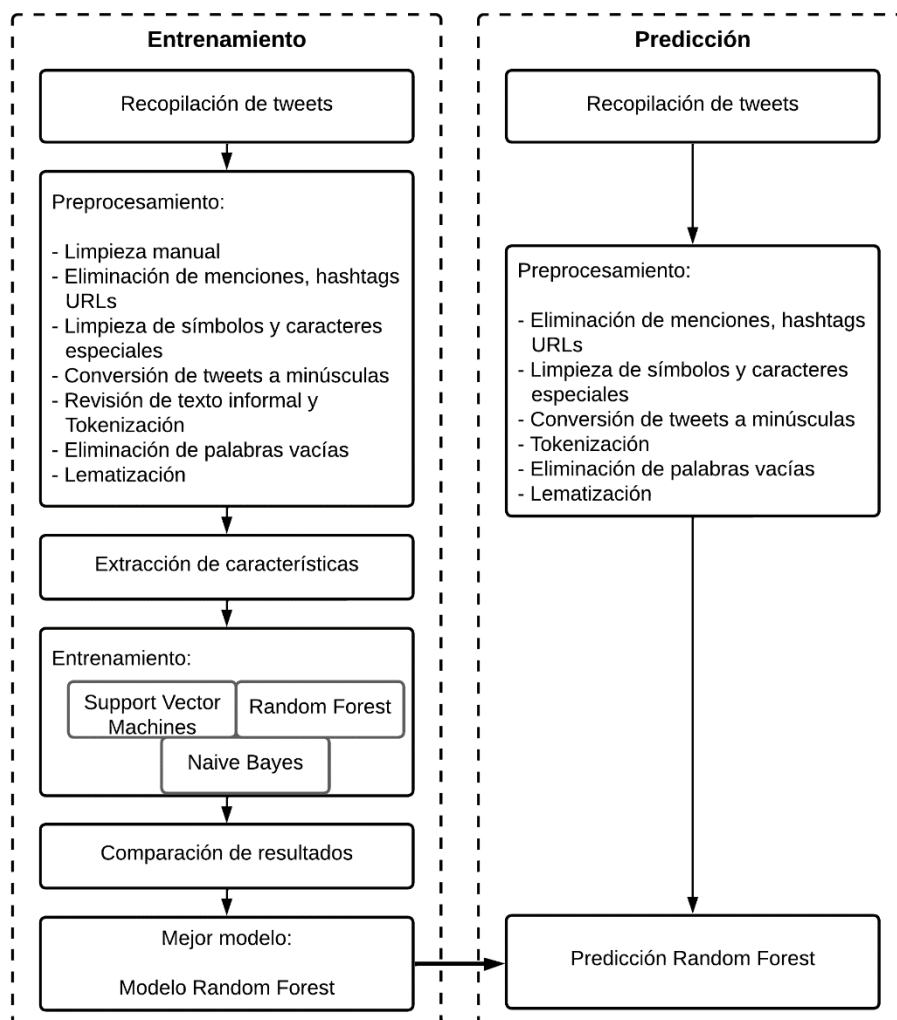
Para asegurar la transparencia del trabajo, todo el proceso de entrenamiento y prueba del modelo Naive Bayes con unigramas, bigramas y trigramas se presenta con mayor detalle en el repositorio<sup>14</sup>.

<sup>14</sup> Código en repositorio del entrenamiento y prueba con Naive Bayes:  
[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/blob/main/Naive\\_Bayes.ipynb](https://github.com/byronmb/Identificacion_Depresion_Ecuador/blob/main/Naive_Bayes.ipynb)

### 6.3. Objetivo 3: Interpretar los resultados obtenidos en el análisis de sentimientos

#### Fase 5. Interpretación/evaluación

En esta sección se realizó el análisis en base a los mejores modelos que fueron obtenidos mediante el entrenamiento en el objetivo 2, usando las métricas de rendimiento y conforme a la cantidad de predicciones realizadas por cada uno de los modelos, además se realizó una predicción a los tweets del 2019 mediante el mejor modelo para finalmente realizar un análisis univariado y bivariado a los conjuntos de datos obtenidos. En la figura 44 se representa el proceso de análisis de sentimientos en base a la metodología kdt (ver figura 11), agregado la comparación de los modelos, así como el proceso de predicción. Los detalles de la comparación y predicción mostrados en esta figura se muestran en detalle en las secciones posteriores.



**Figura 44.** Diagrama de flujo del proceso de entrenamiento de modelos y predicción con los nuevos datos.



### 6.3.1. Tarea: Evaluar el desempeño de los algoritmos mediante métricas de precisión, accuracy, recall y F1 Score.

Se utilizó el modelo con mejor rendimiento de acuerdo a cada tipo de algoritmo, es decir el mejor modelo de Máquinas de soporte vectorial, Random Forest y Naive Bayes para comparar el rendimiento de estos 3 modelos. Cabe recalcar que los modelos con mejor rendimiento para identificar contenido depresivo fueron los que se entrenaron con unigramas y que se obtuvieron en el objetivo 2.

A través de la aplicación de estos 3 modelos de algoritmos de clasificación, se realizó la predicción con el conjunto de datos de prueba, en la figura 45 se muestran la cantidad de tweets del conjunto de datos de prueba por cada clase (depresivo y no depresivo). Cabe recalcar que es el mismo conjunto de datos de prueba para los 3 modelos, por lo que se observa que tienen la misma cantidad de valores. Mientras que en la figura 46 se muestra la cantidad de tweets que fueron predichos por el modelo de Máquinas de soporte vectorial, Random Forest y Naive Bayes respectivamente.

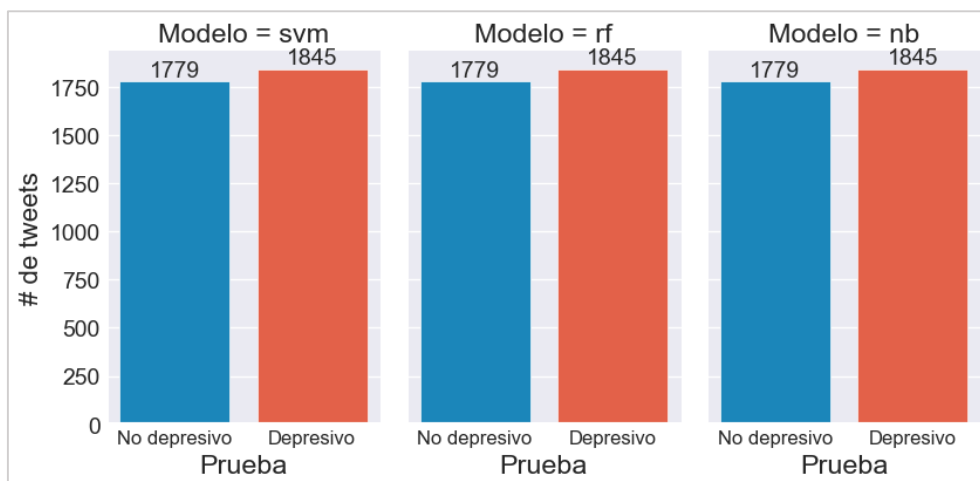


Figura 45. Cantidad de tweets del conjunto de datos de prueba por cada clase.

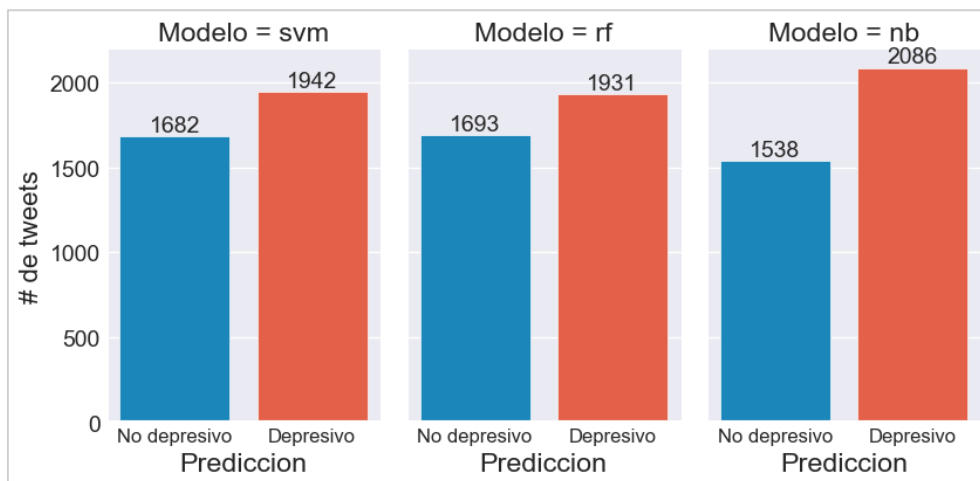
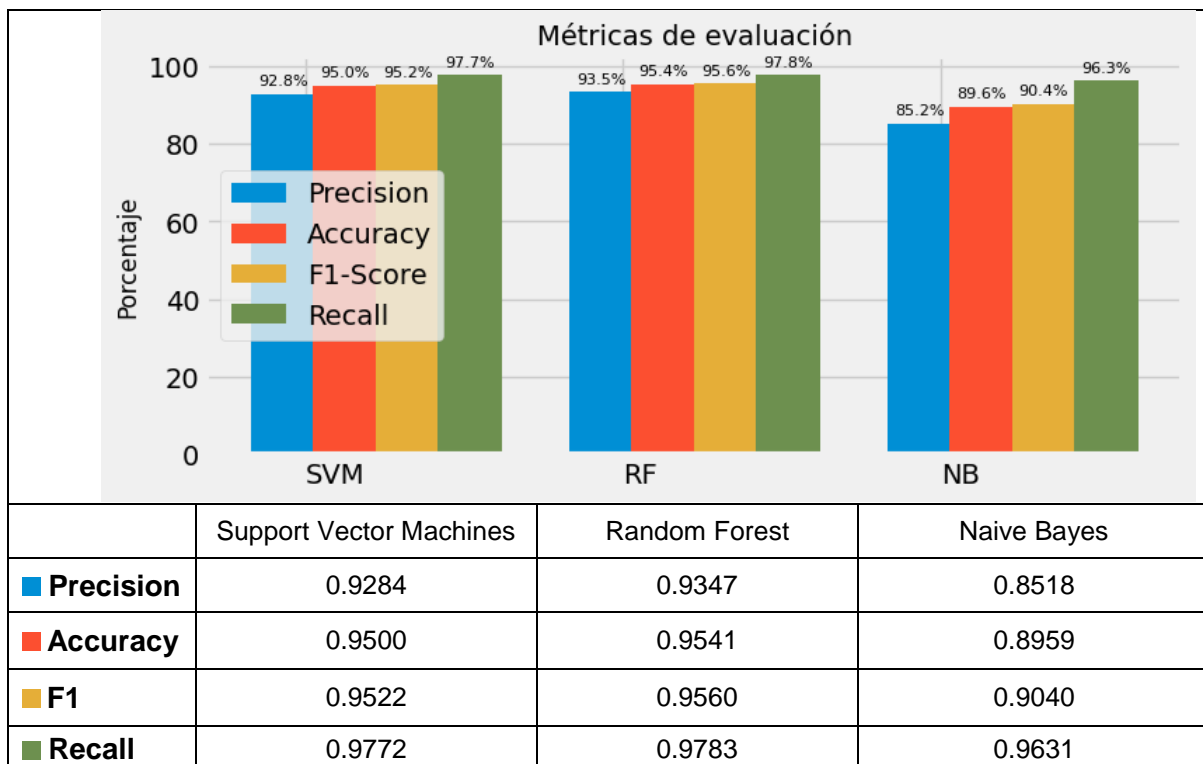


Figura 46. Cantidad de tweets predichos por cada modelo.

En la figura 46 se observa que la cantidad de valores predichos por los modelos de Máquinas de soporte vectorial y Random Forest respecto al conjunto de prueba (figura 45) es bastante cercano, mientras que en el modelo Naive Bayes se ve una variabilidad más notable en la cantidad de datos predichos.

Para tener una mejor perspectiva del rendimiento de los 3 modelos, en la tabla 41 se presenta una comparación del rendimiento de cada uno de ellos utilizando las métricas de Precisión, Exactitud (Accuracy), Puntuación F1 y Recall.

