MODELO BASADO EN TÉCNICAS DE NLP Y ANÁLISIS ACÚSTICO PARA DETECCIÓN TEMPRANA DE POSIBLES PACIENTES CON TENDENCIA AL SUICIDIO

KRISTELL RASHEL URUETA PÁEZ

MONOGRAFÍA PARA OPTAR AL TÍTULO DE MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

DIRECTOR: JOSÉ DUVÁN MÁRQUEZ DÍAZ

UNIVERSIDAD DEL NORTE
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
DEPARTAMENTO DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BARRANQUILLA, COLOMBIA
ENERO-2025

Aprobado por el profesorado de la División de Ingenierías en cumplimiento de los requisitos exigidos para otorgar el título de Magíster en Ingeniería de Sistemas y Computación.

_____

Kristell Rashel Urueta Páez

_____

José Duván Márquez Díaz

_____

Miguel Ángel Jimeno Paba

Barranquilla, 30 de enero del 2025

AGRADECIMIENTOS

Agradezco de manera especial a la Universidad del Norte y a los docentes que me acompañaron durante mis estudios de pregrado y posgrado, pues su dedicación y guía me proporcionaron las bases necesarias para la realización de este trabajo.

Expreso mi profunda gratitud al profesor y asesor de esta tesis, José Duván Márquez, por su constante apoyo, conocimiento y confianza a lo largo de este proceso.

Asimismo, quiero reconocer a mi familia por su respaldo incondicional en cada etapa de mi formación académica avanzada, así como por sus consejos y buenos deseos. En particular, agradezco a mi esposo, futura hija y a mis padres, cuya motivación fue fundamental para la culminación de esta meta.

TABLA DE CONTENIDO

LISTADO DE FIGURAS

LISTADO DE TABLAS

ABREVIATURAS

AI	Artificial Intelligence
API	Application Programming Interface
NLP	Natural Language Processing
ML	Machine Learning
LDA	Latent Dirichlet Allocation
NaN	Not a Number
TP	True Positive
TN	True Negative
TPR	True Positive Rate
FPR	False Positive Rate
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
TF-IDF	Term Frequency – Inverse Document Frequency
CSV	Comma Separated Values
GPT	Generative Pre-trained Transformer
NLTK	Natural Language Toolkit
VADER	Valence Aware Dictionary and sEntiment Reasoner
TCC	Terapia Cognitivo Conductual
MFCC	Mel Frequency Cepstral Coefficients
EHR	Electronic Health Records
GD	Gradient Descent
C-SSRS	Columbia-Suicide Severity Rating Scale
SHAP	SHapley Additive exPlanations
GDPR	General Data Protection Regulation
HIPAA	Health Insurance Portability and Accountability Act
EEG	Electroencephalogram
GPU	Graphics Processing Unit
RNN	Recurrent Neural Networks
BER	Bidirectional Encoder Representations from Transformers
DistilBERT	Distilled BERT
RMS	Root Mean Square
RNN	Recurrent Neural Network
LR	Logistic Regression

RESUMEN

En el ámbito de la salud mental, la detección temprana de comportamientos suicidas es fundamental para prevenir posibles tragedias y brindar apoyo oportuno a las personas en riesgo. No obstante, la complejidad de los factores involucrados y la falta de herramientas de evaluación eficientes hacen que muchos casos pasen inadvertidos. Con el propósito de mitigar este problema, la presente tesis desarrolla un modelo basado en técnicas de Procesamiento del Lenguaje Natural (NLP) y análisis de audio para la detección temprana de posibles tendencias suicidas, proporcionando así un primer canal de apoyo a los pacientes y asistiendo a profesionales de la salud en la toma de decisiones preventivas.

La primera fase de la investigación emplea algoritmos como VADER, metodologías de NLTK y Latent Dirichlet Allocation (LDA), para clasificar e identificar temáticas subyacentes asociadas a preocupaciones específicas. En la segunda fase, se presenta un modelo avanzado que integra la información textual con atributos de audio (entonación, ritmo, índices de energía, entre otros) obtenidos de grabaciones de voz de posibles pacientes y expertos en salud mental. Esta combinación de datos multimodales se procesa mediante técnicas de clasificación (Logistic Regression, Random Forest, entre otras), entrenadas sobre un conjunto de datos cuidadosamente compilados que incluyen tanto textos de redes sociales como audios transcritos. Así, se comparan ambos enfoques, evidenciando un aumento significativo en la capacidad de detección de riesgo gracias a la fusión de características textuales y acústicas.

El flujo de la solución consiste en recibir la información del paciente a través de una aplicación, procesar dichos datos con el modelo unificado y, finalmente, emitir un diagnóstico basado en atributos clave. Los resultados obtenidos demuestran la eficacia de combinar técnicas de NLP con el análisis de audio para identificar indicadores de riesgo suicida de forma anticipada, contribuyendo así a la intervención oportuna y al fortalecimiento de la atención en salud mental.

ABSTRACT

In the field of mental health, early detection of suicidal behavior is essential to prevent potential tragedies and to provide timely support to people at risk. However, the complexity of the factors involved, and the lack of efficient assessment tools mean that many cases go unnoticed. To mitigate this problem, the present thesis develops a model based on Natural Language Processing (NLP) techniques and audio analysis for the early detection of possible suicidal tendencies, thus providing a first channel of support to patients and assisting health professionals in preventive decision-making.

The first phase of the research employs algorithms such as VADER, NLTK, and Latent Dirichlet Allocation (LDA) methodologies, to classify and identify underlying themes associated with specific concerns. In the second phase, an advanced model is presented that integrates textual information with audio attributes (intonation, rhythm, energy indexes, among others) obtained from voice recordings of potential patients and mental health experts. This combination of multimodal data is processed using classification techniques (Logistic Regression, Random Forest, among others), trained on a carefully compiled dataset that includes social network texts and transcribed audio. Thus, both approaches are compared, evidencing a significant increase in risk detection capability thanks to the fusion of textual and acoustic features.

The solution flow consists of receiving patient information through an application, processing this data with the unified model, and issuing a diagnosis based on key attributes. The results obtained demonstrate the effectiveness of combining NLP techniques with audio analysis to identify suicide risk indicators early, thus contributing to timely intervention and strengthening mental health care.

I. INTRODUCCIÓN

La salud mental constituye uno de los pilares fundamentales del bienestar humano y su atención adecuada es esencial para prevenir trastornos que pueden derivar en conductas autodestructivas, como el suicidio. Según la Organización Mundial de la Salud (OMS), anualmente, aproximadamente 700,000 personas fallecen a causa del suicidio, y millones más intentan quitarse la vida, lo que genera un impacto significativo tanto a nivel individual como colectivo [1]. Esta problemática afecta de manera desproporcionada a ciertos grupos demográficos, especialmente a jóvenes de entre 15 y 29 años, donde el suicidio ocupa el cuarto lugar entre las principales causas de muerte. Además, el impacto sociocultural y económico del suicidio trasciende a las víctimas directas, dejando cicatrices profundas en sus familias, comunidades y redes sociales. Estos datos subrayan la urgencia de desarrollar estrategias efectivas para prevenir el suicidio y mitigar sus consecuencias. En este contexto, la detección temprana de comportamientos suicidas emerge como una necesidad crítica, permite intervenir oportunamente y salvar vidas. Sin embargo, los desafíos asociados a la identificación de estos comportamientos residen en la complejidad de los factores psicosociales y la falta de herramientas accesibles y eficientes que permitan un diagnóstico inicial oportuno.

En los últimos años, los avances en tecnologías como el Procesamiento del Lenguaje Natural (Natural Language Processing, NLP) han transformado la forma en que se analiza la comunicación humana, ofreciendo nuevas posibilidades para explorar patrones lingüísticos y emocionales asociados con el riesgo suicida. Estudios recientes han demostrado que ciertas características del lenguaje, como el uso frecuente de términos negativos, referencias al aislamiento social o una marcada falta de esperanza, pueden ser indicadores significativos de pensamientos suicidas [2] [3]. Asimismo, el análisis de datos multimodales, que integra señales acústicas como la entonación y el ritmo del habla, ha mostrado un gran potencial para complementar el diagnóstico al identificar señales emocionales difíciles de capturar solo a través del texto [4].

A pesar de estos avances, el uso de modelos basados en NLP y análisis de audio para la detección temprana de tendencias suicidas aún enfrenta importantes limitaciones, entre las que destacan la disponibilidad de conjuntos de datos representativos, la necesidad de modelos que combinen múltiples fuentes de información y la adaptación de estas herramientas a contextos clínicos reales. Este trabajo se inscribe en ese marco de investigación al desarrollar un modelo basado en técnicas de NLP y análisis acústico, con el propósito de ofrecer una herramienta que potencie la

detección temprana y actúe como un canal de apoyo para pacientes y profesionales de la salud.

Esta investigación deriva del proyecto titulado *"Desarrollo de un Algoritmo basado en el Habla para el Pronóstico de Intento de Suicidio en Adultos Jóvenes Colombianos: Un Estudio Piloto*", el cual fue aprobado por el Comité de Ética en Investigación en el Área de la Salud de la Universidad del Norte el 30 de noviembre de 2023, bajo el acta número 303 y el radicado 2311-607. Este proyecto busca desarrollar un modelo basado en técnicas de NLP y análisis acústico con el propósito de ofrecer una herramienta que potencie la detección temprana y actúe como un canal de apoyo para pacientes y profesionales de la salud.

En este contexto, el presente trabajo tiene como objetivo complementar y fortalecer esta línea de investigación mediante el desarrollo de un modelo que no solo identifique patrones textuales y acústicos asociados al riesgo suicida, sino que también los integre en un enfoque multimodal para mejorar la precisión diagnóstica. A través de una metodología robusta y el uso de herramientas avanzadas como VADER, Latent Dirichlet Allocation (LDA) y clasificadores de aprendizaje automático.

El modelo propuesto podría servir como un insumo clave para la evolución del proyecto anteriormente mencionado, permitiendo mejorar la capacidad de los sistemas de atención para identificar y mitigar riesgos de manera proactiva.

II. GENERALIDADES

1. Justificación

La presente investigación responde a la necesidad de abordar el suicidio desde una perspectiva innovadora, que aproveche las capacidades de las tecnologías emergentes para superar las limitaciones de los métodos tradicionales de detección y prevención. Aunque los enfoques convencionales han contribuido significativamente al entendimiento del comportamiento suicida, su dependencia de herramientas clínicas y su alcance limitado dificultan su aplicabilidad en contextos donde el acceso a servicios especializados es escaso o inexistente. Por ello, resulta imprescindible explorar alternativas tecnológicas que integren análisis automatizados y accesibles para identificar riesgos de manera temprana y eficiente.

El empleo del Procesamiento del Lenguaje Natural (NLP) y el análisis de audio en el campo de la salud mental ofrece una oportunidad sin precedentes para mejorar la detección de tendencias suicidas al analizar patrones complejos en el lenguaje y las señales acústicas. Sin embargo, su integración práctica en sistemas accesibles y adaptados a poblaciones diversas aún es incipiente. Esta brecha tecnológica representa un área crítica de intervención, donde las soluciones basadas en datos multimodales pueden desempeñar un papel transformador.

Además, la tesis tiene un propósito estratégico al contribuir al desarrollo de herramientas escalables y de bajo costo, que puedan ser implementadas tanto en entornos clínicos como en plataformas digitales de apoyo. Este enfoque no solo amplía el alcance de las intervenciones, sino que también abre nuevas posibilidades para la personalización de las estrategias preventivas, permitiendo una atención más precisa y centrada en el individuo.

Finalmente, esta investigación busca no solo mitigar el impacto del suicidio, sino también sentar las bases para el desarrollo de sistemas de salud mental más robustos e inclusivos. Al abordar una problemática global con enfoques interdisciplinarios e innovadores, la tesis se posiciona como una contribución relevante tanto para la comunidad científica como para los esfuerzos prácticos de prevención y cuidado. Es importante destacar que, aunque este estudio puede ser útil para el análisis y apoyo en salud mental; sin embargo, este no sustituye la evaluación y tratamiento proporcionados por profesionales de la salud. Se recomienda utilizar estas herramientas como complemento y, ante cualquier indicio de trastorno mental, o pensamiento que puedan atentar contra su vida, buscar atención especializada.

2. Objetivos

a. Objetivo general

Desarrollar un modelo computacional basado en técnicas de Procesamiento del Lenguaje Natural (NLP) y análisis acústico, capaz de detectar de manera temprana posibles tendencias suicidas en individuos, integrando datos textuales y de audio en un enfoque multimodal que permita identificar patrones lingüísticos y emocionales asociados al riesgo, con el propósito de facilitar la intervención preventiva y contribuir al fortalecimiento de las estrategias de salud mental en contextos clínicos y no clínicos.

b. Objetivos específicos

- Desarrollar un proceso de recopilación y estructuración de datos textuales provenientes de redes sociales, con el propósito de identificar patrones lingüísticos asociados a tendencias suicidas.
- Construir un conjunto de datos multimodal que integre grabaciones de audio de posibles pacientes, personas sin antecedentes de riesgo y expertos en salud mental, capturando parámetros clave relacionados con el estado emocional y mental.
- Implementar técnicas avanzadas de refinamiento y clasificación de datos, tanto textuales como acústicos, utilizando metodologías robustas para garantizar la calidad y representatividad del dataset.
- Diseñar y entrenar un modelo unificado que combine datos textuales y de audio, empleando enfoques multimodales para maximizar la precisión en la detección temprana de tendencias suicidas.
- Validar el desempeño del sistema completo mediante métricas que permitan evaluar su efectividad y aplicabilidad en contextos reales de salud mental bajo un conjunto de supuestos definidos.
- Desarrollar una aplicación con una interfaz intuitiva que permita procesar los datos del usuario final a través del modelo y retornar un diagnóstico confiable, asegurando una experiencia accesible y funcional.

3. Descripción del problema de investigación

La detección temprana de tendencias suicidas es un desafío crítico en la salud mental que carece de herramientas específicas, accesibles y automatizadas para apoyar el proceso de evaluación. Actualmente, las evaluaciones dependen de entrevistas

clínicas y escalas subjetivas que no son escalables ni sistemáticas, lo que limita su capacidad para identificar patrones de riesgo en grandes volúmenes de datos.
A pesar de la disponibilidad de datos textuales y de audio provenientes de mensajes, conversaciones o transcripciones, no existen suficientes sistemas que integren el análisis de características lingüísticas, emocionales y acústicas mediante algoritmos de Machine Learning y Deep Learning. Esta ausencia dificulta el aprovechamiento de estos datos para una detección eficaz y reproducible del riesgo suicida.

Esta investigación busca abordar esta problemática desarrollando un modelo predictivo que combine análisis de texto y audio, diseñado para su uso en entornos digitales y accesible tanto para profesionales especializados como para usuarios con formación básica, contribuyendo a la identificación temprana y oportuna de casos críticos.

III. MARCO TEÓRICO

1. Machine Learning (ML)

Machine Learning es un campo de la inteligencia artificial enfocado en desarrollar algoritmos que permitan a las computadoras aprender de los datos y tomar decisiones o realizar predicciones sin ser explícitamente programadas para cada tarea específica. A lo largo de su evolución, ML ha sido aplicado en diversos dominios, mejorando constantemente en precisión y eficacia mediante el desarrollo de nuevos algoritmos y técnicas. Samuel [5] fue uno de los pioneros en este campo, utilizando el aprendizaje automático para mejorar el rendimiento en juegos como las damas, lo que sentó las bases para estudios futuros en ML.

Dentro del campo de Machine Learning, existen distintos paradigmas para el entrenamiento de modelos, entre los que se encuentran:

a. Aprendizaje supervisado

En esta subcategoría de ML, el modelo se entrena utilizando un conjunto de datos etiquetados. Esto significa que para cada entrada del conjunto de entrenamiento se conoce la salida deseada, y el algoritmo aprende a mapear entradas a salidas basándose en estos ejemplos. Kotsiantis et al. [6] y Singh et al. [7] revisan diversas técnicas y algoritmos de clasificación supervisada, resaltando sus aplicaciones prácticas y comparando la eficacia de distintos métodos. Estas técnicas incluyen, entre otras, árboles de decisión, máquinas de soporte vectorial y redes neuronales, que son comúnmente utilizados para resolver problemas de clasificación y regresión en diferentes ámbitos.

b. Aprendizaje no supervisado

El aprendizaje no supervisado se refiere a métodos de ML en los que los datos no vienen con etiquetas. El objetivo es descubrir patrones, agrupamientos o estructuras inherentes dentro de los datos sin una guía explícita sobre cuál debe ser la salida. Gentleman y Carey [8] analizan casos de estudio en biología computacional utilizando técnicas no supervisadas para identificar patrones en datos biológicos. Entre los algoritmos más conocidos de esta categoría están el clustering (p. ej., K-means, DBSCAN) y técnicas de reducción de dimensionalidad (p. ej., PCA), que ayudan a organizar y representar la información de manera más comprensible.

c. Aprendizaje por refuerzo

Esta rama de ML consiste en la que un agente aprende a tomar decisiones secuenciales interactuando con un entorno. El agente recibe recompensas o penalizaciones basadas en las acciones que toma, y su objetivo es aprender una política que maximice la recompensa acumulada a lo largo del tiempo. Distintos autores [9] [10] proporcionan encuestas y descripciones profundas de los fundamentos teóricos y prácticos del aprendizaje por refuerzo. Estos trabajos explican conceptos como la función de valor, la función de política, la exploración vs. explotación y algoritmos específicos (como Q-learning y métodos basados en políticas) que son esenciales para el desarrollo de agentes inteligentes en una variedad de aplicaciones, desde juegos hasta robótica y sistemas de recomendación.

2. Conjunto de datos de Entrenamiento y Prueba

Un conjunto de datos se divide comúnmente en dos subconjuntos principales: el de entrenamiento y el de prueba. El **conjunto de entrenamiento** es la porción de datos que se utiliza para enseñar al modelo a reconocer patrones y aprender la relación entre las características y la variable objetivo. Una vez entrenado, el modelo se evalúa usando el **conjunto de prueba**, que contiene datos que no han sido vistos durante el entrenamiento. Este proceso permite medir la capacidad de generalización del modelo a nuevos datos, asegurando que no se haya sobreajustado a los datos de entrenamiento [7] [11].

3. Muestras o Instancias

En el contexto del aprendizaje automático, una **muestra** o **instancia** se refiere a un único elemento o punto de datos dentro de un conjunto más amplio. Cada instancia representa una observación individual y se compone de un conjunto de características o atributos. Por ejemplo, en un conjunto de datos sobre pacientes, cada paciente sería una instancia con características como edad, síntomas y diagnósticos [7] [11].

4. Características

Las **características** (también llamadas atributos o variables independientes) son las propiedades medibles o valores descriptivos que se obtienen de cada muestra en un conjunto de datos. Estas características son utilizadas por los modelos de aprendizaje automático para aprender patrones y hacer predicciones. Por ejemplo, en un conjunto de datos para clasificación de flores, las características podrían

incluir la longitud y el ancho de los pétalos y sépalos [11] .La calidad y relevancia de estas características influyen directamente en el rendimiento del modelo.

5. Variable Objetivo

La variable objetivo (también conocida como etiqueta o variable dependiente) es el resultado o valor que el modelo intenta predecir o clasificar. Es la información que se conoce durante el entrenamiento y que se usa como referencia para ajustar el modelo. Por ejemplo, en un problema de detección de enfermedades, la variable objetivo podría ser la presencia o ausencia de una enfermedad en un paciente. El objetivo del aprendizaje supervisado es aprender una función que mapee las características de entrada a esta variable objetivo de manera precisa [7] [11].

6. Clases

En problemas de clasificación, las **clases** representan los distintos grupos o categorías en los que pueden clasificarse las instancias de un conjunto de datos. Cada muestra se asigna a una clase en función de sus características. Por ejemplo, en un problema de clasificación de correos electrónicos, las clases podrían ser "spam" y "no spam" [6] [12].

7. Pipeline

Un **pipeline** en aprendizaje automático es una secuencia de pasos de procesamiento de datos y modelado que automatiza flujos de trabajo complejos. Permite encadenar transformaciones y estimadores, facilitando la reproducibilidad y la optimización del modelo [13].

8. Función de Costo

Es una función J que se utiliza para determinar el error entre los valores estimados ($\hat{y}$) y los valores reales (y), con el fin de optimizar los parámetros de una red neuronal. La función por seleccionar depende del problema que se esté trabajando. La función mayormente utilizada en los problemas de clasificación es la entropía cruzada [14].

9. Entrenamiento del Modelo

El entrenamiento del modelo es el proceso mediante el cual un algoritmo de aprendizaje automático aprende a partir de un conjunto de datos de entrenamiento.

Durante este proceso, el modelo ajusta sus parámetros internos para minimizar el error en sus predicciones, basándose en la relación entre las características de entrada y la variable objetivo. Este proceso se repite iterativamente hasta que el modelo alcanza un rendimiento óptimo. Típicamente, utiliza el algoritmo de gradiente descendente [14] [15] para la obtención de estos valores a través de los procesos de forward propagation y backpropagation.

a. Gradiente Descendente

Es un algoritmo iterativo de optimización de primer orden utilizado para encontrar un mínimo/máximo local de una función determinada [15]. En el contexto del aprendizaje automático se usa para minimizar la función de costo/pérdida (loss).
Durante la ejecución del algoritmo, se calcula iterativamente el siguiente punto usando el gradiente en la posición actual escalado por la tasa de aprendizaje (learning rate) restándolo del valor de la posición actual. El cálculo de los gradientes se realiza mediante los procesos del forward propagation y backpropagation a través de:

$$x_{n+1} = x_n - \alpha \cdot \nabla f(x_n)$$

Este proceso se repite un número definido de iteraciones determinadas por el valor de las épocas.

b. Forward Propagation

Forward propagation es uno de los procesos claves a la hora de realizar el entrenamiento de una red neuronal mediante el algoritmo del gradiente descendente (GD). Durante este proceso la red toma los datos de la capa de entrada y los transmite a la siguiente capa donde se calculan sus propias activaciones. Esta a su vez transmite sus activaciones a la capa siguiente repitiendo el proceso hasta la última capa, obteniendo así la predicción correspondiente en la capa de salida [15].

c. Backpropagation

Es el segundo proceso clave que se realiza durante el entrenamiento de una red neuronal mediante el algoritmo de gradiente descendente (GD). Durante este se busca obtener los gradientes de los parámetros (pesos y sesgos) para optimizar (típicamente minimizar) la función de costo J. Por lo tanto, para lograr este propósito se utiliza la regla de la cadena para calcular las derivadas deseadas [14] [15].

Este proceso toma como principal dato de entrada el valor de las predicciones, que luego utiliza para calcular las derivadas parciales de los parámetros respecto a la función de costo haciendo uso de la regla de la cadena.

10. Métricas

Las **métricas** son medidas cuantitativas utilizadas para evaluar el rendimiento de un modelo de aprendizaje automático. Dependiendo del tipo de problema (clasificación, regresión, etc.), las métricas pueden incluir matriz de confusión, precisión, exactitud, recall, F1-score, entre otras. Estas métricas permiten comparar distintos modelos y ajustar sus hiperparámetros para mejorar el desempeño [16] [12].

a. Matriz de Confusión

Es una matriz que describe el desempeño completo de un modelo [14] [17]. Cada fila de la matriz representa las instancias de los valores (clases) reales, mientras que cada columna representa las instancias de los valores (clases) predichos, aunque también existen variantes en donde se invierten las filas y las columnas. Supongamos una clasificación binaria donde tenemos dos clases (Positivo y Negativo), la matriz de confusión está dividida en cuatro sectores y estos son:

- Verdaderos positivos (TP): Es la cantidad de casos donde el valor real es positivo y el modelo predijo también que era positivo.
- Verdaderos negativos (TN): Es la cantidad de casos donde el valor real es negativo y el modelo predijo también que era negativo.
- Falsos negativos (FN): Es la cantidad de casos donde el valor real es positivo, pero el modelo predijo que era negativo. También es conocido como el error de tipo II.
- Falsos positivos (FP): Es la cantidad de casos donde el valor real es negativo, pero el modelo predijo que era positivo. También es conocido como el error de tipo I.

b. Exactitud (Accuracy)

Se define como la proporción de predicciones correctas respecto al total de muestras [17]. Esta se calcula de la siguiente manera:

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

c. Sensibilidad (Sensitivity or Recall)

Se define como la probabilidad que una muestra positiva sea realmente positiva, o qué porcentaje de los casos positivos fueron capturados; es decir, es la razón de las predicciones de verdaderos positivos sobre la totalidad de las predicciones positivas (suma de las predicciones de verdaderos positivos y falsos negativos) [17].

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

d. Precisión

Se define como la razón de las predicciones de verdaderos positivos sobre la suma de las predicciones de verdaderos positivos y falsos positivos, o qué porcentaje de las predicciones positivas son correctas [17].

$$PPV = \frac{TP}{TP + FP}$$

e. F1-Score

Es la media armónica de la precisión y el recall, su valor está entre 1 y 0, y se calcula de la siguiente manera:

$$F1 \ = \ 2 \times \frac{Pr\,e\,cision \times Recall}{Pr\,e\,cision \ + \ Recall}$$

Esta fórmula permite equilibrar ambos valores, penalizando fuertemente cuando uno de ellos es bajo. Así, el F1-score sólo será alto si tanto la precisión como la exhaustividad son altas [18].

f. Especificidad (Specificity)

Se define como la probabilidad que una muestra negativa sea realmente negativa, es decir, es la razón de las predicciones de verdaderos negativos sobre la totalidad de las predicciones negativas (suma de las predicciones de verdaderos negativos y falsos positivos) [17].

$$TNR = \frac{TN}{TN + FP}$$

g. ROC-AUC

Esta métrica describe el área bajo la curva (AUC) Característica Operativa del Receptor (ROC). Esta curva es una representación gráfica de la sensibilidad (TPR) contra uno menos la especificidad (1 - TNR) para un sistema clasificador binario según se varía un umbral de discriminación, que es el valor a partir del cual decidimos que un caso es un positivo [19].