**Tabla 41.** Comparación de los modelos finales mediante métricas de evaluación.



Como se muestra en la tabla 41, el modelo de Random Forest supera a los otros 2 modelos en todas las métricas mostradas, con valores:

- Precisión = 93.4%
- Exactitud (accuracy) = 95.4%
- Puntuación F1 = 95.6%
- Recall = 97.8%

El modelo de Máquinas de Soporte Vectorial (Support Vector Machines) está detrás con una diferencia relativamente cercana, estos valores son:

- Precisión = 92.8%
- Exactitud (accuracy) = 95%

- Puntuación F1 = 95.2%
- Recall = 97.7%

El modelo Naive Bayes tiene el rendimiento más bajo de los 3 modelos, pero aún con valores relativamente altos, estos son:

- Precisión = 85.2%
- Exactitud (accuracy) = 89.5%
- Puntuación F1 = 90.4%
- Recall = 96.3%

En base al análisis de las métricas de evaluación mostradas, el modelo Random Forest fue seleccionado como el mejor para la clasificación de contenido depresivo, siguiéndole muy de cerca el modelo de Máquinas de Soporte Vectorial, y finalmente el modelo Naive Bayes como el que menor rendimiento tiene entre los 3 modelos.

Para garantizar la reproducibilidad, trazabilidad y transparencia del trabajo, la comparación de los modelos entrenados se la realizó en Jupyter y está disponible para su revisión detallada en el repositorio<sup>15</sup>.

### **6.3.2. Tarea: Predecir contenido depresivo con tweets prepandemia**

A través del modelo que tuvo mejor rendimiento para identificar contenido depresivo en twitter y que fue identificado en la tarea anterior, se realizó la predicción para tweets que se han publicado previo al inicio del COVID-19, específicamente del año 2019. Para esto, se extrajeron tweets de todo el año 2019 utilizando las mismas características que se emplearon para extraer los tweets depresivos usados para el entrenamiento (ver sección 6.1.3), obteniendo un total de 12125 tweets.

Los tweets extraídos se preprocesaron siguiendo el mismo procedimiento realizado para el preprocesamiento de los datos que fueron usados en el entrenamiento de los modelos (ver sección 6.2.1) con excepción de la limpieza manual, esto con el fin de que se pueda usar con distintos datos nuevos sin necesidad de realizar procedimientos manuales. Con los tweets preprocesados se creó un nuevo vector tf-idf usando el vocabulario guardado de los unigramas (ver sección 6.2.2.1) para que el nuevo vector creado tenga la misma longitud de características que tienen los datos entrenados. Cabe mencionar que se usó el vocabulario de los unigramas porque es el que se utilizó para el modelo que tuvo mejor rendimiento.

---

<sup>15</sup> Código en repositorio de la comparación de los modelos:  
[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/blob/main/4.Comparacion\\_y\\_Analisis\\_Modelos/1.Comparacion\\_Analisis\\_Mejores\\_Modelos.ipynb](https://github.com/byronmb/Identificacion_Depresion_Ecuador/blob/main/4.Comparacion_y_Analisis_Modelos/1.Comparacion_Analisis_Mejores_Modelos.ipynb)

Por último, se realizó la predicción de tweets depresivos utilizando el modelo con mejor rendimiento guardado con el nombre de “modelo\_RF\_Unigram”. Algunos tweets que fueron clasificados como depresivos por este modelo se presentan en la tabla 42.

**Tabla 42.** Tweets del 2019 clasificados como depresivos por el modelo RF.

<b>Tweet Original</b>	<b>Tweet Preprocesado</b>	<b>Predicción</b>
Ya les conté que la ansiedad me está matando ? Pues buej, aquí me tienen 😊	['conte', 'ansiedad', 'yo', 'matar', 'pues', 'buej', 'aquí', 'yo', 'cara_radiante_con_ojos_sonrientes']	Depresivo
No me aguanto este estrés y ansiedad que cargo.	['no', 'yo', 'aguantar', 'estre', 'ansiedad', 'cargo']	Depresivo
Siento como la hora de la depresión llega poco a poco	['sentir', 'hora', 'depresion', 'llegar']	Depresivo
A veces la depresión y la ansiedad no me deja ver lo que en realidad las personas hacen por mí.	['vez', 'depresion', 'ansiedad', 'no', 'yo', 'dejar', 'ver', 'realidad', 'persona', 'hacer']	Depresivo
Ni el psicólogo me ayuda con mi situación, de seguro y sale loco o sumergido en la depresión conmigo y me acompaña a llorar.	['ni', 'psicologo', 'yo', 'ayuda', 'situacion', 'seguro', 'salir', 'loco', 'sumergido', 'depresion', 'yo', 'yo', 'acompañar', 'llorar']	Depresivo

Los tweets clasificados como depresivos por el modelo se guardaron en un conjunto de datos con el nombre de “Tweets\_Depresivos\_2019\_Predichos”. Para asegurar la transparencia del trabajo, todo el proceso llevado a cabo para extraer y preprocesar los tweets del año 2019, además del proceso de clasificación de los tweets usando el modelo Random Forest se presenta con mayor detalle en el repositorio<sup>16</sup>.

### 6.3.3. Tarea: Realizar análisis univariado y bivariado de los datos para representar y comparar la cantidad de tweets depresivos.

Para el presente análisis se utilizó el dataset que contiene tweets depresivos y que fue previamente limpiado en la fase de preprocesamiento (ver sección 6.2.1), además del dataset que se obtuvo mediante la predicción con el modelo Random Forest (ver sección 6.3.2), para realizar la respectiva comparación. Ambos dataset tienen algunos atributos (columnas) en común que se crean al momento de la extracción de los tweets y que serán usados para realizar el análisis de éstos. En la tabla 43 se muestra un resumen de los campos que contienen ambos dataset.

<sup>16</sup> Código en repositorio de la extracción y preprocesamiento de tweets del 2019: [https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/tree/main/4.Comparacion\\_y\\_Analisis\\_Modelos](https://github.com/byronmb/Identificacion_Depresion_Ecuador/tree/main/4.Comparacion_y_Analisis_Modelos)

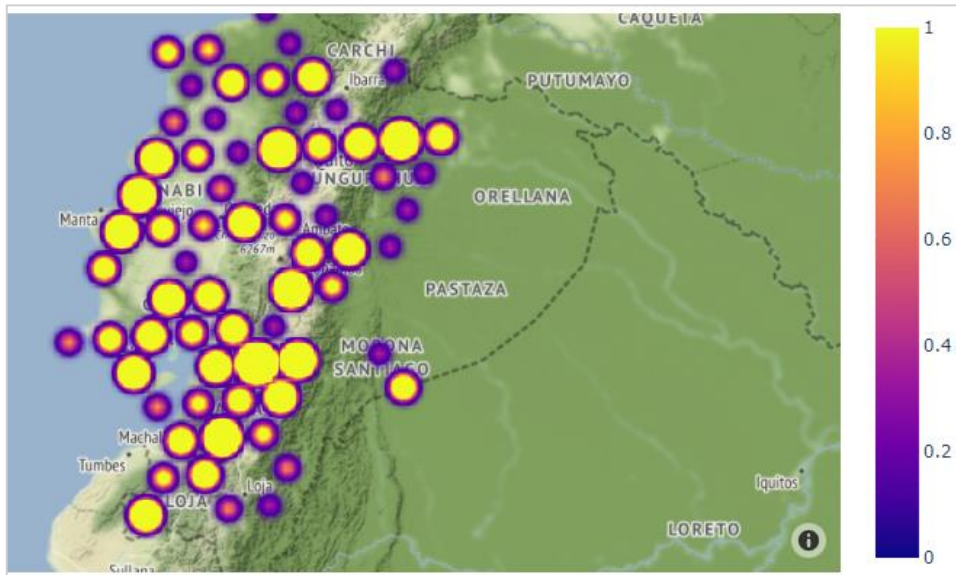
**Tabla 43.** Atributos del conjunto de datos con tweets depresivos.

Nombre de atributo	Descripción	Ejemplo
<b>Id</b>	Contiene un número identificador distinto para cada tweet	1339912665586610000
<b>Date</b>	Contiene la fecha de publicación del tweet	2020-12-18
<b>Time</b>	Contiene la hora de la publicación del tweet	07:37:28
<b>Tweet</b>	Contiene la redacción del tweet	a veces es válido sentirse agobiada desesperada y con una incertidumbre tenaz
<b>Geo</b>	Contiene las coordenadas geográficas de donde se publicó el tweet (latitud, longitud y radio en km)	-1.3695361279507605, -78.08139447223985, 27.98571148655604km

### Zonas geográficas donde hay más tweets depresivos

En primer lugar, mediante el dataset con tweets depresivos que se limpió en la fase de preprocesamiento (tweets en tiempos de COVID-19), se utilizó el atributo llamado “geo” que contiene la latitud, longitud y radio en kilómetros de la ubicación geográfica que fue usada para recolectar determinados tweets.

El atributo “geo” se separó en campos individuales que contienen Latitud, Longitud y Radio respectivamente para graficar las zonas en donde existen una mayor distribución de tweets depresivos en el Ecuador, tal como se muestra en la figura 47.



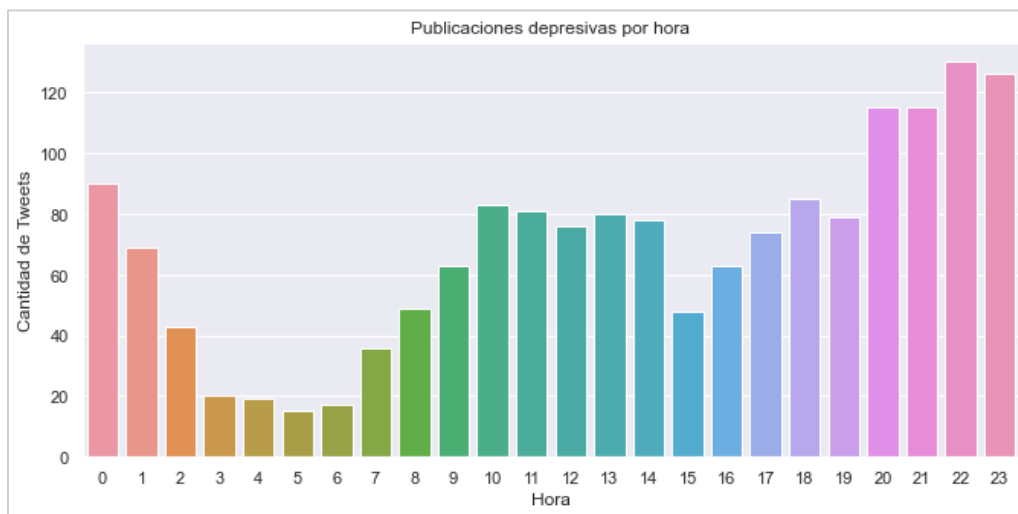
**Figura 47.** Mapa de calor de la distribución de tweets depresivos en Ecuador.

En la figura 47 se observa que hay una mayor cantidad de publicaciones depresivas en Twitter en la región Costa y Sierra, con presencia notoria en ciudades como Cuenca, Riobamba,

Quito, Manabí y Loja, mientras que en la región Amazónica existe una cantidad relativamente baja en cuanto a publicaciones, con cierta presencia evidente en la provincia de Morona Santiago. Cabe recalcar que las coordenadas geográficas están relacionadas a las que se usaron para extraer los tweets (ver anexo 3), por lo que existen varios tweets que tienen la misma ubicación geográfica.

### **Distribución de publicaciones depresivas en base a las horas del día**

En base al atributo “time” del dataset (ver tabla 43), se obtuvo la hora para conocer en qué horas del día existen una mayor o menor cantidad de tweets depresivos en el año 2020 y 2021, tal como se muestra en la figura 48.