11. Scikit-learn

Scikit-learn es una biblioteca de Python ampliamente utilizada para implementar algoritmos de aprendizaje automático. Proporciona herramientas simples y eficientes para el análisis predictivo y la minería de datos, incluyendo clasificación, regresión, clustering, reducción de dimensionalidad y selección de modelos, entre otras funciones. Es conocida por su facilidad de uso y buena documentación, lo que facilita la experimentación y el desarrollo de modelos [13].

12. pipeline.fit()

El método **pipeline.fit()** en scikit-learn entrena (ajusta) todos los pasos definidos en un pipeline sobre un conjunto de datos. Encadena transformaciones y la estimación final, facilitando la experimentación y validación del modelo [13].

13. FeatureUnion

Es un componente de scikit-learn que combina múltiples transformadores en paralelo, uniendo sus salidas. Esto permite extraer y concatenar características de diferentes fuentes o mediante diferentes técnicas para alimentar a un modelo [20].

14. Librosa

Es una biblioteca de Python especializada en el análisis y procesamiento de audio. Proporciona herramientas para extraer características acústicas, realizar transformadas de Fourier y manipular señales de audio, siendo ampliamente usada en investigación musical y procesamiento de audio [21].

15. NLP

El Procesamiento del Lenguaje Natural (NLP) es un campo de la inteligencia artificial que se centra en la interacción entre computadoras y el lenguaje humano. Los sistemas de NLP permiten a las máquinas interpretar, analizar y generar texto y

habla, facilitando tareas como traducción automática, análisis de sentimiento y reconocimiento de voz [22].

16. Análisis de sentimiento

El **análisis de sentimiento** es una técnica de NLP que identifica y clasifica opiniones expresadas en texto para determinar la actitud del autor respecto a un tema, ya sea positiva, negativa o neutral. Se utiliza comúnmente para analizar reseñas de productos, comentarios en redes sociales y feedback de clientes [23].

17. Transformer

Los **Transformers** son una arquitectura de redes neuronales basada en mecanismos de atención que permiten procesar secuencias de datos de forma paralela y capturar dependencias a larga distancia. Han revolucionado el campo del NLP y más allá [24]. Esta arquitectura ha revolucionado el campo del NLP, permitiendo modelos como BERT y GPT, y mejorando significativamente tareas de traducción, generación de texto y análisis semántico [25] [24].

18. Bidirectional Encoder Representations from Transformers (BERT)

BERT es un modelo de aprendizaje profundo basado en la arquitectura Transformer, desarrollado por Google. Se distingue por su capacidad para entender el contexto bidireccional de las palabras en una oración, lo que mejora significativamente el rendimiento en tareas de NLP como la clasificación de texto y la extracción de información [25].

19. Embeddings

Los **embeddings** son representaciones vectoriales de datos (típicamente palabras o frases) en espacios de alta dimensión. Capturan relaciones semánticas y sintácticas, permitiendo que algoritmos de aprendizaje automático trabajen con texto de forma efectiva [26].

20. DistilBERT

DistilBERT es una versión reducida y eficiente de BERT (Bidirectional Encoder Representations from Transformers), obtenida mediante técnicas de destilación del conocimiento. Mantiene gran parte del rendimiento de BERT con menos parámetros, siendo más rápido y ligero [27].

21. Sentence transformers

Sentence Transformers es un marco basado en modelos de lenguaje preentrenados (como BERT) que genera embeddings de oraciones. Estos embeddings facilitan tareas como la similitud semántica y la búsqueda de información, ofreciendo representaciones eficientes de texto [28].

22. VADER (Valence Aware Dictionary and Sentiment Reasoner)

VADER es una herramienta de análisis de sentimiento diseñada para textos provenientes de redes sociales. Utiliza un léxico y reglas afinadas para calcular puntajes de sentimiento, considerando la valencia (positivo o negativo) y la intensidad emocional de las palabras y expresiones [29].

23. Latent Dirichlet Allocation (LDA)

Modelo de tópicos probabilístico que descubre automáticamente temas latentes en una colección de documentos. Asume que cada documento se compone de una mezcla de varios temas y que cada tema se caracteriza por una distribución de palabras, permitiendo la agrupación de documentos por temas subyacentes [30].

24. GridSearch

Técnica para la optimización de hiperparámetros en la que se define un conjunto de valores a probar para cada hiperparámetro y se evalúa el rendimiento del modelo para cada combinación posible. Esto permite identificar la combinación que proporciona el mejor rendimiento según una métrica determinada [13].

25. Whisper

Modelo de reconocimiento de voz desarrollado por OpenAI. Está diseñado para transcribir audio a texto en múltiples idiomas de manera robusta y precisa, utilizando técnicas avanzadas de aprendizaje profundo basadas en arquitecturas de Transformers [31].

26. TF-IDF

TF-IDF es una técnica de ponderación en minería de texto que evalúa la importancia de una palabra en un documento dentro de un corpus. Calcula la frecuencia de

término (TF) multiplicada por la frecuencia inversa de documento (IDF), de modo que resalta palabras relevantes que no son comunes en todo el corpus [32].

Cáculo de TF:

$$tf(t,d) = \sum_{x \in d} fr(x,t)$$

donde t es el término o palabra a buscar en el documento d.

Cáculo de IDF:

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

Donde D es el número total de documentos y $DF(t)$ es el número documentos que contienen el término t. Por tanto,

$$DF(t) = |\{d|t \in d\}|$$

Nota = log es logaritmo natural

Cálculo de $TF{-}IDF$ en sklearn (Predeterminado)

$$tfidf(t,d) = tf(t,d) \times idf(t,D)$$

27. Regresión logística

Algoritmo estadístico y de aprendizaje automático utilizado para problemas de clasificación binaria. Modela la probabilidad de que una instancia pertenezca a una clase específica utilizando una función logística (sigmoide) sobre una combinación lineal de las características de entrada [33].

28. Redes neuronales recurrentes

Clase de redes neuronales diseñadas para procesar secuencias de datos. Poseen ciclos internos que les permiten retener información pasada, lo que las hace especialmente útiles para tareas de NLP, reconocimiento de voz y series temporales [34].

29. Hiperparámetros

Los **hiperparámetros** son parámetros ajustables que se establecen antes del proceso de entrenamiento de un modelo de aprendizaje automático. A diferencia de los parámetros internos del modelo que se aprenden a partir de los datos, los hiperparámetros controlan aspectos como la complejidad del modelo, la tasa de aprendizaje y la estructura de la red neuronal. Su elección adecuada es crucial para optimizar el rendimiento del modelo [6] [12].

a. Tasa de aprendizaje

Es un valor que controla la velocidad de las actualizaciones de los parámetros durante la ejecución del algoritmo de gradiente descendente. Afecta directamente la velocidad a la que el algoritmo converge en el punto óptimo, es decir:

- Si la tasa de aprendizaje es muy pequeña, el algoritmo tardará demasiado tiempo en converger.
- Si la tasa de aprendizaje es muy grande, puede que no se alcance el punto óptimo o que el algoritmo diverja.

b. Épocas

Es la cantidad de iteraciones que realiza el algoritmo de aprendizaje a través del conjunto de datos de entrenamiento.

c. Arquitectura

Es el hiperparámetro que determina la forma de la red neuronal, es decir, su estructura: cantidad y tipo de capas ocultas, sus conexiones y el número de neuronas y las funciones de activación de cada capa.

d. ngram_range en TD-IDF

Es un hiperparámetro en el vectorizador TF-IDF que define el rango de n-gramas (secuencias de n palabras) a considerar. Por ejemplo, ngram_range=(1,2) incluye unigramas y bigramas [35].

e. C en LogisticRegression

Hiperparámetro en **LogisticRegression** que controla la fuerza de la regularización. Valores más pequeños implican mayor regularización, ayudando a evitar el sobreajuste [35].

30. Random forest

Algoritmo de aprendizaje automático basado en la construcción de múltiples árboles de decisión durante el entrenamiento y la agregación de sus predicciones para mejorar la precisión y controlar el sobreajuste. Cada árbol se construye a partir de una muestra aleatoria del conjunto de datos y considera un subconjunto aleatorio de características en cada división [36].

31. Fusión de características y fusión de decisiones

La **fusión de características** (feature fusion) combina múltiples tipos de características extraídas de diferentes fuentes o modalidades antes de la fase de decisión, mientras que la **fusión de decisiones** (decision fusion) combina las salidas de varios clasificadores independientes para tomar una decisión final. Ambas técnicas buscan mejorar la robustez y precisión del sistema integrando información complementaria [37].

32. Reinforcement learning

Área del aprendizaje automático donde un agente aprende a tomar decisiones secuenciales a partir de la interacción con un entorno. El agente recibe recompensas o castigos basados en sus acciones y busca maximizar la recompensa total a largo plazo. Se utiliza en robótica, juegos y sistemas de recomendación, entre otros [10].

33. Scrapping

El **scrapping** (o web scraping) es la técnica de extraer información de sitios web de manera automatizada. Mediante scrapping se recolectan datos no estructurados de la web para su análisis posterior [38].

34. Polaridad de sentimientos

La **polaridad de sentimiento** se refiere a la orientación (positiva, negativa o neutral) de una opinión expresada en texto. Es una tarea en procesamiento de lenguaje natural (NLP) que clasifica las emociones o actitudes subyacentes en textos [39].

35. GPU

Una **GPU** (Unidad de Procesamiento Gráfico) es un dispositivo especializado en cálculos paralelos masivos. Es ampliamente utilizado en entrenamiento de modelos

de aprendizaje profundo debido a su capacidad para acelerar operaciones matriciales y de tensor [14].

36. Transfer Learning

El **Transfer Learning** (aprendizaje por transferencia) implica reutilizar un modelo preentrenado en una tarea similar y adaptarlo a una nueva tarea, ahorrando recursos y tiempo de entrenamiento y aprovechando conocimientos previos [40].

37. Rule-based systems

Los **sistemas basados en reglas** son programas informáticos que utilizan un conjunto de reglas lógicas predefinidas para tomar decisiones o inferir conclusiones a partir de datos. Son interpretables y transparentes, aunque menos flexibles que los modelos de aprendizaje automático cuando se enfrentan a datos complejos o no estructurados [41].

38. Minería de datos

Proceso de descubrir patrones, correlaciones y estructuras interesantes en grandes conjuntos de datos mediante técnicas de análisis, estadística y aprendizaje automático. Su objetivo es extraer conocimiento útil que apoye la toma de decisiones en diversas áreas [42].

IV. ESTADO DEL ARTE

En esta sección se presentan los trabajos más recientes y los avances realizados en las áreas de Procesamiento del Lenguaje Natural (NLP), análisis acústico y el desarrollo de modelos de inteligencia artificial (IA) entrenados específicamente para la detección temprana de riesgos en salud mental, con énfasis en la identificación de tendencias suicidas. Estas áreas han evolucionado considerablemente gracias al auge de las tecnologías de aprendizaje automático y al acceso a grandes volúmenes de datos provenientes de diversas fuentes, como redes sociales, registros médicos y grabaciones de voz.

1. Panorama General de Procesamiento del Lenguaje Natural y Análisis Multimodal

Los avances recientes en Procesamiento del Lenguaje Natural (NLP) se han visto reforzados por la creciente disponibilidad de datos y la sofisticación de modelos de aprendizaje profundo (deep learning) [43]. En paralelo, el análisis acústico y el uso de señales multimodales (voz, texto, imágenes, etc.) han cobrado relevancia para la identificación de marcadores de depresión y tendencias suicidas [44]. La confluencia de estos campos ha dado lugar a sistemas cada vez más precisos y robustos.

a. Contribución de los Modelos de Lenguaje

Los grandes modelos de lenguaje, basados en arquitecturas tipo Transformer (p. ej., BERT, RoBERTa, GPT-3), han mejorado la capacidad de extraer patrones semánticos y emocionales complejos a partir de texto [25]. Estos modelos, entrenados inicialmente de forma auto-supervisada con grandes corpus, se refinan mediante estrategias de fine-tuning para la detección de tendencias suicidas o estados depresivos en plataformas como Twitter, Reddit u otros foros en línea [45]. Los autores destacan el uso de técnicas como:

i. **Regularización y Fine-Tuning:** Ajuste de hiperparámetros y aplicación de dropout o weight decay para evitar sobreajuste.
ii. **Ensamble de Modelos:** Combinación de diferentes modelos (e.g., BERT + BiLSTM) para reforzar la capacidad de clasificación.

b. Análisis acústico y señales biométricas

La extracción de características acústicas, como MFCC (Mel Frequency Cepstral Coefficients), prosodia o voice pitch, se integra con modelos de aprendizaje profundo para la clasificación de estados emocionales [46]. Algunos enfoques

recientes incorporan además señales fisiológicas (ritmo cardíaco, conductancia de la piel), lo que permite una caracterización más precisa del estado mental del individuo [47]. Sin embargo, la interpretación de estas señales requiere una cuidadosa ingeniería de características y metodologías de fusión (feature-level o decision-level) para combinar efectivamente distintas modalidades.