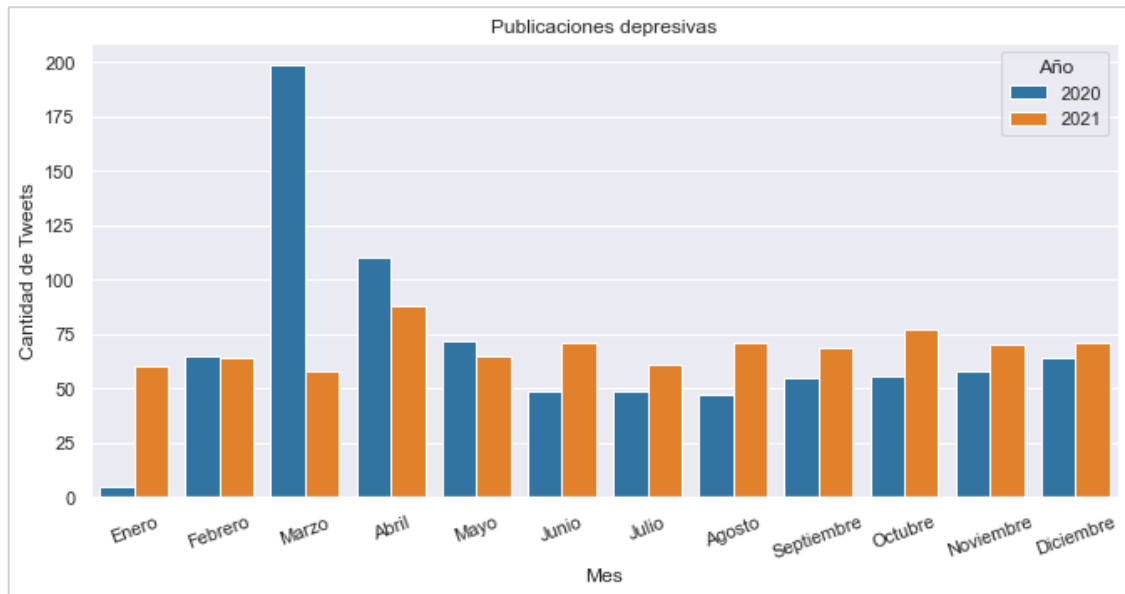


**Figura 48.** Distribución de publicaciones depresivas por hora.

Como se observa en la figura 48, las horas donde existen más publicaciones relacionadas a depresión son en la noche, teniendo una mayor cantidad entre las 20h00 y 00h00, mientras que en horas de la madrugada aproximadamente entre las 3h00 y 6h00 existe una menor cantidad de publicaciones depresivas.

### **Distribución de publicaciones depresivas en base a los meses de los años 2020 y 2021.**

En base al atributo “date” del dataset (ver tabla 43), se separó el campo para obtener los meses y años en campos individuales para graficar la distribución de publicaciones por cada mes de los años 2020 y 2021, como se muestra en la figura 49.



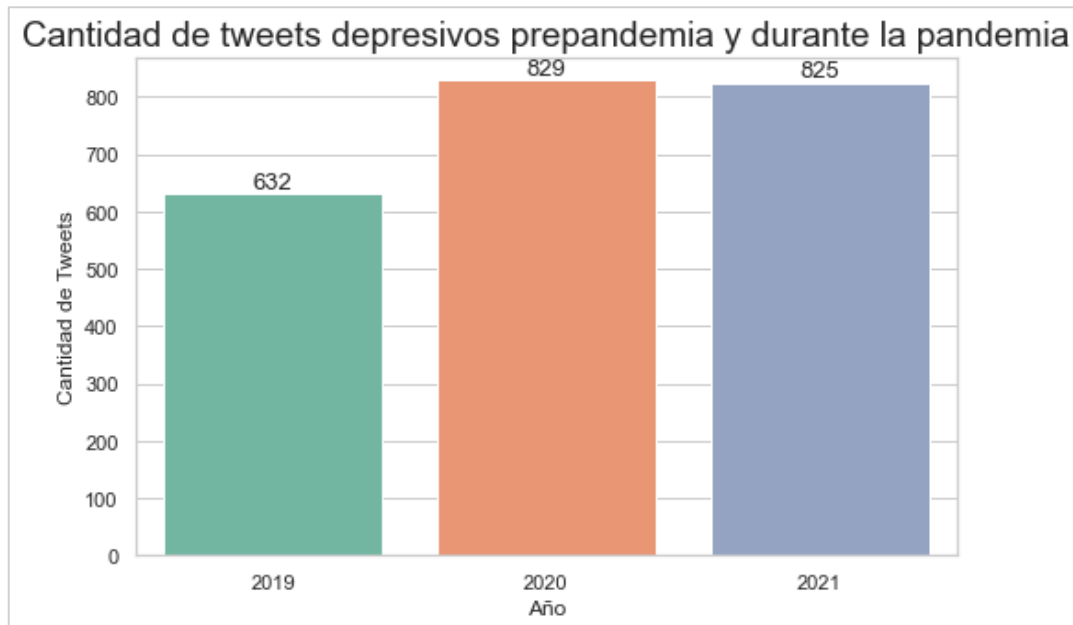
**Figura 49.** Publicaciones depresivas por meses de los años 2020 y 2021.

Como se observa en la figura 49, en el mes de marzo de 2020 es donde existe una mayor cantidad de publicaciones depresivas, esto concuerda con el inicio del confinamiento decretado en Ecuador por el coronavirus el 11 de marzo del 2020 [86], como segundo mes con mayor cantidad de publicaciones está el mes de abril del mismo año. Además, en los meses posteriores se puede ver una cantidad más baja pero relativamente constante en la cantidad de publicaciones sin presentar mayor variación. Sin embargo, se puede notar que en estos meses la cantidad de publicaciones del año 2021 son mayores que la cantidad de publicaciones del año 2020.

Cabe recalcar que en el mes de enero del 2020 no se muestra la representación de publicaciones ya que se recolectaron a partir del 30 de enero del 2020 (ver sección 6.1.2).

### **Variación porcentual en las publicaciones depresivas en prepandemia y durante la pandemia**

Para ver la variación porcentual en las publicaciones depresivas, se utilizó el dataset que se obtuvo mediante la predicción con el modelo Random Forest (ver sección 6.3.2) junto al dataset con tweets depresivos obtenido en la fase de preprocesamiento (ver sección 6.2.1) para comparar la cantidad de publicaciones en base a cada año. En la figura 50 se observa la cantidad de publicaciones depresivas del año 2019, 2020 y 2021 respectivamente.



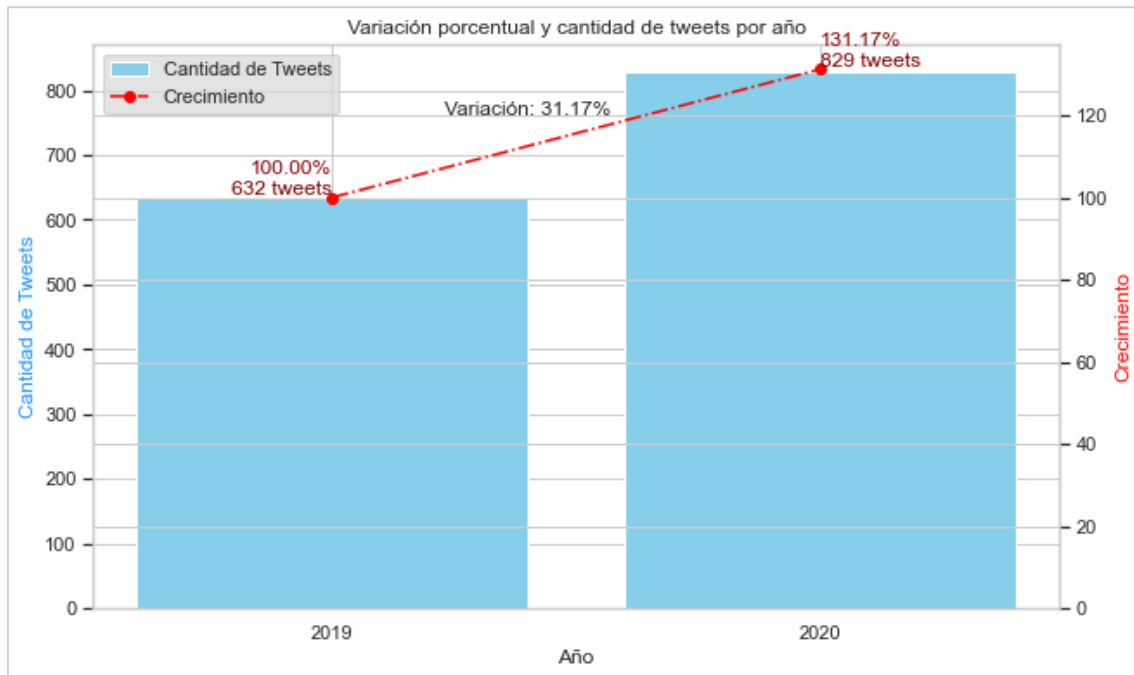
**Figura 50.** Cantidad de tweets depresivos de los años 2019, 2020 y 2021.

Como se observa en la figura 50, si se compara los 3 años, el 2019 (publicaciones prepandemia), 2020 y 2021 (publicaciones durante la pandemia), se puede notar que existe un incremento considerable en la cantidad de publicaciones depresivas realizadas en twitter en los años 2020 y 2021.

En el 2019 existe un total de 632 tweets, en el año 2021 existen 829 publicaciones y en el año 2021 hay un total de 825 publicaciones. Entre los años 2020 y 2021 podemos ver una diferencia relativamente baja en la cantidad de publicaciones depresivas, siendo el año 2020 el que tiene una cantidad un poco más alta. Sin embargo, si se compara el crecimiento desde el año 2019 hasta el 2020, existe una mayor diferencia en cuanto a la cantidad de publicaciones depresivas.

Para analizar con mayor precisión la diferencia en la cantidad de publicaciones depresivas entre el año 2019 y 2020, se presenta en la figura 51 la variación en términos porcentuales.





**Figura 51.** Variación porcentual en la cantidad de tweets entre 2019 y 2020.

Además, para respaldar los datos presentados en la figura 51, se utilizó la fórmula de variación porcentual que se muestra en la ecuación 10.

$$\frac{|\text{valor nuevo} - \text{valor antiguo}|}{|\text{valor antiguo}|} \times 100\%$$

$$\frac{829 - 632}{632} \times 100\% = 31.17\% \quad (10)$$

La hipótesis planteada en la fase de Minería de Texto y Construcción de Hipótesis (ver sección 6.2.3) de que existe un aumento en las publicaciones depresivas en Twitter durante tiempos de COVID-19, fue respaldada por el resultado mostrado en la figura 51 y respaldado a partir de la ecuación 10. Específicamente, se demostró que hubo un aumento del 31.17% en el número de publicaciones depresivas en Twitter en el año 2020 en comparación con el 2019. Este resultado sugiere que la pandemia COVID-19 ha tenido un impacto negativo en la salud mental de los usuarios de Twitter en Ecuador, lo que se refleja en el aumento de publicaciones de este tipo en la red social.

Para asegurar la transparencia del trabajo, todo el código utilizado para realizar este análisis comparativo, así como las gráficas resultantes, están disponibles en el repositorio de GitHub<sup>17</sup> para su consulta y revisión en detalle.

<sup>17</sup> Código en repositorio de la comparación y análisis de los modelos: [https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/tree/main/4.Comparacion\\_y\\_Analisis\\_Modelos](https://github.com/byronmb/Identificacion_Depresion_Ecuador/tree/main/4.Comparacion_y_Analisis_Modelos)

## 7. Discusión

En esta sección se analizan los resultados obtenidos en relación a los objetivos establecidos en el presente Trabajo de Titulación (TT). Estos resultados se desglosan por cada uno de los objetivos en concordancia con las fases de la metodología KDT y se destacan los aspectos más relevantes de cada proceso, incluyendo sus limitaciones y aportes comparativos con otros trabajos relacionados.

### 7.1. Objetivo 1: Construir un conjunto de datos a partir de las publicaciones de Twitter.

Los resultados de este objetivo permitieron construir los conjuntos de datos mediante las publicaciones en twitter desde el 30 de enero de 2020 hasta el 31 de diciembre del 2021, recopilando un total de 31978 tweets relacionados con depresión y 10000 tweets de carácter aleatorio. Inicialmente, se intentó utilizar la herramienta Tweepy para realizar el raspado o scraping de los tweets, sin embargo, ésta presenta algunas limitaciones debido a que utiliza la API de Twitter y esto conlleva que solo se puedan recopilar los últimos 3200 tweets, además de tener que crear una cuenta de desarrollador y autenticarse obligatoriamente. Eventualmente con la revisión de literatura se encontró la herramienta Twint [51] [87] [88], que permitió superar las limitaciones de Tweepy, sobre todo en la cantidad de publicaciones, permitiendo extraer todos los tweets relacionados en el rango de fecha establecido y cumpliendo satisfactoriamente con el propósito principal del objetivo.