2. Modelos de IA Aplicados a la Detección de Riesgo Suicida

Los trabajos más recientes en la literatura se dividen en varios enfoques, según la naturaleza de los datos de entrada y el tipo de modelado estadístico:

a. Clasificación en Redes Sociales

Modelos entrenados con publicaciones y comentarios en plataformas como Reddit y Twitter han demostrado alta precisión (superior al 90% en algunos casos) en la identificación de tendencias suicidas o trastornos mentales [45]. Se observan, no obstante, importantes desafíos:

i. **Ética y Privacidad:** El acceso y uso de datos personales en redes sociales conlleva implicaciones éticas y legales, por lo que la anonimización y el consentimiento informado resultan cruciales.
ii. **Representatividad:** Las personas que no utilizan estas plataformas o lo hacen de forma limitada quedan fuera de las muestras, pudiendo introducir sesgos significativos en los modelos [43].

b. Registros Clínicos Electrónicos (EHR)

La minería de datos en historiales médicos para detección de factores de riesgo ha sido explorada por sistemas como VSAIL (Virtual Suicidal Alert Intelligent System), donde se implementan algoritmos predictivos que generan alertas tempranas [48]. Estos métodos suelen combinar técnicas de ensemble learning (Random Forests, Gradient Boosting) y modelos neuronales (RNAs o Transformers), tomando como entrada variables demográficas, antecedentes clínicos y resultados de pruebas psicológicas [49].

A pesar de la alta eficacia en entornos clínicos controlados, la adaptación de estos modelos a diferentes sistemas de salud requiere una integración robusta, así como estrategias de encriptación y anonimización de datos que garanticen la privacidad [50].

i. **Enfoques Multimodales:** En la literatura se destacan modelos que integran señales acústicas, textuales e incluso visuales (expresiones faciales), logrando precisiones de hasta el 98% en la clasificación de intenciones suicidas [46]. Técnicamente, se utilizan técnicas de feature fusion (concatenación a nivel de embeddings) o decision fusion (votación de la predicción final). Estudios recientes demuestran la eficacia de este planteamiento, en el que la salida de modelos como RoBERTa se combina con capas convolucionales sobre espectrogramas para la detección de patrones en la voz [51].

ii. **ChatGPT y Herramientas Basadas en Modelos de Lenguaje:** ChatGPT ha sido propuesto como herramienta de cribado o apoyo psicológico [52]; sin embargo, estudios comparativos evidencian que su capacidad de evaluación del riesgo suicida no cumple con los estándares clínicos de escalas como la Columbia-Suicide Severity Rating Scale (C-SSRS) [53]. Por ello, se enfatiza que la IA no debe substituir la intervención humana directa en escenarios críticos.

3. Plataformas Basadas en IA para el Apoyo en Salud Mental

a. **Youper**: emplea algoritmos de IA para adaptar estrategias de terapia cognitivo-conductual (TCC) y gestionar de forma personalizada el estrés y la ansiedad [54]. Desde el punto de vista técnico, recurre a técnicas de reinforcement learning para ajustar las recomendaciones al progreso del usuario, aunque en casos severos la efectividad puede verse limitada por la interacción activa requerida.

b. Wysa: opera como un chatbot que integra patrones de TCC en un entorno de mensajería anónima [55]. Su diseño gira en torno a una arquitectura de diálogo secuencial, donde la IA propone "rutinas" de afrontamiento basadas en el historial de interacción del usuario. Sin embargo, los escenarios complejos superan las capacidades de su modelo estadístico, lo que remarca la necesidad de intervención profesional.

c. **Replika**: se basa en conversaciones personalizadas para fomentar la reflexión emocional, pero carece de un marco terapéutico formal que posibilite diagnósticos o intervenciones más profundas [55]. La técnica principal de modelado del lenguaje se centra en redes neuronales recurrentes y, más recientemente, en Transformers preentrenados, aunque las respuestas generadas están más orientadas a la empatía que a la clínica.

d. **Woebot**: adopta la TCC como eje principal de intervención, presentando al usuario módulos estructurados de autoayuda [55]. La IA implementa rule-based systems combinados con algoritmos supervisados de intención, lo que resulta

útil para manejo diario de emociones, aunque con interacciones y contenidos predefinidos.

e. **Aimentia Health:** este sistema emplea minería de datos para detectar patrones asociados con trastornos mentales, generando diagnósticos preliminares [56]. La eficacia depende en gran medida de la calidad y diversidad de los datos capturados, exigiendo un riguroso control de sesgos y un seguimiento continuo de la evolución de los usuarios.

4. Relevancia y Limitaciones Críticas

a. **Interpretabilidad**: El empleo de modelos basados en aprendizaje profundo supone un reto de "caja negra". Se están explorando métodos de interpretabilidad como Grad-CAM o SHAP (SHapley Additive exPlanations) para justificar decisiones, lo que reviste especial importancia en contextos clínicos [45].

b. **Calidad de los Datos:** El rendimiento de los modelos está condicionado por la diversidad y representatividad de los conjuntos de datos. La escasez de datos etiquetados sobre suicidio y depresión grave dificulta el entrenamiento y la validación, generando riesgo de sobreajuste [43] [46].

c. **Ética y Privacidad:** Dado que los datos tratados (grabaciones de voz, historiales clínicos, publicaciones en redes sociales) pueden ser altamente sensibles, es vital adoptar medidas sólidas de encriptación, anonimización y cumplimiento de normativas (GDPR, HIPAA), así como garantizar la transparencia en el uso de la información [47] [48].

d. **Generalización y Transferencia:** Los modelos entrenados en un entorno o población específica pueden no generalizar bien a otros, requiriendo técnicas de domain adaptation o aprendizaje por transferencia (transfer learning) para ajustarse a nuevas realidades [49].

5. Perspectivas de Futuro

Las líneas de investigación actuales apuntan a:

a. **Aprendizaje Semi-supervisado y Auto-supervisado:** Reducir la dependencia de datos etiquetados y mejorar la robustez del modelo ante nuevas variables contextuales.

b. **Herramientas de Diagnóstico en Tiempo Real:** Integrar sistemas de IA en aplicaciones móviles y entornos clínicos, permitiendo alertas instantáneas y remisiones más ágiles.

c. **Enfoque Multimodal Extendido:** Incorpora biomarcadores adicionales (EEG, análisis de movimiento, etc.) para lograr sistemas aún más precisos y explicables.

V. SOLUCIÓN PROPUESTA

En la presente sección se describirán las distintas fases para la elaboración del modelo multimodal final (Ilustración 1), desde la construcción de un dataset propio para esta investigación, implementación de cada uno de los modelos de diagnóstico (texto, multimodal (texto+audio), unificado), así como su proceso de entrenamiento y construcción de entorno para usuario final donde se pondrá a prueba dicho modelo final. Esto con el fin verificar la eficiencia de técnicas de NLP y componentes de Machine Learning para la clasificación y diagnóstico de posibles riesgos de suicidio. Para la evaluación, se tomará como medida las métricas de exactitud, la sensibilidad, la precisión, matriz de confusión y la curva ROC-AUC de cada uno de los modelos.

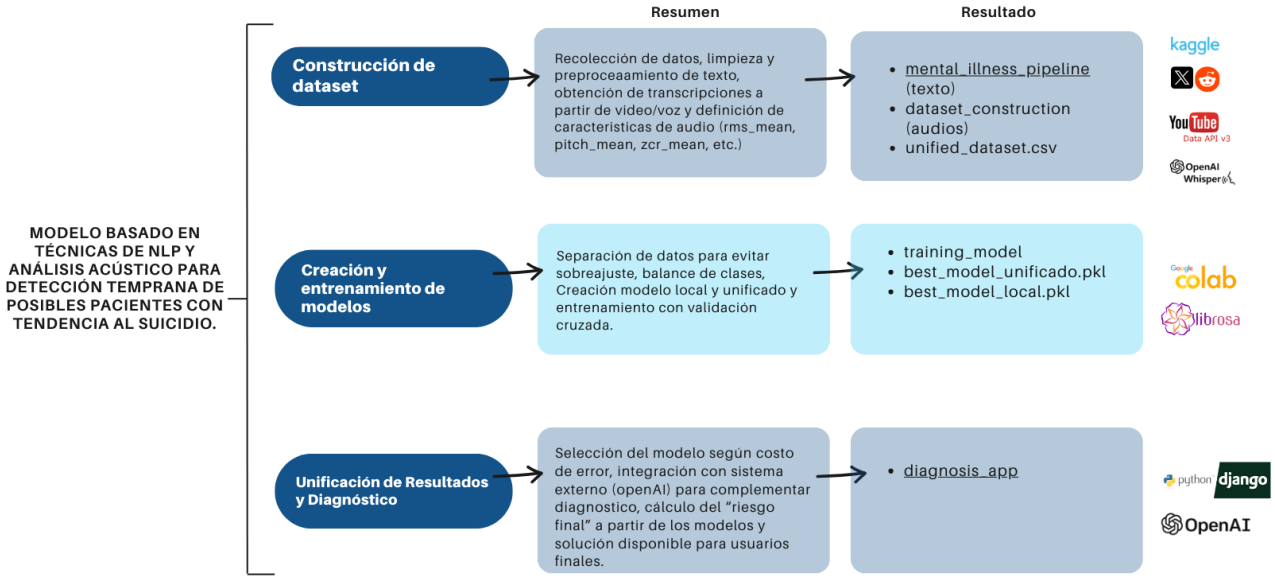

Ilustración 1 Resumen de la solución.

1. Metodología para la Obtención del Dataset

a. Selección y procesamiento de texto

En el proceso de construcción del dataset de texto, se recopiló un conjunto de datos dataset_eng.xlsx que consta de 23,731 mensajes recopilados de redes sociales y repositorios como Kaggle. Estos mensajes abarcan una amplia gama de emociones —positivas, negativas y neutrales—(Tabla 1), capturando interacciones reales en plataformas como Twitter y Reddit. La diversidad y cantidad de datos permiten

entrenar modelos robustos de análisis de sentimientos y detectar patrones emocionales complejos.

Tabla 1 Distribución de sentimientos

Sentimiento	Cantidad	Porcentaje
Negativo	16718	70.45
Positivo	6960	29.33
Neutral	53	0.22
Total	23731	100.00

i. Preprocesamiento con TF-IDF

El preprocesamiento del texto constituyó una etapa crítica en la construcción del dataset, ya que garantiza que los datos sean adecuados para su análisis posterior. En este trabajo, el proceso se alineó con las mejores prácticas descritas en [50], donde se destacaron la importancia de la limpieza y normalización de textos provenientes de redes sociales. Este procedimiento incluyó la eliminación de ruido textual, como URLs, menciones y caracteres especiales, seguido de técnicas de lematización y eliminación de palabras vacías.

Para convertir los datos textuales en representaciones numéricas adecuadas para modelos computacionales, se utilizó la técnica de TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF, ampliamente reconocido en la literatura por su capacidad de ponderar la relevancia de los términos en documentos específicos [51], permitió transformar los mensajes en vectores que capturan características semánticas clave, facilitando el análisis automatizado posterior.

ii. Análisis de Sentimientos y Modelado de Tópicos

La identificación de emociones y temas subyacentes en los textos recopilados se realizó mediante una combinación de técnicas de análisis de sentimientos y modelado de tópicos. Este enfoque sigue las estrategias de modelos avanzados como BERT y VADER.

(1). Análisis de Sentimientos

BERT: Se utilizó para capturar sutilezas contextuales en los textos y proporcionar representaciones semánticas enriquecidas. Este modelo demostró ser

particularmente útil para detectar emociones complejas en mensajes de redes sociales. [25]
VADER: Herramienta eficiente para analizar textos cortos y coloquiales, como tweets y publicaciones en Reddit. VADER complementó las capacidades de BERT al capturar matices como sarcasmo y jerga. [29]
La combinación de ambos modelos en un enfoque de ensamblaje permitió mejorar la precisión del análisis, mitigando las limitaciones individuales de cada técnica [30].