Además de obtener la herramienta apropiada de twitter scraping, la revisión de literatura también permitió conocer la metodología más adecuada a usarse para cumplir adecuadamente con todo el proceso de análisis de sentimientos. Se descubrió que varias metodologías generales son planteadas de forma particular por cada autor en sus respectivos trabajos (ver sección 6.1.1.2), por lo que no existen muchas referencias en cuanto a una metodología en común que sea ampliamente utilizada. No obstante, en base a la investigación se seleccionó a la metodología KDT [62] [64] [89], puesto que sus distintas fases son muy similares a las metodologías generales propuestas en trabajos previos, y principalmente porque todas las actividades planteadas para este proyecto (ver figura 12) se adaptaron adecuadamente a esta metodología, lo que permitió que todas las actividades se ejecuten de forma correcta y se cubran todos los aspectos necesarios. La primera fase de la metodología de análisis de sentimientos KDT fue completada en este objetivo.

Un aspecto limitante en la extracción de tweets es que, si bien se puede obtener información como el nombre de usuario, la ubicación y el contenido de los tweets, así como otros detalles públicos de los perfiles, hay datos que podrían ser muy útiles para lograr un análisis más

completo de las publicaciones, por ejemplo, la edad de los usuarios. Sin embargo, estos datos son considerados por Twitter como información privada y no está disponible para obtenerla mediante el scraping directo.

Otro aspecto limitante radica en la abundancia de información irrelevante, por ejemplo, existen muchas publicaciones que son usadas con fines promocionales y que contienen enlaces, además es muy frecuente el uso de retweets que por lo general no reflejan la opinión del autor que la comparte. Por lo tanto, es fundamental considerar estas particularidades ya sea en el preprocesamiento o al momento de configurar la herramienta de twitter scraping como es el caso del presente TT, con el fin de obtener las publicaciones más adecuadas desde un inicio.

## **7.2. Objetivo 2: Aplicar el análisis de sentimientos mediante una técnica basada en Machine Learning.**

En este objetivo se ejecutaron las fases 2, 3 y 4 de la metodología KDT, que corresponden al preprocesamiento de datos, transformación de datos y detección de sentimiento. Durante el preprocesamiento, la limpieza manual fue fundamental para garantizar que los tweets que se utilizaron como base para el entrenamiento, sobre todo los relacionados a depresión, sean los más óptimos posibles; esto se logró seleccionando solo los tweets que contenían depresión emocional y se descartaron los restantes, todo esto con la ayuda de un experto en el área como la Dra. Sandra Otero<sup>18</sup>. A través de esto se conoció que muchos usuarios en sus publicaciones depresivas suelen mencionar ciertas características similares, por ejemplo, mencionan síntomas físicos y emocionales como cansancio, nervios, insomnio, pérdida de apetito o que están desmotivados, sin mencionar necesariamente términos de depresión en sus publicaciones. Por otro lado, algunos autores sí hacen mención directa a la depresión y al uso de antidepresivos.

Como resultado de la limpieza manual, la cantidad de tweets recopilados inicialmente disminuyó significativamente, pasando de los 31978 tweets relacionados con depresión obtenidos al inicio a 5328. Además, durante este proceso se descubrió que muchos tweets están repetidos, debido principalmente al uso de mapa de mosaicos H3 para obtener toda la zona geográfica de Ecuador (ver sección 6.1.3), donde el radio de búsqueda se ajusta a cada celda, por lo que existen áreas en donde muchas publicaciones son recopiladas más de una vez. Por lo tanto, la cantidad de tweets se redujo aún más, quedando finalmente 1654 tweets que fueron utilizados. Con esto se demostró que la técnica de recopilación no es la más

---

<sup>18</sup> Evidencia del proceso de validación con la Dra. Sandra Otero:  
[https://drive.google.com/drive/folders/1WKKCqOgM\\_PI1aYCbVBTIHoH6-4Zu8vhW?usp=sharing](https://drive.google.com/drive/folders/1WKKCqOgM_PI1aYCbVBTIHoH6-4Zu8vhW?usp=sharing)

efectiva para obtener una mayor cantidad de publicaciones válidas en comparación con la que se obtuvo inicialmente. Sin embargo, tomando en cuenta que se obtuvieron publicaciones de usuarios geográficamente dispersos en todo el territorio de Ecuador, y que se usaron palabras clave para recopilar tweets en una red social que se caracteriza por la amplia variedad de publicaciones, se logró obtener un resultado aceptable para ser considerado como fuente de entrenamiento, y que es representativo para los 1.45 millones de usuarios que están activos en Twitter en Ecuador hasta la fecha de recopilación.

En el preprocesamiento se utilizó inicialmente la librería Spacy para llevar a cabo la lematización de texto en español, sin embargo, se identificaron errores al encontrar el lema de varias palabras. Por esta razón, se decidió explorar una alternativa que fue identificada en la revisión de literatura que es Stanza [74] [90], que además ofrece algunas ventajas respecto a la primera opción (ver sección 6.1.1.4.4), la tabla 44 muestra los resultados que se obtuvieron en algunas palabras al ejecutar ambas herramientas.

**Tabla 44.** Obtención del lema de algunas palabras usando Spacy y Stanza.

Palabra	Lema	
	Spacy	Stanza
cerre	cerrir	cerre
tragarte	tragar tu	tragarte
jajaja	jajajar	jajaja
atras	atra	atros
podriendo	podreir	podriendo
chiste	chistir	chiste
tomarlo	tomar el	tomarlo
estres	estr	estre

Como se observa en la tabla 44, ambas herramientas presentan conflictos en determinadas palabras. Sin embargo, los resultados de Stanza demuestran que tiene una cantidad menor de errores en la identificación del lema de las palabras en comparación con Spacy. Debido a esto, se decidió utilizar Stanza para la lematización del texto, ya que, aunque también presenta algunos errores, preserva el significado en una gran cantidad de palabras.

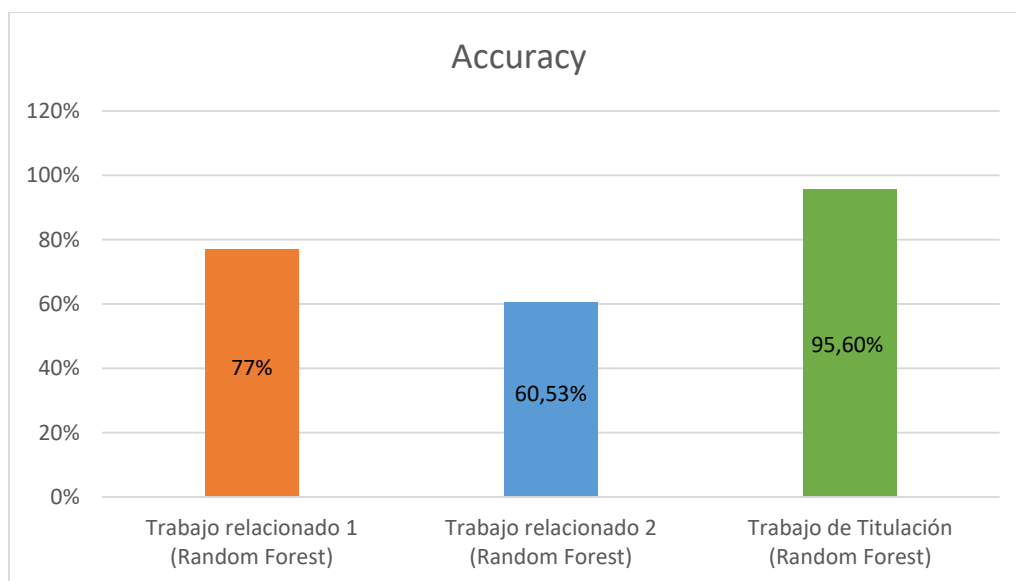
Durante la fase de transformación de datos, se aplicó TF-IDF para determinar la relevancia de las palabras en el conjunto de datos a través de una medida numérica, conjuntamente con unigramas (palabras individuales), bigramas (parejas de palabras consecutivas) y trigramas (tres palabras consecutivas), con el fin de comprobar cuál de estas opciones nos brinda un mejor rendimiento. Los modelos de Support Vector Machines, Random Forest y Naive Bayes se entrenaron con la exactitud (accuracy) como principal criterio de aceptación, con la finalidad de optar por el mejor entre todos ellos. Entre los 3 modelos con unigramas se obtuvo una exactitud que va desde el 90%, desde el 82% con bigramas y desde el 62% con trigramas,

por lo tanto, se comprobó que en los 3 modelos la exactitud disminuía cuando se utilizaban con bigramas y trigramas respectivamente; esto confirma la naturaleza informal de los tweets, ya que no siguen una estructura de redacción uniforme en las distintas publicaciones.

Además, la aplicación del equilibrio de datos usando la técnica SMOTE durante la transformación de los datos fue crucial para evitar el sesgo de los algoritmos de clasificación hacia la clase mayoritaria. Gracias a esta técnica, se logró aumentar significativamente el número de datos en la clase depresiva, pasando de 1654 a 9060 registros, lo que proporcionó una mayor cantidad de información y mejoró la representatividad de esta clase en el conjunto de entrenamiento. Asimismo, la combinación de esta estrategia con la validación cruzada permitió confirmar la utilidad y efectividad de SMOTE, asegurando la variabilidad del conjunto de datos y la fiabilidad de los modelos entrenados.

### 7.3. Objetivo 3: Interpretar los resultados obtenidos en el análisis de sentimientos.

Al realizar este objetivo se ejecutó la fase 5 de la metodología KDT, por lo que se comprobó de forma gráfica el rendimiento de los mejores modelos entrenados en el objetivo 2, obteniendo como mejor modelo a Random Forest con una exactitud (accuracy) del 95.6%, este resultado se puede contrastar con los trabajos relacionados [35], en donde se identificó a Random Forest como el mejor modelo de predicción para el análisis de sentimientos en las publicaciones de Twitter, y en [58] obtuvieron un menor rendimiento usando el mismo modelo, en la figura 52 se presentan los porcentajes de rendimiento de los trabajos relacionados y del presente TT.



**Figura 52.** Comparación del rendimiento del modelo Random Forest con otros modelos en trabajos relacionados.

En la figura 52 se observa que los resultados de los trabajos relacionados que usaron el mismo modelo del presente TT para el análisis de sentimientos obtuvieron una exactitud de 77% y 60.53% respectivamente, estos resultados fueron mejorados en el presente TT con una exactitud de 95,6%. Sin embargo, cabe mencionar que los trabajos relacionados se enfocaron en publicaciones con texto en inglés, mientras en el presente proyecto se utilizó texto en español. Además, es necesario tener en cuenta que la cantidad de datos utilizados para el entrenamiento varía en cada uno de los trabajos.

Con el mejor modelo para clasificar contenido depresivo se predijeron tweets de tiempos prepandemia, específicamente del año 2019, en donde se pudo dar contestación a la pregunta de investigación: ¿En qué medida existen tweets con contenido depresivo en Twitter en tiempos de covid-19 en Ecuador?. Se comparó los datos del 2019 con los extraídos del año 2020 y 2021 (tiempos de pandemia), en donde se pudo comprobar y medir de forma porcentual que respecto al año anterior a la pandemia (2019), la cantidad de publicaciones depresivas incrementó en un 31,17% en el año 2020 y se mantuvo casi en la misma cantidad en el año 2021.

Un factor limitante de este objetivo es que por la cantidad limitada de datos que se tuvo para el entrenamiento, no se aprovecharon los modelos de aprendizaje automático y no se logró alcanzar su máximo potencial para que se pueda reducir la probabilidad de proporcionar predicciones erróneas, especialmente al analizar nuevos datos que contienen características informales mencionadas en la sección 6.2.1. Además, es importante tener en cuenta que la precisión del análisis de sentimientos puede verse afectada por el contexto y el uso de sarcasmo e ironía en los tweets [91], lo que podría llevar a interpretaciones erróneas.