(2). Modelado de Tópicos:

Para descubrir temas recurrentes en los mensajes, se empleó el modelo Latent Dirichlet Allocation (LDA). Este enfoque estadístico, mencionado en [50], facilitó la identificación de tópicos predominantes en los textos negativos, proporcionando información clave sobre patrones temáticos relacionados con las emociones detectadas (Tabla 2).

Tabla 2 Cálculo de consistencia por número de tópicos

Número de tópicos	Consistencia
2	0.455
3	0.438
4	0.453
5	0.428
6	0.457
7	0.457
8	0.456
9	0.431
10	0.441

iii. Visualización y Evaluación de Resultados

La visualización y evaluación de los resultados desempeñaron un papel crucial en este estudio, facilitando una comprensión más profunda de las emociones y temas identificados. Para este proceso, se generaron herramientas visuales como grafos de polaridad emocional y nubes de palabras.

Grafo de Polaridad Emocional: Este grafo (Ilustración 2) permitió observar la interrelación entre los diferentes tópicos y sus cargas emocionales. Los resultados mostraron cómo ciertos temas estaban estrechamente vinculados a emociones

negativas predominantes, destacando patrones críticos que podrían ser indicadores de estados emocionales complejos.

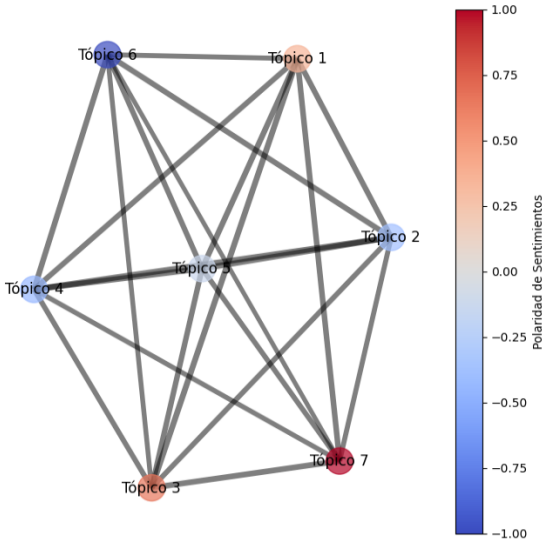

Ilustración 2 Grafo de polaridad de sentimientos

Nubes de Palabras: Estas representaciones gráficas ilustran los términos más relevantes en cada tópico, facilitando la interpretación de los temas subyacentes. Estas visualizaciones ayudan a identificar términos clave que informan sobre las preocupaciones principales expresadas en los textos (Ilustración 3).

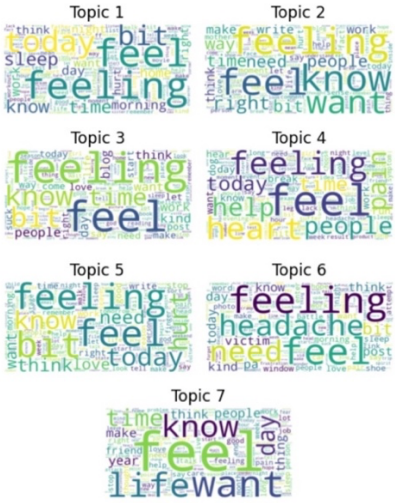

Ilustración 3 Nube de palabras por tópico

Interpretaciones Avanzadas: Se integraron modelos de OpenAI, como GPT-3.5 y GPT-4, para enriquecer el análisis con interpretaciones contextuales más profundas.

Estos modelos proporcionaron perspectivas adicionales al relacionar los tópicos con posibles condiciones emocionales, demostrando la utilidad de combinar técnicas tradicionales con herramientas avanzadas de inteligencia artificial (Ilustración 4).

	Topic	Query	gpt-3.5-turbo-0125	gpt-4
0	Topic 1		Based on the given topical words, it appears t...	1. Anxiety, Panic, Fear, Worry, Stress: These ...
1	Topic 2		Based on the given topical words, it appears t...	1. Anxiety\n2. Depression\n3. Obsessive-Compul...
2	Topic 3		Based on the given topical words, it appears t...	1. Anxiety, Panic, Fear, Worry, Stress: These ...
3	Topic 4		Based on the given topical words, it appears t...	Obsessive-Compulsive Disorder, Anxiety, Depres...
4	Topic 5		Based on the given topical words, it appears t...	Apologies, but you haven't provided any topica...
5	Topic 6		Based on the given topical words, it appears t...	1. Anxiety, Panic, Fear, Worry, Stress: These ...
6	Topic 7		Based on the given topical words, it appears t...	You didn't provide any topical words. Please p...

Ilustración 4 Diagnóstico preliminar de dataset de texto con openAI

La combinación de estas técnicas de análisis y visualización no solo validó la robustez del enfoque metodológico adoptado, sino que también ofreció una base sólida para extender el estudio hacia aplicaciones prácticas en la evaluación de patrones emocionales. Los resultados obtenidos destacaron la utilidad del procesamiento del lenguaje natural en el análisis de datos textuales provenientes de redes sociales que sirvieron de apoyo para la implementación del enfoque multimodal planteado en este trabajo.

b. Selección y procesamiento de Videos/audios

Para conformar el dataset de audios, se seleccionaron alrededor de 500 videos de YouTube y Vimeo utilizando queries específicas y herramientas como scrapping para obtener entradas orientadas a temas relacionados con el suicidio y la salud mental. Posteriormente, los audios de los videos fueron extraídos con sus características acústicas correspondientes, y se clasificaron en dos categorías principales:

Videos con Connotación Negativa: Contenidos que reflejan experiencias personales relacionadas con pensamientos suicidas, desesperanza o depresión. Estos videos frecuentemente contienen testimonios de personas que han convivido con la depresión o manifiestan haber tenido intentos de suicidio.

Videos con Connotación Positiva o Neutra: Testimonios que evocan un mensaje de esperanza, superación, o presentaciones de profesionales de la salud hablando sobre estrategias de prevención del suicidio y factores que conducen a esta problemática.

i. Extracción y procesamiento de audio

Se implementó un pipeline de procesamiento que incluye:

Extracción de Audio: Utilizando herramientas como ffmpeg, se extrajo la pista de audio de cada video descargado, garantizando la calidad y compatibilidad para el procesamiento posterior.

Filtrado de Contenido: Se excluyeron aquellos videos que no contenían discurso humano, asegurando que el dataset final solo contenga materiales con contenido vocal relevante.

ii. Transcripción y Extracción de Características

Una vez extraídos los audios, se empleó el modelo Whisper de OpenAI para obtener las transcripciones de cada archivo de audio. A partir de estas transcripciones y el procesamiento acústico, se calcularon diversas características que son **fundamentales para el análisis** emocional y de riesgo.

Las características extraídas para cada video incluyen:

- **Transcripción:** Texto resultante de la conversión del discurso a texto.
- **RMS Mean:** La media de la raíz cuadrática media, que indica la energía promedio del audio. Es útil para detectar variaciones en la intensidad vocal que pueden correlacionarse con estados emocionales [63].
- **Spectral Centroid Mean:** Centroide espectral promedio, que da una idea de la "brillantez" o frecuencia predominante del audio. Variaciones en este parámetro pueden reflejar tensiones emocionales o cambios de tono [64].
- **Spectral Bandwidth Mean:** Ancho de banda espectral medio, relacionado con la dispersión de frecuencias en la señal, lo que puede indicar cambios en la articulación y emoción del discurso [65].
- **Spectral Rolloff Mean:** Frecuencia de roll-off, que determina el punto donde se encuentra una cierta proporción de la energía total del espectro. Puede correlacionarse con características de la claridad y presencia en el habla [66].
- **Zero-Crossing Rate Mean:** Tasa media de cruces por cero, un indicador del nivel de ruido en la señal y cambios rápidos de fase, relacionados con emociones agudas o estados de agitación [67].
- **Pitch Mean:** Frecuencia fundamental promedio, asociada con el tono de voz; variaciones en el pitch son indicadores directos de estados emocionales como tristeza o ansiedad [68].

- **Spectral Contrast Mean:** Contraste espectral medio, que refleja la diferencia entre picos y valles en el espectro sonoro, relacionado con la percepción y proyección emocional en el discurso [69].
- **Spectral Flatness Mean:** Medida de planicidad espectral, que identifica si una señal es más tonal o ruidosa. Un mayor valor puede indicar falta de tonalidad, asociada a estados depresivos [66].
- **Tempo:** Ritmo o velocidad media del habla, que puede asociarse a estados de agitación o letargo en el discurso.
- **Etiqueta (label):** Clasificación del video como contenido de riesgo o no, basada en las consultas realizadas para su obtención.

iii. Relevancia de las Características Seleccionadas

Las características acústicas seleccionadas son fundamentales para el análisis del estado emocional y psicológico de los individuos en los audios. La RMS, el pitch, y el tempo, entre otros, proveen indicadores cuantitativos que, combinados con la transcripción textual, permiten una comprensión más profunda del discurso emocional. Estudios previos han demostrado que ciertos patrones en estas características pueden correlacionarse fuertemente con estados depresivos y ansiedad [52] [53]. Por ejemplo: Un tono monótono y baja variación en la intensidad suelen asociarse con depresión [54] o por otro lado, cambios en el centroide espectral y el ancho de banda pueden reflejar tensiones y emociones negativas, lo que es relevante para el diagnóstico de riesgo [55] [56].

La construcción de este dataset se realizó bajo consideraciones éticas, asegurando el respeto a la privacidad y la confidencialidad de las personas entrevistadas. Los videos son de acceso público en YouTube y se utilizaron con fines educativos e investigativos, alineándose con la doctrina del "uso justo" para investigación.

El dataset de audio resultante, provee una base sólida para el entrenamiento de modelos de aprendizaje automático destinados a la predicción de estados emocionales y riesgos asociados al suicidio. Las características seleccionadas facilitan un análisis detallado que, apoyado en técnicas de procesamiento de señales y NLP, promete contribuir significativamente a la identificación temprana de señales de riesgo y a la mejora de intervenciones en salud mental.

c. Dataset Unificado

Como dataset unificado final para el análisis de riesgo en salud mental, objetivo de esta monografía, se recopila la data de los datasets anteriores, por consiguiente,

consta de información tanto textual como acústica para el entrenamiento de modelos de diagnóstico. Concretamente, el conjunto incluye las características descritas en (Ilustración 5):

```
Columnas del DataFrame:

['transcription',
 'rms_mean',
 'spectral_centroid_mean',
 'spectral_bandwidth_mean',
 'rolloff_mean',
 'zcr_mean',
 'pitch_mean',
 'spectral_contrast_mean',
 'flatness_mean',
 'tempo',
 'label']
Forma del DataFrame: (23852, 11)
```

Ilustración 5 Características de dataset

La **forma** del dataset, (23852, 11), indica que se dispone de 23 852 muestras totales, con 11 variables asociadas. De estas, 10 se emplean como variables predictoras y una se utiliza como objetivo de clasificación, columna **label** de la clase objetivo-binaria que indica si la muestra presenta rasgos de ideación suicida (1) o no (0). Dentro del mismo se encuentran contenidos **161 ítems de audio,** cada uno con su respectiva transcripción y un conjunto de características acústicas extraídas, tales como RMS, espectro, tonalidad y tempo, que aportan información relevante sobre las emociones y estados psicológicos del hablante.

2. Enfoque unificado para diagnóstico y análisis: Creación de Modelos

La solución implementada para el análisis y diagnóstico de emociones en textos provenientes de redes sociales, integra un enfoque unificado que combina tres fuentes principales de predicción: un **Modelo Local**, un **Modelo Unificado**, y un **Modelo Complementario de ChatGPT**. Este diseño asegura una evaluación robusta y balanceada, aprovechando las capacidades estadísticas tradicionales y la comprensión avanzada proporcionada por modelos generativos, con el objetivo de detectar tempranamente posibles tendencias suicidas en pacientes.

La solución no se limita únicamente al análisis de texto escrito, sino que también incorpora características multimodales al integrar información textual (como transcripciones de audio o mensajes escritos) y características acústicas (como rasgos del audio), además de un análisis detallado de sentimientos. Esto permite abordar los casos con una perspectiva más completa y precisa. Una vez validado y refinado

el dataset inicial, se diseñaron dos modelos clave que, dada una entrada de texto o audio, retornan un diagnóstico binario: **0** en caso de no detectar indicios de riesgo suicida y **1** en caso de identificar un riesgo probable.