Es importante destacar que el estigma que rodea a la salud mental y la falta de acceso a la atención pueden tener consecuencias devastadoras para aquellos que sufren de depresión y otros trastornos relacionados con la salud mental. Además, la pandemia de COVID-19 ha destacado la necesidad de soluciones efectivas de detección y tratamiento de la depresión, así como la necesidad de apoyo a la salud mental en general. Por lo tanto, con el análisis de sentimientos para detectar la depresión en Twitter se pretende brindar una herramienta útil que contribuya para abordar este problema, pero debe ser parte de un enfoque más amplio para abordar la salud mental y el bienestar en general. Asimismo, es importante tener en cuenta que este análisis de sentimientos no debe considerarse una herramienta de diagnóstico de depresión, sino una herramienta complementaria, ya que los datos utilizados para el entrenamiento de acuerdo a la Dra. Sandra Otero, se pueden tomar como una alerta, pero es necesario profundizar en los casos y realizar una evaluación clínica completa para confirmar un diagnóstico de depresión.

Finalmente, es fundamental mencionar que todos los datasets, modelos y código generado durante todo el proceso del presente trabajo está disponible en su totalidad en el repositorio de Github<sup>19</sup>. Además, se proporcionan videos<sup>20</sup> explicativos detallando como se llevó a cabo todo este proceso, para garantizar tanto la reproducibilidad como la transparencia de esta investigación.

## 7.4. Valoración técnica, económica, ambiental y social

### 7.4.1. Valoración técnica

El presente TT se valora técnicamente a razón de las múltiples herramientas de hardware y software que fueron utilizadas en la construcción del modelo para identificar depresión. Algunas de las librerías más relevantes usadas son: Twint para la extracción y construcción del dataset con tweets, Scikit Learn para la utilización de los algoritmos de Máquinas de Soporte Vectorial, Random Forest y Naive Bayes, NLTK para realizar un adecuado preprocesamiento de los datos sobre todo en la tokenización y eliminación de stopwords, Stanza para el preprocesamiento en la lematización de los datos, NumPy para realizar cálculos numéricos y almacenamiento de los vectores de características generados y la librería Imblearn para realizar el sobremuestreo de las clases desequilibradas. Todas estas librerías desempeñaron un papel esencial para lograr el cumplimiento de todas las fases del análisis de sentimientos.

### 7.4.2. Valoración económica

En el desarrollo del presente TT, fue necesaria la inversión de talento humano, recursos de hardware y software, los cuales se presentan en la tabla 45, tabla 46 y tabla 47.

**Tabla 45.** Presupuesto - recursos humanos

Talento Humano	Justificación	N°. horas	Valor por Hora	Subtotal
Autor del proyecto	Estudiante a cargo de la ejecución del proyecto	400	\$5	\$2000
Personal de apoyo	Docente director en la elaboración y supervisión del proyecto	80	\$12.50	\$1000

<sup>19</sup> Código en repositorio del análisis de sentimientos:  
[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador](https://github.com/byronmb/Identificacion_Depresion_Ecuador)

<sup>20</sup> Lista de videos explicativos del proceso de análisis de sentimientos:  
<https://youtube.com/playlist?list=PLC9uZGfnbZtj3i-IIZmPHm4MXuWYxvaWf>

	Docente guía de la materia de Trabajo de Titulación de la Carrera de Ingeniería en Sistemas	40	\$12.50	\$500
--	---	----	---------	-------

**Tabla 46.** Presupuesto - recursos HW, SW y TICs

Recursos HW/SW	Justificación	N meses	Valor por mes	Subtotal
Internet	Revisión de literatura y Comunicación.	5	\$40	\$200
Computadora	Herramienta para el trabajo a desarrollar	5	\$10	\$50
Mendeley	Gestión de referencias bibliográficas	5	\$0	\$0
Python	Creación de código y modelos de Machine Learning	5	\$0	\$0
Jupyter	Creación de código y modelos de Machine Learning	5	\$0	\$0
<b>Total</b>				<b>\$250</b>

**Tabla 47.** Presupuesto total del PTT

Recursos	Subtotal
Recursos Humanos	\$3500
Recursos Software. Hardware y TICs	\$250
Insumos	\$15
Subtotal	\$3765
Imprevistos (+10% del Subtotal) \$376,50	\$376.50
Presupuesto total del Proyecto	\$4141.50

#### 7.4.3. Valoración ambiental

El presente TT se realizó en su totalidad con recursos tecnológicos y digitales que no tienen un mayor impacto al medio ambiente, además que se tuvo un bajo consumo de recursos materiales o de otros elementos que puedan llegar a perjudicar al medio ambiente.

#### 7.4.4. Valoración social

El presente TT ofrece un aporte social ya que se pudo clasificar el sentimiento depresivo en publicaciones de Twitter, con lo cual la información y resultados generados podrían ser usados en el campo de la salud como una herramienta complementaria y de concientización que sirva como un aporte en la mitigación de problemas de salud mental en el Ecuador,



además en la educación puede ser una valiosa aportación como una fuente de información para las instituciones educativas, especialmente en colegios y universidades, ya que según menciona la Dra. Ximena Amaya Valarezo en una entrevista (ver anexo 1), en Ecuador nos hace falta culturizarnos acerca de la importancia de la salud mental. Por lo tanto, esta información puede ayudar a las instituciones a desarrollar nuevas propuestas y metodologías para realizar una intervención preventiva en la salud mental de sus estudiantes.

## 8. Conclusiones

- A través de la revisión de literatura se determinó cuáles son las herramientas de lenguaje natural más adecuadas para trabajar con texto en español, ya que muchas existentes están limitadas en cuanto al idioma, lo que permitió conocer y comprobar que las herramientas NLTK y Stanza son las más adecuadas brindando una mejor precisión, además de ser multilingües y permitir trabajar con texto en idioma español.
- Ejecutar el entrenamiento de todos los algoritmos con N-gramas permitió conocer el comportamiento de cada uno de estos, comprobándose que los modelos de Máquinas de Soporte Vectorial, Random Forest y Naive Bayes, tienen un mejor rendimiento cuando están entrenados con unigramas, disminuyendo este rendimiento si se entrenan con bigramas y trigramas respectivamente, por lo que se puede confirmar que el texto en los tweets no tienen una estructura estándar en la redacción, evidenciando de esta forma la característica informal en la mayoría de las publicaciones de Twitter en Ecuador.
- Comparando los resultados del rendimiento de cada algoritmo de clasificación, se concluye que el algoritmo con mejor resultado en las predicciones es Random Forest con una exactitud de 95.4%, sin dejar de lado al algoritmo Máquinas de Soporte Vectorial que ofrece un rendimiento relativamente cercano con un 95% de exactitud, por lo que también puede ser considerado para clasificar sentimiento depresivo.
- El análisis de sentimientos durante los dos primeros años de la pandemia permitió obtener una visión integral y contextualizada del impacto emocional en términos de depresión. Se comprobó un significativo crecimiento porcentual del 31.17% en la cantidad de tweets depresivos durante el año 2020 en comparación con el año anterior. Este resultado sugiere que desde el inicio de la pandemia de COVID-19, tuvo un impacto negativo en la salud mental de los usuarios activos de Twitter en Ecuador.

## 9. Recomendaciones

- Se recomienda que antes de entrenar los modelos y realizar la detección de sentimiento, se aborde el problema del desequilibrio de clases a través de la generación de muestras sintéticas de las clases minoritarias. Esto permitirá establecer un equilibrio entre ellas, aumentar la cantidad de datos disponibles, y evitar que los modelos presenten un sesgo hacia la clase mayoritaria, especialmente en situaciones en las que la clase minoritaria cuenta con pocos datos.
- En caso de hacer predicciones con nuevos datos (tweets), se recomienda llevar a cabo un proceso de limpieza manual que abarque la eliminación eficiente de la información ruidosa, errores gramaticales, abreviaturas o mezcla de diferentes lenguajes que a menudo caracterizan a las publicaciones de twitter y que pueden afectar negativamente la clasificación adecuada del sentimiento.
- Para futuros trabajos, se sugiere explorar diferentes fuentes de datos para detectar riesgos de depresión, por ejemplo, datos biométricos, expresiones faciales y patrones de habla de los usuarios. Además, se podría ampliar el alcance del análisis a otras redes sociales como Facebook e Instagram, lo cual proporcionaría una visión más completa y diversificada de las tendencias de sentimientos en línea.
- También como trabajo futuro, se recomienda expandir el alcance del análisis de sentimientos para incluir otros trastornos mentales, como el estrés y el trastorno bipolar. De esta manera, se podría ofrecer una herramienta más completa y versátil que brinde una comprensión más profunda y detallada de la salud mental de los usuarios de las redes sociales.

## 10. Bibliografía

- [1] World Health Organization, “Depression,” 2021. <https://www.who.int/news-room/fact-sheets/detail/depression> (accessed Dec. 31, 2021).
- [2] S. A. Alharthi, “Empirical Study of Features and Unsupervised Sentiment Analysis Techniques for Depression Detection in Social Media,” *Adv. Comput. Sci. Adv. Comput. Sci.*, pp. 3–4, 2020, [Online]. Available: <https://www.boffinaccess.com/journals/advances-in-computer-sciences/acs>.
- [3] S. Zahoor and R. Rohilla, “Twitter Sentiment Analysis Using Machine Learning Algorithms: A Case Study,” in *2020 International Conference on Advances in Computing, Communication Materials (ICACCM)*, 2020, pp. 194–199, doi: 10.1109/ICACCM50413.2020.9213011.
- [4] OPS, “Depresión y otros trastornos mentales comunes,” *Organ. Panam. la Salud*, pp. 1–24, 2017, [Online]. Available: <http://iris.paho.org/xmlui/bitstream/handle/123456789/34006/PAHONMH17005-spa.pdf>.
- [5] A. Tusev, L. Tonon, and M. Capella, “Efectos Iniciales en la Salud Mental por la Pandemia de Covid-19 en algunas Provincias de Ecuador,” *Investigatio*, vol. 15, no. 15, pp. 11–22, 2020, doi: 10.31095/investigatio.2020.15.2.
- [6] S. Li, Y. Wang, J. Xue, N. Zhao, and T. Zhu, “The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users,” *Int. J. Environ. Res. Public Health*, vol. 17, no. 6, 2020, doi: 10.3390/ijerph17062032.
- [7] OMS, “Plan de acción sobre salud mental 2013-2020,” *Organización Mundial de la Salud*. p. 54, 2013.
- [8] World Health Organization, “COVID 19 Public Health Emergency of International Concern (PHEIC) Global Research and Innovation Forum: Towards a Research Roadmap,” *Glob. Res. Collab. Infect. Dis. Prep.*, pp. 1–10, 2020, [Online]. Available: [https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-\(pheic\)-global-research-and-innovation-forum](https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum).
- [9] Confederación Salud Mental España, “Salud mental y COVID-19. Un año de pandemia,” *Confed. Salud Ment. España*, pp. 1–17, 2021.
- [10] K. Chaudhary, M. Alam, M. S. Al-Rakhami, and A. Gumaei, “Machine learning-based mathematical modelling for prediction of social media consumer behavior using big data analytics,” *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00466-2.
- [11] A. Sapountzi and K. E. Psannis, “Social networking data analysis tools & challenges,” *Futur. Gener. Comput. Syst.*, vol. 86, pp. 893–913, 2018, doi: 10.1016/j.future.2016.10.019.
- [12] A. Deshwal and S. K. Sharma, “Twitter Sentiment Analysis using various Classification Algorithms,” in *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2016, pp. 251–257, doi: 10.1109/ICRITO.2016.7784960.
- [13] R. D. Desai, “Sentiment Analysis of Twitter Data,” in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 114–117, doi: 10.1109/ICCONS.2018.8662942.
- [14] J. P. Alcázar, “Ecuador Estado Digital Ene/21,” in *Mentinno – Innovation & Lifetime Value Partners*, 2021, p. 101, [Online]. Available:

<https://www.mentinno.com/informesdigitalesecuador>.