Este enfoque integrado garantiza no solo la detección temprana de patrones preocupantes en los datos, sino también la robustez del sistema al combinar técnicas de procesamiento de lenguaje natural y análisis de audio con herramientas avanzadas de diagnóstico generativo. Al incorporar diversas modalidades de entrada y técnicas de evaluación complementarias, se optimiza la capacidad para identificar señales de alerta en contextos emocionales complejos, contribuyendo así al monitoreo y prevención en escenarios clínicos y sociales.

a. Modelo local

El Modelo Local se basa en técnicas clásicas procesamiento de tecto y lenguaje natural (NLP), incluyendo el uso de representaciones vectoriales **como TF-IDF** y algoritmos estadísticos como **Logistic Regression,** para la clasificación de emociones. Este modelo, si bien es limitado en su capacidad para captar sutilezas contextuales complejas, proporciona una base sólida para identificar patrones emocionales recurrentes en datos estructurados. Además, su desempeño es altamente interpretable, lo que facilita la validación de resultados en etapas iniciales del diagnóstico.

El **Modelo Local** se construye mediante un *pipeline* que aplica TfidfVectorizer sobre la columna de texto transcription, seguido de LogisticRegression. La **representación TF-IDF** (Term Frequency – Inverse Document Frequency) [51] mide la relevancia de cada término en un documento y se define de forma simplificada como:

$$tfidf(t,d) = tf(t,d) \times log\left(\frac{N}{df(t)}\right)$$

donde:

- $tf(t,d)$ es la frecuencia del término t en el documento d.
- $df(t)$ es el número de documentos donde aparece t.
- N es el total de documentos de la colección.

Con GridSearchCV se optimizan hiperparámetros como ngram_range [35] en $tfidf$ y C en LogisticRegression [40], maximizando la métrica f1_score. Este proceso está soportado por **scikit-learn** [57], que permite la búsqueda exhaustiva y una validación cruzada.

Para el paso de la **Regresión Logística** [58], se realiza la clasificación binaria a partir de la función sigmoide:

$$\hat{y} = \sigma(w^T x + b), donde\ \sigma(z) = \frac{1}{1 + e^{-z}}$$

Donde,

- x es el vector TF-IDF resultante.
- w y b son los parámetros ajustados durante el entrenamiento.
- $\hat{y}$ oscila entre 0 y 1, lo que facilita la predicción de "riesgo" (1) o "no riesgo" (0).

b. Modelo Unificado

El Modelo Unificado, por otro lado, busca no solo considerar como variables de estudio las entradas de tipo textual sino también involucrar a las entradas acústicas como parte de su entrenamiento, por consiguiente, combina múltiples enfoques basados en embeddings avanzados y técnicas de ensamblaje, como la integración de modelos como **BERT** y **VADER**. En donde mientras que BERT aporta un análisis contextual profundo mediante su arquitectura bidireccional, VADER complementa con una detección rápida y eficiente de emociones explícitas en textos cortos y coloquiales. Este modelo actúa como un puente entre técnicas estadísticas tradicionales y capacidades modernas de procesamiento del lenguaje natural, proporcionando una evaluación balanceada y eficiente. Para su implementación, en este se combina:

1. Embeddings semánticos de la transcripción (mediante Sentence Transformers [28]).
2. Variables numéricas derivadas del audio (energía RMS, centroides espectrales, pitch, tempo, etc.).
3. Polaridad de sentimiento, estimada con un pipeline de DistilBERT [59].

Esta integración se realiza a través de FeatureUnion, uniendo dos *pipelines* parciales (texto y numérico) y culminando en un RandomForestClassifier.

Se emplea librosa [21] para calcular indicadores como:

- rms_mean (Root Mean Square)
- spectral_centroid_mean (Centroide espectral)
- spectral_bandwidth_mean

- rolloff_mean
- zcr_mean (Tasa de Cruces por Cero)
- pitch_mean (Frecuencia fundamental estimada)
- spectral_contrast_mean
- flatness_mean
- tempo

Estos rasgos ayudan a modelar patrones de voz potencialmente asociados a estados emocionales (por ejemplo, tristeza, estrés, etc.).

Para analizar la polaridad de sentimiento en textos en inglés (idioma escogido para el dataset), se define una función llamada compute_sentiment_polarity(). Esta función utiliza un pipeline de análisis de sentimiento, conjunto ordenado de pasos automatizados diseñados para procesar texto y determinar si la opinión expresada es positiva o negativa. El resultado inicial del pipeline es una etiqueta como "POSITIVE" o "NEGATIVE". Esta etiqueta se convierte luego en un valor numérico que oscila entre −1 y 1, donde -1 representa un sentimiento completamente negativo, 1 un sentimiento completamente positivo, y 0 una neutralidad.

Este análisis de sentimiento se basa en técnicas avanzadas de Transfer Learning y Transformers. El Transfer Learning permite reutilizar conocimientos adquiridos por modelos de lenguaje previamente entrenados en tareas similares, adaptándolos a nuestro problema sin tener que empezar desde cero. Los Transformers por su parte, son una arquitectura de redes neuronales que han revolucionado el procesamiento del lenguaje natural, permitiendo que el modelo entienda contextos complejos y relaciones a lo largo de un texto mediante mecanismos de atención [24].

Una vez que se obtiene el vector numérico que representa la polaridad del sentimiento, este se utiliza como parte de un conjunto de características extendido para entrenar un modelo de clasificación. En este caso, se emplea RandomForestClassifier, un algoritmo de aprendizaje automático que combina múltiples árboles de decisión. Cada árbol en el bosque realiza una predicción, y la decisión final se toma en función del consenso entre ellos. Este enfoque ayuda a reducir el sobreajuste (cuando un modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad de generalización) y mejora la precisión de las predicciones [36].

$$RF(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x),$$

donde h_m es la predicción de cada árbol, y se hace un voto mayoritario o promedio para la clasificación.

El entrenamiento del *Modelo Unificado* tiende a llevar más tiempo que el *Modelo Local*, especialmente por la generación de **embeddings** en GPU y el procesamiento de audio. Sin embargo, su capacidad de **abarcar múltiples modalidades** (texto, sonido y sentimiento) lo hace **significativamente más robusto** en dominios sensibles, como la detección de riesgo suicida [57] [60].

c. Modelo complementario (ChatGPT)

Con el fin de ofrecer un diagnóstico lo más aterrizado a la realidad y que genere mayor exactitud a la hora de retornar una respuesta o posible diagnóstico, se decide añadir la incorporación de ChatGPT como componente complementario de la solución respondiendo a su capacidad única para ofrecer una comprensión semántica avanzada, lo que resulta especialmente valioso en el análisis de textos ambiguos o complejos. Las razones para considerarlo como principal incluyen:

- **Desempeño humano-simil:** ChatGPT ha demostrado capacidades comparables a evaluadores humanos en tareas de diagnóstico emocional, especialmente en escenarios que requieren interpretaciones contextuales profundas [61].
- **Captura de matices lingüísticos:** Su entrenamiento masivo en datos de lenguaje natural le permite identificar matices contextuales y relaciones implícitas entre palabras y frases, incluso cuando los mensajes analizados son ambiguos o carecen de estructura formal.
- **Flexibilidad y adaptabilidad:** A diferencia de los modelos tradicionales, ChatGPT puede interpretar y contextualizar información más allá de patrones predefinidos, generando respuestas que integran tanto conocimiento general como específico.

3. Entrenamiento y Validación

a. División de datos

Para evitar *overfitting*, se separa el dataset unificado en:

- Entrenamiento (≈ 70%): Aproximadamente 16,867 ítems, usados para que los modelos aprendan patrones y relaciones inherentes a los datos.

- Validación (≈ 15%): Alrededor de 3,615 ítems, destinados a afinar hiperparámetros, realizar selección de modelo y prevenir sobreajuste, sin comprometer la evaluación final.

- Prueba (≈ 15%): Cerca de 3,614 ítems, reservados para evaluar el desempeño final de los modelos en datos no vistos, proporcionando una estimación realista de su capacidad de generalización.

La distribución de los tres datasets se muestra en la Ilustración 6. Esta metodología se justifica para afinar hiperparámetros en la fase de validación y garantizar una evaluación objetiva en la fase de prueba [57].

```
=== Partición de datos ===
Train shape: (16695, 10)
Val   shape: (3579, 10)
Test  shape: (3578, 10)
```

Ilustración 6 Partición de datos para entrenamiento

b. Consideraciones en la Partición de Datos

Dado que el dataset unificado, incluye dos tipos principales de datos (texto y audio), se emplearon técnicas de estratificación para asegurar que los conjuntos de validación y prueba contengan una representación proporcional de ambos tipos. Esta estrategia ayuda a preservar la distribución original de clases y formatos en cada subconjunto, garantizando que los modelos se evalúen de manera justa y consistente [62]. Adicionalmente, ante la posible existencia de desequilibrios entre clases (por ejemplo, un mayor número de casos "no riesgo" frente a "riesgo"), se aplicaron métodos de muestreo y ajuste de pesos durante el entrenamiento. Estas técnicas son fundamentales para evitar sesgos en la predicción y asegurar que los modelos aprendan a identificar correctamente ambas clases [63], por último, para el modelo unificado, que integra tanto datos de texto como de audio, se prestó especial atención a la consistencia en la extracción de características acústicas a través de todos los subconjuntos (entrenamiento, validación y prueba). Mantener esta consistencia es crucial para la robustez del modelo, ya que asegura que las representaciones acústicas sean comparables y fiables en todas las fases del entrenamiento y evaluación [64].

c. Entrenamiento de modelos

El modelo local en esta configuración utiliza el método GridSearchCV para optimizar los parámetros a través de validación cruzada, utilizando tres folds para

garantizar una evaluación robusta. Se exploran varios n-grams, específicamente (1,1) y (1,2), y se prueban distintos valores de la constante de regularización C, que incluyen 0.1, 1.0 y 10, permitiendo así ajustar el modelo de manera precisa según las características del dataset. En contraste, el modelo unificado adopta un enfoque más simplificado, entrenándose directamente con el método pipeline.fit(). Esta técnica omite inicialmente la utilización de Grid Search, priorizando la simplicidad y la eficiencia operativa, aunque se mantiene la opción de expandir este método a futuras iteraciones para realizar búsquedas más exhaustivas de parámetros. Complementariamente, se integra ChatGPT como herramienta de análisis avanzado dentro de la solución, donde se aprovechan sus capacidades sin adentrarse en los detalles técnicos de su estructura o entrenamiento. Tratando a ChatGPT como una caja negra, se beneficia de su poderosa capacidad generativa y su sofisticada comprensión lingüística, facilitando así el análisis sin requerir un entendimiento exhaustivo del modelo. Este enfoque pragmático permite utilizar ChatGPT de manera efectiva dentro de la infraestructura existente, maximizando los recursos y optimizando los resultados del análisis.

4. Integración Final y "Ensemble"

Como parte final de la solución se realiza una integración final o ensemble entre los tres modelos. La asignación de pesos en el diagnóstico unificado responde a un equilibrio cuidadosamente diseñado entre las fortalezas y limitaciones de cada modelo utilizado. Este enfoque asegura que cada componente contribuya de manera efectiva al análisis global, aprovechando sus capacidades específicas para maximizar la precisión y robustez del sistema.

```
=== Comparación Modelos (Validación) ===
+----+------------------+------------+-------------+----------+------------+----------+
|    | model_name       |   accuracy |   precision |   recall |   f1_score |      auc |
|----+------------------+------------+-------------+----------+------------+----------|
|  0 | Modelo Local     |   0.857502 |    0.881459 | 0.921366 |   0.900971 | 0.911049 |
|  1 | Modelo Unificado |   0.808606 |    0.801183 | 0.968229 |   0.876821 | 0.843681 |
|  2 | Modelo ChatGPT   |   0.48     |    0.833333 | 0.294118 |   0.434783 | 0        |
+----+------------------+------------+-------------+----------+------------+----------+
```

Ilustración 7 Comparación de los tres modelos revisados

Los **Modelos Local y Unificado**, con un peso de **0.45, 0.35 respectivamente**, se destacan por su eficacia en la detección de patrones estadísticos y en la captura de relaciones contextuales básicas. Estas técnicas proporcionan una base sólida para el análisis inicial, especialmente en casos donde los datos son estructurados y las emociones explícitas. Sin embargo, su cobertura semántica es limitada en comparación con modelos más avanzados, como ChatGPT. Si bien ofrecen

predicciones precisas, tienden a carecer de la profundidad necesaria para interpretar mensajes ambiguos o contextualmente complejos [65].