- [15] B. Liu, *Sentiment Analysis and Opinion Mining*, vol. 5, no. 1. Morgan & Claypool, 2012.
- [16] C. Xavier and M. Souza, "A Basic Approach for Extracting and Analyzing Data from Twitter," in *Special Topics in Multimedia, IoT and Web Technologies*, 2020, pp. 185–211.
- [17] P. Boucher, *Artificial intelligence: How does it work, why does it matter, and what can we do about it?*, no. June. 2020.
- [18] T. Jo, *Text Mining: Concepts, Implementation, and Big Data Challenge*, vol. 26, no. 2. Seoul, Korea (Republic of), 2019.
- [19] Z. Rybchak and O. Basystiuk, "Analysis of methods and means of text mining," *ECONTECHMOD. An Int. Q. J.*, vol. 6, no. 2, pp. 73–78, 2017.
- [20] S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "A survey of text mining in social media: Facebook and Twitter perspectives," *Adv. Sci. Technol. Eng. Syst. J.*, vol. 2, no. 1, pp. 127–133, 2017, doi: 10.25046/aj020115.
- [21] R. Talib, M. Kashif, S. Ayesha, and F. Fatima, "Text Mining: Techniques, Applications and Issues," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, 2016, doi: 10.14569/IJACSA.2016.071153.
- [22] M. Chistol and M. Danubianu, "Survey of Text Mining Research Methods and Their Innovative Applicability," *J. Danubian Stud. Res.*, vol. 11, no. 1, pp. 225–233, 2021.
- [23] M. Delgado, M. J. Martín-Bautista, D. Sánchez, and M. A. Vila, "Mining text data: Special features and patterns," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2447, pp. 140–153, 2002, doi: 10.1007/3-540-45728-3\_11.
- [24] S. A. Salloum, M. Al-Emran, and K. Shaalan, "Mining Text in News Channels: A Case Study from Facebook," *Int. J. Inf. Technol. Lang. Stud.*, vol. 1, no. 1, pp. 1–9, 2017.
- [25] V. Tyag and S. Singh, "Sentiment Analysis to Detect Mental Depression Based on Twitter Data," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 8, no. 9, pp. 268–274, 2020, doi: 10.22214/ijraset.2020.31392.
- [26] S. Yang, Z. Ning, and Y. Wu, "NLP Based on Twitter Information: A Survey Report," *Proc. - 2020 2nd Int. Conf. Inf. Technol. Comput. Appl. ITCA 2020*, pp. 620–625, 2020, doi: 10.1109/ITCA52113.2020.00135.
- [27] V. A. Kharde and S. Sonawane, "Sentiment Analysis of Twitter Data : A Survey of Techniques," *Int. J. Comput. Appl. (0975)*, vol. abs/1601.0, 2016, [Online]. Available: <http://arxiv.org/abs/1601.06971>.
- [28] C. Diamantini, A. Mircoli, D. Potena, and E. Storti, "Social information discovery enhanced by sentiment analysis techniques," *Futur. Gener. Comput. Syst.*, vol. 95, pp. 816–828, 2018, doi: 10.1016/j.future.2018.01.051.
- [29] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment Analysis and Affective Computing for depression monitoring," *Proc. - 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017*, vol. 2017-Janua, pp. 1988–1995, 2017, doi: 10.1109/BIBM.2017.8217966.
- [30] A. Malik, Y. T. Javeri, M. Shah, and R. Mangrulkar, "Impact analysis of COVID-19 news headlines on global economy," in *Cyber-Physical Systems*, R. C. Poonia, B. Agarwal, S. Kumar, M. S. Khan, G. Marques, and J. Nayak, Eds. Academic Press,

- 2022, pp. 189–206.
- [31] A. Giachanou and F. Crestani, “Like It or Not: A survey of Twitter Sentiment Analysis Methods,” *ACM Comput. Surv.*, vol. 49, pp. 28:1-28:40, 2016, doi: 10.1145/2938640.
- [32] J. F. Raisa, M. Ulfat, A. Al Mueed, and S. M. S. Reza, “A Review on Twitter Sentiment Analysis Approaches,” in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2021, pp. 375–379, doi: 10.1109/ICICT4SD50815.2021.9396915.
- [33] I. El Naqa and M. J. Murphy, “What Is Machine Learning?,” *Mach. Learn. Radiat. Oncol.*, pp. 3–11, 2015, doi: 10.1007/978-3-319-18305-3\_1.
- [34] S. K. Khanna, “Machine Learning v/s Deep Learning,” *Int. Res. J. Eng. Technol.*, vol. 6, no. 2, pp. 455–458, 2019.
- [35] F. Azam, M. Agro, M. Sami, M. H. Abro, and A. Dewani, “Identifying Depression among Twitter Users using Sentiment Analysis,” *2021 Int. Conf. Artif. Intell. ICAI 2021*, pp. 44–49, 2021, doi: 10.1109/ICAI52203.2021.9445271.
- [36] A. P. Jain and P. Dandannavar, “Application of Machine Learning techniques to Sentiment Analysis,” *Proc. 2016 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. iCATccT 2016*, pp. 628–632, 2017, doi: 10.1109/ICATCCT.2016.7912076.
- [37] D. A. Musleh *et al.*, “Twitter Arabic Sentiment Analysis to detect Depression using Machine Learning,” *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 3463–3477, 2022, doi: 10.32604/cmc.2022.022508.
- [38] M. T. Khan, M. Durrani, A. Ali, I. Inayat, S. Khalid, and K. H. Khan, “Sentiment analysis and the complex natural language,” *Complex Adapt. Syst. Model.*, vol. 4, no. 1, 2016, doi: 10.1186/s40294-016-0016-9.
- [39] R. Srinivasan and C. N. Subalalitha, “Sentimental analysis from imbalanced code-mixed data using machine learning approaches,” *Distrib. Parallel Databases*, pp. 1–16, 2021, doi: 10.1007/s10619-021-07331-4.
- [40] K. Borowska and J. Stepaniuk, “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data Gustavo,” *ACM SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004, doi: 10.1145/1007730.1007735.
- [41] A. Kulkarni, D. Chong, and F. A. Batarseh, “Foundations of data imbalance and solutions for a data democracy,” *CoRR*, pp. 83–106, 2021, doi: 10.48550/arXiv.2108.00071.
- [42] A. Vijayvargiya, C. Prakash, R. Kumar, S. Bansal, and J. M. Tavares, “Human knee abnormality detection from Imbalanced sEMG data,” *Biomed. Signal Process. Control*, vol. 66, no. April, pp. 0–35, 2021, doi: 10.1016/j.bspc.2021.102406.
- [43] D. M. Aguilera, “Sistema de Detección y Clasificación de Melanomas a través de Imágenes,” Universidad Politécnica de Madrid, 2021.
- [44] R. Arghandeh *et al.*, “Data Mining Techniques and Tools for Synchrophasor Data,” *North American SynchroPhasor Initiative (NASPI)*. p. 45, 2019, doi: 10.13140/RG.2.2.22389.22242.
- [45] I. Stancin and A. Jovic, “An overview and comparison of free Python libraries for data mining and big data analysis,” *2019 42nd Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2019 - Proc.*, pp. 977–982, 2019, doi: 10.23919/MIPRO.2019.8757088.
- [46] K. M. Mendez, L. Pritchard, S. N. Reinke, and D. I. Broadhurst, “Toward collaborative

- open data science in metabolomics using Jupyter Notebooks and cloud computing,” *Metabolomics*, vol. 15, no. 10, pp. 1–16, 2019, doi: 10.1007/s11306-019-1588-0.
- [47] O. Widhyartha, “Engaging People’s Enthusiasm in 2020 Population Census by Scrapping Social Media,” *Expert Meet. Dissem. Commun. Stat.*, no. 1, 2021.
- [48] M. Aljabri *et al.*, “Sentiment analysis of Arabic tweets regarding distance learning in Saudi Arabia during the COVID-19 pandemic,” *Sensors*, vol. 21, no. 16, 2021, doi: 10.3390/s21165431.
- [49] J. L. Martínez, “Identificación de depresión mediante el análisis de sentimientos,” Universidad de Extremadura, 2019.
- [50] Z. Chen and M. Sokolova, “Sentiment Analysis of the COVID-related r/Depression Posts.” arXiv, 2021, doi: 10.48550/ARXIV.2108.06215.
- [51] C. Bhargava, S. Poornima, S. Mahur, and M. Pushpalatha, “Depression Detection Using Sentiment Analysis of Tweets,” *Turkish J. Comput. Math. Educ. Res. Artic.*, vol. 12, no. 11, pp. 5411–5418, 2021.
- [52] A. Sood, M. Hooda, S. Dhira, and M. Bhatia, “An Initiative To Identify Depression Using Sentiment Analysis: a Machine Learning Approach,” *Indian J. Sci. Technol.*, vol. 11, no. 4, pp. 1–20, 2018, doi: 10.17485/ijst/2018/v11i4/119594.
- [53] Y. Castán, “Introducción al Metodo Cientifico y Sus Etapas,” *Inst. Aragon. Ciencias La Salud*, vol. 2, pp. 1–6, 2006.
- [54] J. L. Abreu, “El Método de la Investigación,” *Daena Int. J. Good Conscienc.*, vol. 9, no. 3, pp. 195–204, 2014.
- [55] C. A. Espinoza, *Metodología de investigación tecnológica*. 2014.
- [56] M. Genero, J. A. Cruz-Lemus, and M. G. Piattini, *Métodos de investigación en ingeniería del software*. 2014.
- [57] T. Baviera Puig, “Técnicas para el Análisis de Sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength,” *Dígitos Rev. Comun. Digit.*, vol. 1, no. 3, pp. 33–50, 2017.
- [58] A. Saha, A. Al Marouf, and R. Hossain, “Sentiment Analysis from Depression-Related User-Generated Contents from Social Media,” in *2021 8th International Conference on Computer and Communication Engineering (ICCCCE)*, 2021, pp. 259–264, doi: 10.1109/ICCCCE50029.2021.9467214.
- [59] M. I. Sajib, S. Mahmud Shargo, and M. A. Hossain, “Comparison of the efficiency of Machine Learning algorithms on Twitter Sentiment Analysis of Pathao,” in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 2019, pp. 1–6, doi: 10.1109/ICCIT48885.2019.9038208.
- [60] S. Tiwari, A. Verma, P. Garg, and D. Bansal, “Social Media Sentiment Analysis On Twitter Datasets,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 925–927, doi: 10.1109/ICACCS48705.2020.9074208.
- [61] M. Deshpande and V. Rao, “Depression detection using emotion artificial intelligence,” *Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2017*, no. Iciss, pp. 858–862, 2018, doi: 10.1109/ISS1.2017.8389299.
- [62] J. A. Mansilla, “Minado de texto aplicado en Twitter para obtener la polaridad en opiniones de usuarios acerca den nuevo Proyecto de ley que regula la migración en Chile,” Universidad del Bío-Bío, 2018.

- [63] L. Liu, *Encyclopedia of Database Systems*, 1st ed. Springer, 2009.
- [64] W. Elmenreich, J. T. Machado, and I. J. Rudas, *Intelligent Systems at the Service of the Mankind*, vol. 1. 2003.
- [65] S. M. Fonseca *et al.*, "An approach based on text mining for knowledge acquisition in diagnostic systems," 2007.
- [66] S. J. Pachouly, G. Raut, K. Bute, R. Tambe, and S. Bhavsar, "Depression Detection on Social Media Network (Twitter) using Sentiment Analysis," *Int. Res. J. Eng. Technol.*, vol. 08, no. 01, 2021, [Online]. Available: [www.irjet.net](http://www.irjet.net).
- [67] B. A. Eclarin, A. C. Fajardo, and R. P. Medina, "A novel feature hashing with efficient collision resolution for bag-of-words representation of text data," *ACM Int. Conf. Proceeding Ser.*, pp. 12–16, 2018, doi: 10.1145/3278293.3278301.
- [68] A. Tripathy, A. Agrawal, and S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Syst. Appl.*, vol. 57, pp. 117–126, 2016, doi: 10.1016/j.eswa.2016.03.028.
- [69] N. V. Babu and E. G. M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review," *SN Comput. Sci.*, vol. 3, no. 74, pp. 1–20, 2021, doi: 10.1007/s42979-021-00958-1.
- [70] A. P. Tirtopangarsa and W. Maharani, "Sentiment Analysis of Depression Detection on Twitter Social Media Users Using the K-Nearest Neighbor Method," *Inov. Teknol. dan Pengolah. Inf. untuk Mendukung Transform. Digit.*, vol. 1, no. 1, pp. 247–258, 2021.
- [71] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2017.
- [72] B. Bondaruk, S. A. Roberts, and C. Robertson, "Assessing the state of the art in Discrete Global Grid Systems: OGC criteria and present functionality," *Geomatica*, vol. 74, no. 1, pp. 9–30, 2020, doi: 10.1139/geomat-2019-0015.
- [73] I. Lauriola, A. Lavelli, and F. Aielli, "An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, pp. 443–456, 2021, doi: 10.1016/j.neucom.2021.05.103.
- [74] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages," *CoRR*, pp. 101–108, 2020, doi: 10.18653/v1/2020.acl-demos.14.
- [75] L. Sohmen and L. Rossenova, "Open refine to wikibase: a new data upload pipeline," in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, 2022, no. 53, pp. 1–2, doi: 10.1145/3529372.3530919.
- [76] O. Kramer, "Scikit-Learn," in *Machine Learning for Evolution Strategies*, Springer International Publishing, 2016, pp. 45–53.
- [77] G. Zhao, Y. Liu, W. Zhang, and Y. Wang, "TFIDF based feature words extraction and topic modeling for short text," *ACM Int. Conf. Proceeding Ser.*, pp. 188–191, 2018, doi: 10.1145/3180374.3181354.
- [78] D. Sarkar, "Feature Engineering for Text Representation," in *Text Analytics with Python: A Practitioner's Guide to Natural Language Processing*, Apress, 2019, pp. 201–273.
- [79] J. Brownlee, *XGBoost With Python: Gradient Boosted Trees with XGBoost an scikit-learn*, vol. v1.10. 2018.



- [80] P. Singh and A. Manure, *Learn TensorFlow 2.0: Implement Machine Learning and Deep Learning Models with Python*. Bangalore: Apress, 2020.
- [81] U. Saeed, S. Ullah Jan, Y. D. Lee, and I. Koo, "Machine Learning-based Real-Time Sensor Drift Fault Detection using Raspberry Pi," *2020 Int. Conf. Electron. Information, Commun. ICEIC 2020*, vol. 2020-Janua, pp. 4–10, 2020, doi: 10.1109/ICEIC49074.2020.9102342.
- [82] A. Leis, F. Ronzano, M. A. Mayer, L. I. Furlong, and F. Sanz, "Detecting signs of depression in tweets in Spanish: Behavioral and linguistic analysis," *J. Med. Internet Res.*, vol. 21, no. 6, 2019, doi: 10.2196/14199.
- [83] M. Park, C. Cha, and M. Cha, "Depressive moods of users portrayed in Twitter," *Proc. 18th ACM Int. Conf. Knowl. Discov. Data Mining, SIGKDD 2012*, pp. 1–8, 2012.
- [84] M. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting Depression via Social Media," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2013, vol. 7, no. 1, pp. 128–137, doi: 10.3109/01460862.2013.798190.
- [85] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?," in *Machine Learning and Data Mining in Pattern Recognition*, no. July 2012, Springer Berlin Heidelberg, 2012, pp. 154–168.
- [86] Servicio Nacional de Gestión de Riesgo y Emergencias, "Informe de situación COVID-19 Ecuador 16 de Marzo de 2020," 2020. [Online]. Available: <https://www.gestionderiesgos.gob.ec/wp-content/uploads/2020/03/Informe-de-Situación-No008-Casos-Coronavirus-Ecuador-16032020-20h00.pdf>.
- [87] S. Singh *et al.*, "A Novel Approach for Data Scraping and Sentiment Analysis from Twitter using Machine Learning Techniques," *Am. J. Electron. Commun.*, vol. 2, no. 3, pp. 1–6, 2022, doi: 10.15864/ajec.2301.
- [88] M. S. Muneshwara, M. S. Swetha, M. P. Rohidekar, and A. B. Pranove, "Implementation of Therapy Bot for Potential Users With Depression During Covid-19 Using Sentiment Analysis," *J. Posit. Sch. Psychol.*, vol. 6, no. 3, pp. 7816–7826, 2022.
- [89] R. R. Rivera-Guamán, O. M. Cumbicus-Pineda, R. A. López-Lapo, and L. A. Neyra-Romero, "Sentiment Analysis Related of International Festival of Living Arts Loja-Ecuador Employing Knowledge Discovery in Text," in *Applied Technologies*, 2021, pp. 327–339.
- [90] M. J. Haris, A. Upreti, M. Kurtaran, F. Ginter, S. Lafond, and A. Sepinoud, "Identifying gender bias in blockbuster movies through the lens of machine learning," *Humanit. Soc. Sci. Commun.*, pp. 1–8, 2022, doi: 10.1057/s41599-023-01576-3.
- [91] Karthika K, "An efficient approach for sarcasm detection in tweets using polarity flip," *Int. J. Adv. Res. Ideas Innov. Technol.*, vol. 5, no. 2, pp. 1822–1826, 2019.
- [92] M. Ebner, "Microblogs," *SAGE Encycl. Internet*, pp. 640–641, 2018, doi: 10.4135/9781473960367.n183.
- [93] J. Gutiérrez-Rexach, *ENCICLOPEDIA DE LINGÜÍSTICA HISPÁNICA*, 1st ed. Routledge, 2016.

## 11. Anexos

### **Anexo 1.** Entrevista dirigida a Psicóloga Clínica.

Entrevista realizada con el propósito de sustentar el presente TT, tanto a nivel social como académico. Como apoyo de la información captada en la entrevista, se adjunta la grabación de la misma en el siguiente enlace:

<https://drive.google.com/drive/folders/1XNtTVpQkyxXpBouewN1hgeLNijl6cflq?usp=sharing>

A continuación, se redacta fielmente las respuestas vertidas por su autor.

**Cargo:** Psicóloga Clínica

**Nombre:** Ximena Dennisse Amaya Valarezo

**Fecha de entrevista:** 21-01-2022

### **Descripción:**

#### **1. ¿Considera a la depresión como un problema de salud serio?**

Claro que sí. De hecho, dentro de los nueve años que tengo ejerciendo la carrera es uno de los trastornos que más se trabaja dentro de sesión. Lo dice la Organización Mundial de la Salud, que dentro de la población en general, en los adultos, dentro de toda la población hay un **5%** de personas que lo padecen. Entonces es una enfermedad que lleva a tener bastantes alteraciones psicológicas, inclusive lleva a cometer el acto de suicidarse. Por eso es muy importante detectar a tiempo los síntomas para poder trabajar con los pacientes que padezcan episodios depresivos o que conlleven a un trastorno depresivo mayor o crónico que tenga como finalidad la muerte. Creo que es una de las enfermedades con las que más se batalla últimamente y que en los últimos tiempos se ha logrado detectar con esta situación de la pandemia.

#### **2. ¿Cree que las medidas de restricción tomadas como consecuencia del covid ha repercutido negativamente sobre la salud mental de las personas en Ecuador?**

En cierta parte, creo que el Ecuador, o bueno, quizá algunos países también como Colombia, como Cuba, como Puerto Rico, son países que creo que les hace falta bastante culturizarnos con el tema de la prevención y de la convivencia. Y eso nos jugó un papel importante en la pandemia en el 2020, porque hubo muchos hogares en donde les afectó bastante el tema de la convivencia. Los divorcios aumentaron muchísimo y lógicamente que aumentó en el tema de la convivencia y para las personas que padecieron la enfermedad, ahí hubo casos de depresión, de ansiedad y luego estrés postraumático. Personas que también no tuvieron la enfermedad, pero sí tuvieron casos de tuvieron episodios de ansiedad y de ataques de pánico

porque tenían miedo de padecer lo que es el virus. Ahora la situación ha cambiado un poquito porque quizás antes como no habían vacunas para la enfermedad, era como que más latente y los síntomas eran más fuertes y hacía que nosotros nos enfermemos más.

Pero tiene que ver mucho la parte mental, porque sí nosotros los seres humanos no batallamos o no trabajamos con nuestros pensamientos. Es lo que nos lleva a tener estos ataques de pánico y ansiedad. Y los pensamientos recurrentes en el 2020 eran sobre ¿Y si me enfermo? ¿Y si contagio en mi familia? ¿Y si me llego a morir? Eran ya pensamientos catastróficos. Ahora, ya que existen las vacunas, quizás hay personas que se han contagiado pero los síntomas son un poquito más leves y como que quizá podría ser que nos hemos acostumbrado un poquito a la pandemia, pero aún hace falta trabajar en el tema de la prevención y de la cultura. Creo que sí ha afectado en las familias. El año pasado atendí bastantes casos de divorcio y bastantes casos en donde los pacientes somatizaban todo lo que sentían, el estrés, la ansiedad que tenían dentro del día. Lo somatizaban con enfermedades como insomnio, como problemas gastrointestinales, como dolores de cabeza recurrentes que provenían del estrés y de la ansiedad que no fue tratada a tiempo. Por eso es muy importante acudir al psicólogo.

### **3. ¿Existe estigma o renuencia asociado con buscar ayuda profesional para los problemas mentales, incluyendo la depresión?**

Hablando de 100% de población en cuanto a Ecuador, yo diría que el 40%, por no decir el 35%, todavía considera que ir al psicólogo es para los “locos”. No tomamos en cuenta de que, así como nuestra salud física es muy importante también la salud mental. De qué nos sirve trabajar en nuestro cuerpo, en nuestra alimentación, en nuestro autocontrol, si no cuidamos lo que tenemos en nuestra mente. Nosotros, los seres humanos en el día procesamos entre dos mil a tres mil pensamientos al día. Ahora hay que ver si en las personas, esos pensamientos son más negativos o más positivos que lo que nos decimos a nosotros mismos. Entonces es tan importante ir al psicólogo, no solamente para cuando tenemos algún problema, sino también para saber cómo estamos llevando nuestra vida. Entonces, hoy por hoy, si hay personas que consideran de que ir al psicólogo es solamente cuando es algo está mal y todavía hace falta culturizarnos un poquito más acerca de la importancia del bienestar mental.

### **4. ¿Considera que las redes sociales como Twitter pueden ser un medio para que las personas que sufren depresión puedan expresar sus sentimientos y opiniones?**

Como profesional de la salud mental, siempre yo voy a sugerir acudir a un terapeuta, porque no es lo mismo escuchar un consejo que te dicen en internet o que te lo dice la vecina, o que te lo dice tu mamá, tu papá o algún hermano o hermana, a que te lo diga un profesional que con bases y herramientas psicológicas te puede atender y te puede ayudar a buscar una solución correcta ante los problemas que tú tienes. Entonces el tema de las redes sociales es un tema complejo porque hay un pro y un contra. Quizá el pro es que nosotros podemos utilizar las redes sociales cuando tenemos mucha inteligencia emocional, ¿A qué me refiero?, a que podemos tener un equilibrio entre qué es lo bueno y qué es lo malo y tener una madurez como para poder aceptar los comentarios y las opiniones ajenas a lo que nosotros sentimos. Ahora, ¿cuál es el contra?, cuando personas que tienen alguna alteración psicológica como, por ejemplo, están atravesando alguna crisis de ansiedad, o están pasando por un cuadro psicológico o depresivo, o quizá tienen depresión, pero no se han dado cuenta, no tienen esa sabiduría o esa inteligencia emocional para poder proyectar o para poder percibir un comentario, ¿a qué me refiero?, hay muchas personas que tienen depresión, arrastran episodios depresivos en ciclos de su vida, pero como no lo han detectado tiempo, no han sido tratados. Entonces eso conlleva que tomen malas decisiones, a que tengan intolerancia a la frustración, a que se enfermen con mayor facilidad y no puedan gestionar sus emociones.

Entonces, si un depresivo crónico, quizá en su medio de su alteración, lee dentro de las redes sociales la vida no sirve de nada, de que es una persona insensible o lo atacan con comentarios negativos, quizá no tenga el suficiente valor o la suficiente madurez como para poder aceptar ese comentario. Y puede jugar un papel importante como para que tome la decisión de quitarse la vida o tome la decisión que los otros están haciendo, cuando en realidad lo importante es poder tener un equilibrio y poder gestionar nuestras emociones de la forma más adecuada. Creo que las redes sociales nos ayudarían, siempre y cuando nosotros tengamos una madurez emocional equilibrada. Si nosotros no tenemos la madurez emocional que se requiere, como aceptar la opinión de la otra persona sin que me afecte a mí, o leer un comentario brusco hacia mi persona pero que no me llegue a afectar como para atentar contra mi vida o bajar mis niveles de autoestima creo que está bien, pero si me afecta, me duele y me va a conllevar a que yo tenga ataques de ira, a que me desquite con la persona que está al lado, a que tome decisiones fatales en mi vida. Yo creo que no estoy preparada para usar las redes sociales, entonces las redes sociales también incluyen el hecho de que nosotros podamos trabajarlos en familia. Hay muchos hogares, padres de familia, en donde no hay la supervisión de las redes sociales en los teléfonos de los chicos.