Por otro lado, y haciendo referencia a la (Ilustración 7) **ChatGPT (4.0)**, al que se le asigna un peso de **0.20**, demuestra una alta precisión (83.33%) en la tabla comparativa, lo que sugiere que es confiable al identificar riesgos positivos. Sin embargo, su bajo recall (29.41%) indica que omite una cantidad significativa de casos relevantes, lo que puede deberse a su dependencia de un umbral fijo para las decisiones y a su falta de ajuste específico al dataset. Aunque ChatGPT destaca por su capacidad para integrar matices lingüísticos y capturar significados contextuales en texto ambiguo, su desempeño limitado en este caso podría atribuirse a dos factores principales:

- **Falta de ajuste al contexto específico:** A diferencia de los otros modelos entrenados directamente con el dataset, ChatGPT no tiene entrenamiento específico para los datos utilizados, lo que puede afectar su capacidad para identificar patrones específicos del dominio.
- **Dependencia de un umbral fijo:** La evaluación de ChatGPT se basa en convertir su salida (un puntaje de riesgo) en una predicción binaria mediante un umbral (0.6). Esto simplifica su desempeño, pero restringe su capacidad para ajustarse dinámicamente a los datos.
- **Evaluación en un subconjunto reducido del dataset:** La evaluación de ChatGPT se realizó sobre una muestra limitada del conjunto de validación (50 ejemplos), para evitar excesivas llamadas al ser de uso pago, lo que puede no reflejar toda la diversidad de los datos. Un subconjunto pequeño podría no capturar adecuadamente la distribución de clases o la complejidad de los patrones presentes en el dataset completo, afectando negativamente métricas como el recall.

Esta estrategia de ponderación busca equilibrar las capacidades complementarias de los tres modelos. Los Modelos Local y Unificado aportan predicciones estructuradas y consistentes, mientras que ChatGPT contribuye con una perspectiva más rica en contexto y matices lingüísticos. En conjunto, estos pesos asignados aseguran un diagnóstico robusto, adaptativo y capaz de manejar una amplia variedad de entradas textuales, incluyendo aquellas con desafíos lingüísticos significativos. La combinación estratégica permite aprovechar la precisión estructurada de los modelos entrenados directamente y la riqueza semántica inherente a ChatGPT.

Cálculo del Promedio Ponderado

La probabilidad final (p_{final}) es:

$$(p_{final}) = (p_{local} \times 0.45) + (p_{unificado} \times 0.35) + (p_{chatgpt} \times 0.20)$$

Cada probabilidad ($p_{local}, p_{unificado}, p_{chatgpt}$) oscila entre 0 y 1.

En esta ocasión, se consideró un umbral >= 0.6 para etiquetar como "RIESGO", aunque puede cambiarse según las necesidades de **sensibilidad/especificidad** [66].

VI. EVALUACIÓN Y ANÁLISIS DE RESULTADOS

Cada modelo construido fue evaluado mediante **métricas estándar** de la literatura en clasificación binaria.

- Accuracy
- *Precision*
- *Recall (Sensibilidad)*
- *F1-score*
- *AUC (Área bajo la curva ROC)*
- *MCC (Matthews Correlation Coefficient)*
- *Balanced Accuracy* y otras.

Para una visión más detallada a nivel de clase (0 = "no suicida" y 1 = "tendencia suicida"), se generaron **reportes de clasificación**. Asimismo, se incluyeron **matrices de confusión** y **curvas ROC** que permiten interpretar el comportamiento de cada modelo en diferentes umbrales de decisión [18].

En la Ilustración 8, se evidencian gráficamente las diferencias en las métricas por cada modelo.

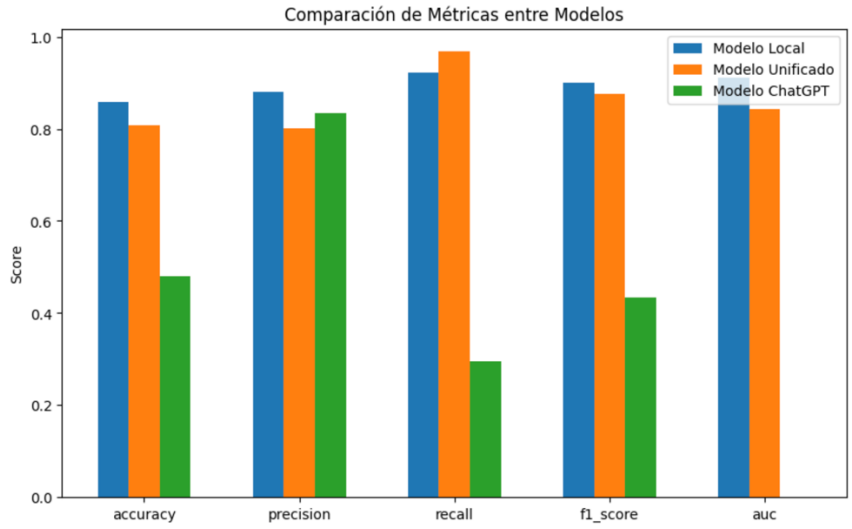

Ilustración 8 Diagrama de barras métricas vs score por modelo

1. Evaluación y resultados por modelo

A continuación, se expone la comparación detallada (Tabla 3) de los resultados principales de los modelos local y unificado, segmentando por métrica y por clase cuando sea necesario.

Tabla 3 Comparación de los principales indicadores de desempeño de los modelos Local y Unificado.

Modelo	Accuracy	Precision	Recall	F1-score	AUC
Modelo Local	0.86	0.88	0.92	0.90	0.91
Modelo Unificado	0.81	0.80	0.97	0.88	0.84

2. Análisis por métrica

a. Accuracy

- El **Modelo Local** obtiene un 0.86 frente a 0.81 del **Modelo Unificado**. Esto sugiere que, si consideramos todas las instancias (positivas y negativas) en conjunto, el Modelo Local clasifica correctamente a un mayor porcentaje de ellas.
- Pese a que la diferencia de 5 puntos porcentuales pudiera parecer moderada, en un escenario de gran volumen de datos (más de 23 000 muestras), este margen representa un número significativo de casos correctamente predichos.

b. Precision

- El Modelo Local exhibe una *precision* de 0.88, mientras que el Unificado se queda en 0.80. La **precision** es crucial cuando el coste de los falsos positivos es alto (por ejemplo, etiquetar como "riesgo suicida" a un paciente que no lo es puede causar alarmas innecesarias, uso extra de recursos clínicos, etc.).
- La diferencia de 0.08 (8 puntos porcentuales) indica que el Modelo Unificado tiende a **elevar su tasa de falsos positivos** para poder ganar sensibilidad en la clase 1.

c. Recall

- Aquí sobresale el **Modelo Unificado**, con un *recall* de 0.97 frente a 0.92 del Modelo Local. Si bien el Modelo Local no es "malo" (92 % de los positivos reales detectados), el Unificado logra capturar casi todos los casos de la clase 1, sacrificando sin embargo su desempeño en la clase 0.
- En el **contexto clínico** de un riesgo suicida, este valor tan alto de *recall* puede ser sumamente relevante, ya que **minimiza los falsos negativos** (p. ej., la posibilidad de dejar pasar un paciente en alto riesgo).

d. F1-score

- El **F1-score** conjuga la información de *precision* y *recall* en un único valor. El Modelo Local (0.90) alcanza un F1 superior al del Unificado (0.88), evidenciando un mejor **equilibrio** entre ambas métricas.
- Esta métrica cobra sentido especial cuando la distribución de clases está desbalanceada, pues un modelo podría tener alta *accuracy* ignorando casi por completo la clase minoritaria [67]. El F1 "penaliza" fuertemente el desbalance de *precision* y *recall*, por lo que el Modelo Local mantiene un punto a favor en la compensación de errores.

e. AUC

- El **Área bajo la curva ROC** mide la probabilidad de que, al tomar una muestra de la clase 1 y una de la clase 0, el modelo puntúe más alto la de la clase 1 que la de la 0 [18]. El Modelo Local consigue un AUC de 0.91, superando al Unificado (0.84).
- Un AUC cercano a 1.0 indica mayor capacidad de discriminación, a lo largo de diversos umbrales de decisión. Así, el Modelo Local, en promedio, separa mejor ambas clases a lo largo de todo el rango de clasificaciones.

3. Análisis por Clase

Los reportes de clasificación detallan rendimiento para cada clase:

- Clase 0: Paciente sin tendencia suicida.
- Clase 1: Paciente con rasgos o signos de suicidio.

A continuación (Ilustración 8), se profundiza en las cifras usualmente reportadas (precision, recall, F1-score) para cada clase.

a. Modelo Local

```
=== Reporte de Clasificación: Modelo Local ===
              precision    recall  f1-score   support

           0       0.79      0.71      0.75      1061
           1       0.88      0.92      0.90      2518

    accuracy                           0.86      3579
   macro avg       0.84      0.81      0.82      3579
weighted avg       0.85      0.86      0.86      3579
```

Ilustración 9 Reporte de clasificación: modelo local

De acuerdo con el reporte de clasificación del modelo local (Ilustración 8), se concluye:

i. Clase 0

- Precision (0.79): De todos los pacientes clasificados como "no suicidas", un 79% realmente son negativos.
- Recall (0.71): El modelo reconoce un 71% de los casos negativos reales, dejando un 29% de falsos positivos.
- F1-score (0.75): Balance entre la capacidad de no sobredimensionar la clase 0 y la de detectarla adecuadamente.

ii. Clase 1

- Precision (0.88): La proporción de predicciones de "riesgo suicida" que realmente son positivas es del 88%.
- Recall (0.92): Captura al 92% de los casos que efectivamente muestran rasgos de suicidio.
- F1-score (0.90): Indica un desempeño robusto en la clase prioritaria (evitando falsos negativos y falsos positivos de manera equilibrada).

En conjunto, la clase 1 está mejor atendida que la clase 0. No obstante, la brecha no es tan acentuada, lo cual sugiere buena estabilidad.

b. Modelo Unificado

```
=== Reporte de Clasificación: Modelo Unificado ===
              precision    recall  f1-score   support

           0       0.85      0.43      0.57      1061
           1       0.80      0.97      0.88      2518

    accuracy                           0.81      3579
   macro avg       0.82      0.70      0.72      3579
weighted avg       0.82      0.81      0.79      3579
```

Ilustración 10 Reporte de clasificación: modelo unificado

De acuerdo con el reporte de clasificación del modelo local (Ilustración 9), se concluye:

i. Clase 0

- Precision (0.85): Cuando el modelo predice "no suicida", acierta un 85% de las veces.
- Recall (0.43): Solamente reconoce un 43% de los negativos reales, generando un 57% de falsos positivos. Este es un punto crítico: más de la mitad de los verdaderos negativos son etiquetados como positivos erróneamente.
- F1-score (0.57): Relativamente bajo debido a esa brecha entre precision y recall.

ii. Clase 1

- Precision (0.80): El 80% de las instancias clasificadas como "en riesgo" lo están de manera correcta.
- Recall (0.97): Apenas un 3% de los casos con rasgos suicidas se quedan sin detectar; esto refleja la filosofía detrás del Modelo Unificado: se prioriza capturar todos los positivos a costa de incrementar falsos positivos.
- F1-score (0.88): Alto, evidenciando una solidez en la detección de la clase 1, a pesar de su precisión un tanto reducida.

Estos números sustentan la idea de que el Modelo Unificado es **más agresivo** en la detección de la clase suicida, sacrificando la correcta clasificación de la clase negativa.

4. Matrices de Confusión

Las matrices de confusión complementan los reportes de clasificación al mostrar exactamente cuántas observaciones caen en cada categoría de (predicción vs. realidad):

a. Modelo local

Tabla 4 Resultados matriz de confusión: modelo local

	Predicción=0	**Predicción=1**
Real=0 (TN)	749	312 (FP)
Real=1 (FN)	198	2320 (TP)

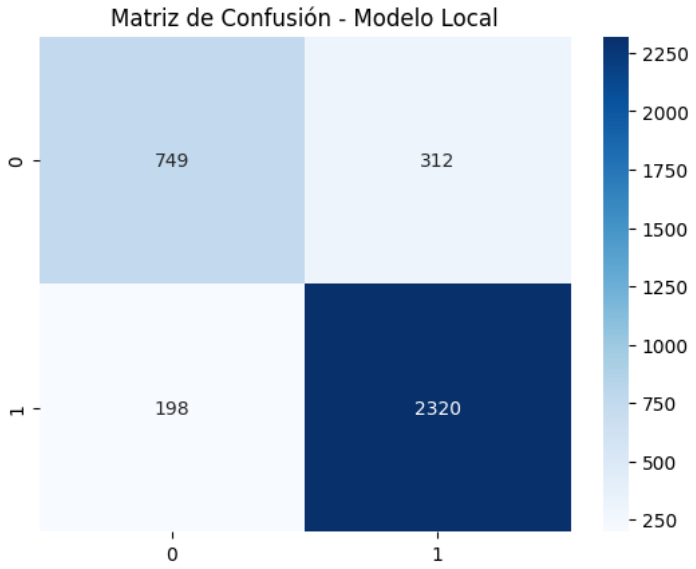

Ilustración 11 Matriz de confusión modelo local

Los resultados de la evaluación de la matriz de confusión para el modelo local (Tabla 4, Ilustración 10), indican:

Falsos Positivos (FP) = 312.

Falsos Negativos (FN) = 198.

A pesar de tener 312 FP y 198 FN, el modelo logra un equilibrio razonable: un número considerable de verdaderos positivos (2320) y verdaderos negativos (749).