Hablando de que de que un niño de 8 años en adelante debe tener supervisión para tener un teléfono y tener una red social porque no tiene la madurez necesaria. Y ahora en las redes sociales hay bastantes plataformas sexuales, bastantes plataformas en donde conllevan a

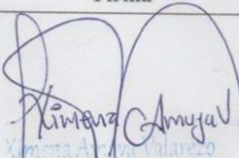
que los niños hagan daño o que haya bastante. Hay bastantes actos de violencia y los niños cuando no tienen una buena estructura mental porque hay deficiencia en su hogar como problemas, conflictos, adicciones, conllevan a que no tengan una estabilidad emocional y una buena gestión de sus emociones y que cometan actos atroces como matar, hacerse daño. Hace como dos años hubo una situación de un juego virtual que hacían con una ballena, de que mientras dibujaban una ballena ellos tenían que dibujarse la ballena en el brazo y se iban cortando el brazo, pero ellos no se daban cuenta de que se estaban haciendo daño e inclusive podían provocarse la muerte. Entonces yo creo que el tema de las redes sociales es un tema bastante complejo. Tenemos que llevar mucha supervisión en el caso de los niños y adolescentes y creo que podría ser conllevado a que a que tengamos una buena familiarización con las redes sociales, siempre y cuando nosotros tengamos. Inteligencia emocional, y sepamos qué es lo bueno, que es lo malo y que podamos aceptar las opiniones del resto.



**UNIVERSIDAD NACIONAL DE LOJA**  
Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables  
Carrera de Ingeniería en Sistemas

Constancia de entrevista:

Entrevista realizada por: Byron Stalin Montaña Beltrán

Nombre	Cargo	Firma
Ximena Dennisse Amaya Valarezo	Psicóloga Clínica	 Ximena Amaya Valarezo PSICÓLOGA CLÍNICA No. Reg. 1052-13-4359

## **Anexo 2.** Encuesta dirigida al Psicólogo de la Unidad de Bienestar Estudiantil de la UNL.

### **Encuesta**

**Cargo:** PSICOLOGO DE LA UNIDAD DE BIENESTAR UNIVERSITARIO

**Nombre:** JORGE FERNANDO JIMENEZ SANCHEZ

**Fecha de encuesta:** 17 DE ENERO DE 2022

### **Descripción:**

#### **1. ¿Considera a la depresión como un problema de salud serio?**

La depresión es causada por una combinación de factores genéticos, biológicos, ambientales y psicológicos. Por lo tanto, es un serio problema de salud.

#### **2. ¿Cree que las medidas de restricción tomadas como consecuencia del covid ha repercutido negativamente sobre la salud mental de las personas en Ecuador?**

Si. Por no haber un plan de contingencia adecuado, en donde la parte psicológica acompañe a la médica para poder reducir el número de personas deprimidas, ansiosas, nerviosas, acompañadas de llanto etc. Apoyados con planes terapéuticos.

#### **3. ¿Existe estigma o renuencia asociado con buscar ayuda profesional para los problemas mentales, incluyendo la depresión?**

Si existe, debido a que todo el mundo piensa que no necesita de ayuda profesional y por ende en estos casos las personas se guardan sus problemas emocionales lo que en poco tiempo se agravan hasta llegar a la depresión.

#### **4. ¿Considera que las redes sociales como Twitter pueden ser un medio para que las personas que sufren depresión puedan expresar sus sentimientos y opiniones?**

Las personas depresivas tienen un lenguaje más pobre: sienten menos interés por lo que les rodea y eso hace que no usen frases enteras para expresarse.

También son personas que retuitean menos, interactúan menos con su entorno, y que usan palabras con polaridad negativa.




## UNIVERSIDAD NACIONAL DE LOJA

Facultad de la Energía, las Industrias y los Recursos Naturales no Renovables

Carrera de Ingeniería en Sistemas

Constancia de encuesta:

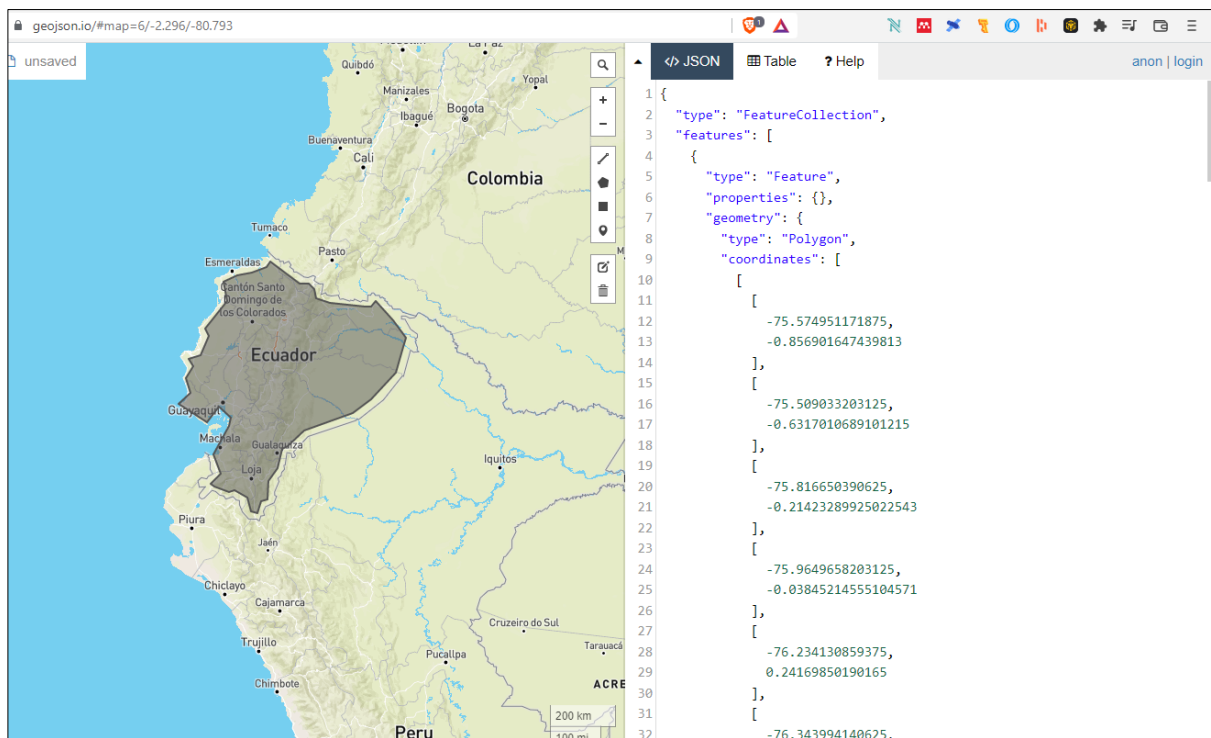
**Encuesta realizada por:** Byron Stalin Montaña Beltran

Nombre	Cargo	Firma
Jorge Fernando Jiménez Sánchez Mg. Sc.	Psicólogo de la Unidad de Bienestar Universitario	 Firmado electrónicamente por: JORGE FERNANDO JIMENEZ SANCHEZ



### Anexo 3. Geodatos del territorio de Ecuador.

Geodatos generados en Geojson.io para el territorio de Ecuador.



**Figura 53.** Geodatos de Ecuador obtenidos mediante la herramienta geojson.io

Para ver el archivo Geojson que se generó del territorio de Ecuador que se muestra en la figura 53, está disponible en el siguiente link:

[https://github.com/byronmb/Identificacion\\_Depresion\\_Ecuador/blob/main/1.Extraccion\\_Tweets/area\\_ecuador.geojson](https://github.com/byronmb/Identificacion_Depresion_Ecuador/blob/main/1.Extraccion_Tweets/area_ecuador.geojson)



## Anexo 5. Certificado de traducción.

**TITLE:** Analysis of Feelings on Twitter for the Identification of Depression in Times of COVID-19 in Ecuador.

**Author:** Byron Stalin Montaña Beltran

**Tutor:** Luis Antonio Chamba Eras, Ph.D.

### ABSTRACT

The COVID-19 pandemic drastically changed the way people lived, wreaking further devastation on issues like the economy and health, coupled with restrictions like lockdown and social distancing. In addition to the increase in death toll, it negatively affected people's lives. Mental health has become a growing concern due to the termination of employment, loss of income, family stress and other uncertainties that have led to serious consequences such as depression, which is considered by the WHO as the main factor that contributes to global disability. Social networks such as Twitter have currently become a universal means of communication and expression of opinions and feelings, which is why they have played a fundamental role during closures and quarantines. For this reason, this Thesis is focused on developing automatic learning models that make it possible to identify linguistic patterns obtained from Twitter posts and determine if a user is depressed or not. For this purpose, tweets from the year 2020 and 2021 were collected during the pandemic, using keywords through the Twint scraping tool, and they were labeled as depressive and random (not depressive); in addition, a manual cleaning was carried out to guarantee the quality of the collected data set. The Support Vector Machines, Random Forest and Naive Bayes algorithms were applied to the tf-idf representation of the text data and the characteristic of the texts was obtained together with N-grams to evaluate the models with the best performance for each algorithm type. The performance of the best models was compared by applying the Precision, Recall, F1-score and Accuracy evaluation metrics. The results showed that the Random Forest classifier presents acceptable results, registering an F1 score of 95.4% and an accuracy of 95.6%. Finally, the model with the best performance was used to predict the feelings in the unlabeled tweets of the year 2019, and it was possible to determine that from the year 2019 to the year 2020 there was a percentage increase in depressive publications of 31.17%.

**Keywords:** mental health, quarantine, twint, machine learning, tf-idf, N-grams.

I, Giovanni Arias Yapool, qualified expert translator with ID card number 1723708101, hereby certify that the above is a faithful translation to the original document - from Spanish into English. Phone: +593989243685. Email: [geovannyarias yapool@outlook.es](mailto:geovannyarias yapool@outlook.es) - [www.geovannyarias yapool.com](http://www.geovannyarias yapool.com). Senescyt accreditation number: 1005-2017-1821565. You can validate the professional qualification on the following web site: <https://www.senescyt.gob.ec/web/guest/consultas>

Lic. Giovanni Arias Yapool  
Expert Translator  
ID: 1723708101  
Senescyt Accreditation  
N° 1005 - 2017 - 1821565

Quito, March 25, 2023



**CERTIFICATE OF OFFICIAL TRANSLATION**  
**CERTIFICADO DE TRADUCCIÓN OFICIAL**

I, Giovanni Arias Yapool, qualified expert translator with ID card number 1723708101, hereby certify that the above is a faithful translation to the original document - from Spanish into English. Phone: +593989243685. Email: [geovannyariasyapool@outlook.es](mailto:geovannyariasyapool@outlook.es) – [www.giovannyariasyapool.com](http://www.giovannyariasyapool.com). Senescyt accreditation number: 1005-2017-1821565. You can validate the professional qualification on the following web site: <https://www.senescyt.gob.ec/web/guest/consultas>

Yo, Lic. Giovanni Arias Yapool, traductor calificado con número de cédula: 1723708101, certifico que lo anterior es una traducción español-inglés exacta y fiel a la del documento original. Telf.: 0989243685. Email: [geovannyariasyapool@outlook.es](mailto:geovannyariasyapool@outlook.es). [www.giovannyariasyapool.com](http://www.giovannyariasyapool.com). Acreditación Senescyt número: 1005-2017-1821565. Usted puede verificar el título habilitante profesional en la siguiente página web: <https://www.senescyt.gob.ec/web/guest/consultas>

Documents - Documentos	Nº of pages - páginas
<b>Abstract - Title:</b> Analysis of Feelings on Twitter for the Identification of Depression in Times of COVID-19 in Ecuador. <b>Author:</b> Byron Stalin Montaña Beltran. <b>Tutor:</b> Luis Antonio Chamba Eras, Ph.D.	1 page - página
<b>TOTAL:</b>	<b>1 pages - páginas</b>



Validation of qualified expert translator: <https://www.senescyt.gob.ec/web/guest/consultas>

Senescyt Registration Number: 1005-2017-1821565

Número de registro Senescyt: 1005-2017-1821565

Quito – Ecuador



Lic. Giovanni Arias Yapool  
Expert Translator  
ID: 1723708101  
Senescyt Accreditation  
Nº 1005 - 2017 - 1821565