El recuento indica que el Modelo Local tiende a "errarle" más en la clase 0 (312 FP) que en la clase 1 (198 FN), aunque las cifras no son extremadamente elevadas.

b. Modelo unificado

Los resultados de la evaluación de la matriz de confusión para el modelo unificado (Tabla 4, Ilustración 10), indican:

Tabla 5 Resultados matriz de confusión modelo unificado

	Predicción=0	**Predicción=1**
Real=0 (TN)	455	606 (FP)
Real=1 (FN)	81	2437 (TP)

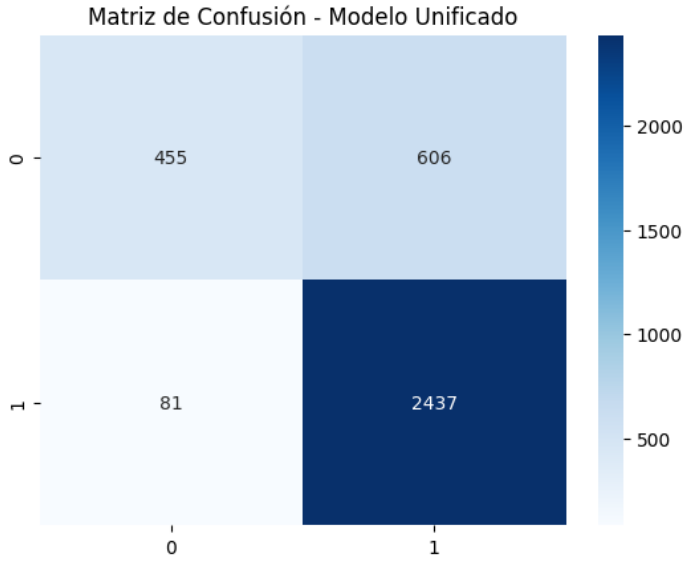

Ilustración 12 Matriz de confusión modelo unificado

Falsos Positivos (FP) = 606.

Falsos Negativos (FN) = 81.

Aquí observamos claramente el énfasis del modelo: **minimizar FN**, incluso si eso ocasiona casi el doble de FP que el Modelo Local (606 vs. 312). En aplicaciones clínicas, esto se traduciría en un alto porcentaje de personas clasificadas como "en riesgo" siendo realmente "no suicidas", lo cual puede sobrecargar los recursos si la confirmación diagnóstica requiere un examen más profundo por parte de un profesional.

5. Curvas ROC y AUC

Las **curvas ROC** grafican la sensibilidad (TPR) frente a la razón de falsos positivos (FPR) a lo largo de distintos umbrales de decisión [18] [19]. El **AUC** resume esta curva en un valor entre 0 y 1. En (Ilustración 12) y (Ilustración 13) se detallan los resultados por modelo evaluado.

a. Modelo Local

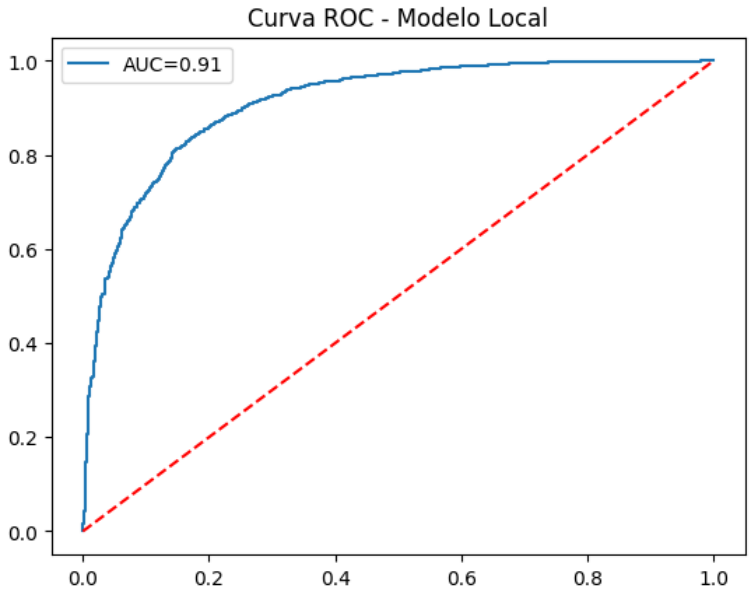

Ilustración 13 Curva ROC- AUC modelo local

- AUC = 0.91: Indica que, en promedio, el modelo discrimina muy bien entre clases, evidenciando un alto grado de separación de las distribuciones de puntajes de la clase 0 y 1.
- La curva ROC se sitúa cerca del eje vertical y luego se curva hacia la esquina superior izquierda, mostrando un buen balance entre TPR y FPR para múltiples cortes.

b. Modelo Unificado

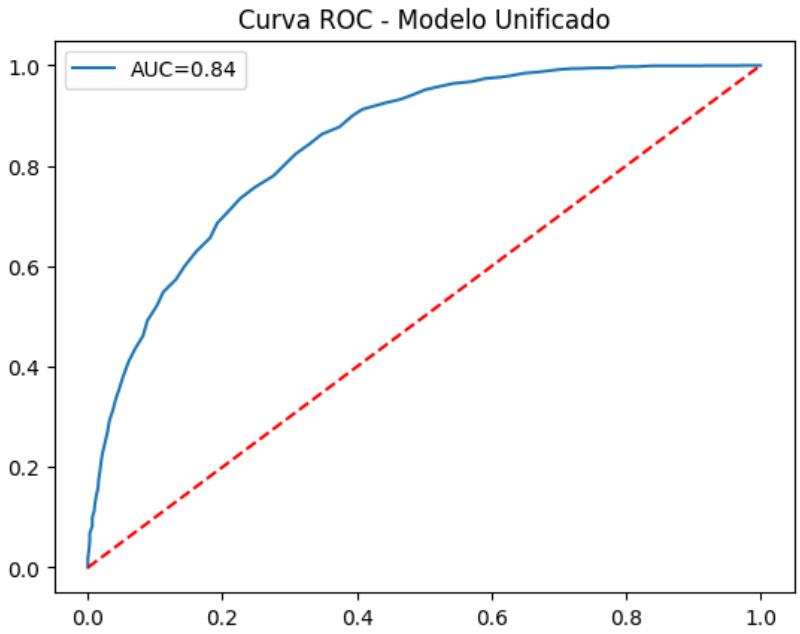

Ilustración 14 Curva ROC-AUC modelo unificado

- AUC = 0.84: Menor que la del Modelo Local, aunque aceptable en términos generales.
- Su curva ROC está fuertemente inclinada en la parte superior (maximizando TPR), pero no logra tan buena especificidad, lo que baja su área global.

6. Ventajas y Limitaciones de modelos construidos

Dado que el dataset se compone de muestras de voz y sus respectivas transcripciones, la multimodalidad constituye una ventaja al combinar señas acústicas (ej: energía, pitch, tempo) con rasgos textuales (a partir de la columna transcription). Sin embargo, cabe destacar otras significativas:

- Variedad de instancias: Permite a los algoritmos generar patrones más estables y robustos estadísticamente.
- Información multimodal: Combina transcripciones con características acústicas. En problemas de salud mental, los indicadores de voz pueden ayudar a detectar señales de angustia, cambios de entonación o vacilaciones típicas de un estado depresivo [68].

No obstante, existen potenciales limitaciones:

- **Distribución de clases**: Es habitual que la cantidad de casos "positivos" (suicidas) sea menor frente a "negativos" (no suicidas). Este posible desbalance puede complicar el aprendizaje de los modelos y requerir técnicas de re-muestreo o ajuste de umbral.
- **Calidad de la anotación**: El etiquetado "label" depende del criterio clínico y/o autoinforme del paciente. Pueden existir falsos positivos en la base de datos o sesgos relacionados con el contexto de recolección (edad, género, condiciones médicas asociadas).
- **Heterogeneidad lingüística**: Si las transcripciones provienen de diversos hablantes con dialectos o idiomas distintos, surgen retos de normalización textual y adaptación acústica.
- **Aspectos éticos**: El manejo de información sensible de salud mental requiere protocolos estrictos de confidencialidad, anonimización y cumplimiento normativo.

En definitiva, **la selección del modelo** dependerá fuertemente de la **política de riesgo** que se desee adoptar en el entorno hospitalario o clínico de aplicación. En un contexto de salud mental, por lo general, se prefiere **maximizar la detección de individuos en riesgo** (clase 1), pero es vital medir las consecuencias éticas, clínicas y económicas de un alto volumen de falsos positivos.

VII. HERRAMIENTA DE DIAGNÓSTICO

Como último paso de la solución, se describe la herramienta que unifica los resultados e integra los distintos modelos (Local, Unificado y ChatGPT) para ofrecer un **diagnóstico ponderado** de riesgo suicida a partir de entradas de texto o audio transcrito (Ilustración 14). Esta solución se enmarca en un **enfoque multimodal**, combinando aspectos clásicos de análisis de texto (TF-IDF y Regresión Logística) con rasgos acústicos y modelos de lenguaje avanzado (p. ej., BERT, VADER y ChatGPT).

La implementación se desarrolla bajo **Django**, lo que permite la interacción con el usuario a través de formularios y vistas web, y la orquestación de la lógica de predicción en segundo plano.

La solución propuesta consiste en una herramienta que, a partir de entradas de texto o audio, integra tanto el uso de modelos de aprendizaje automático tradicionales, como un componente de inteligencia generativa avanzada, con el fin de identificar tempranamente el riesgo de suicidio en pacientes. Para el análisis textual (Ilustración15), se recurre a un modelo local basado en representaciones TF-IDF combinadas con Regresión Logística (best_model_local.pkl resultante del paso anterior), cuyo enfoque interpretativo facilita la detección de patrones léxicos relacionados con la ideación suicida, mientras que la transcripción de audio se procesa mediante una arquitectura de reconocimiento automático del habla que produce texto sobre el cual se aplican las técnicas anteriores. Simultáneamente, se incorpora un modelo unificado (best_model_unificado.pkl resultante del paso anterior) que fusiona rasgos acústicos (por ejemplo, energía de la señal, frecuencia fundamental, tasa de cruces por cero) y métodos de ensamblaje con algoritmos como el Random Forest, permitiendo detectar aspectos paralingüísticos y emocionales en la voz (Ilustración 17). Posteriormente, la solución añade un módulo de ChatGPT, que actúa como componente complementario para la evaluación semántica y contextual de los textos transcritos o introducidos por el usuario, asignándole una ponderación mayor en el cálculo de la probabilidad de riesgo. Esta votación ponderada, que combina las salidas del modelo local, el modelo unificado y ChatGPT, culmina en un diagnóstico binario (“riesgo” o “no riesgo”) y una probabilidad promedio de ideación suicida, ofreciendo así un enfoque multimodal robusto que integra el análisis estadístico clásico, la extracción de rasgos prosódicos y la comprensión profunda del lenguaje natural a fin de maximizar la sensibilidad y la precisión en un escenario clínico donde la detección oportuna de señales de alerta resulta esencial (Ilustración 16, 18).

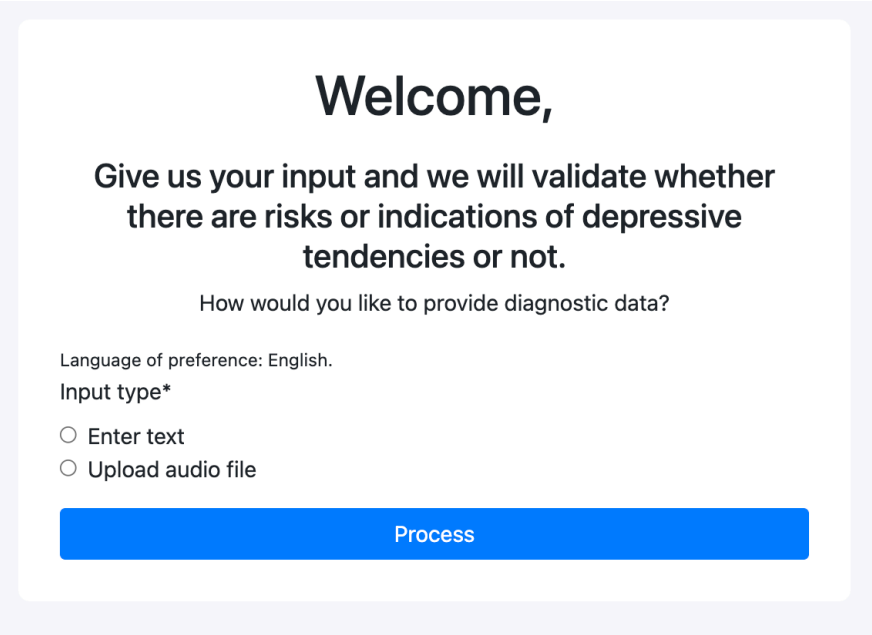

Ilustración 15 Menú principal de aplicación para diagnóstico.

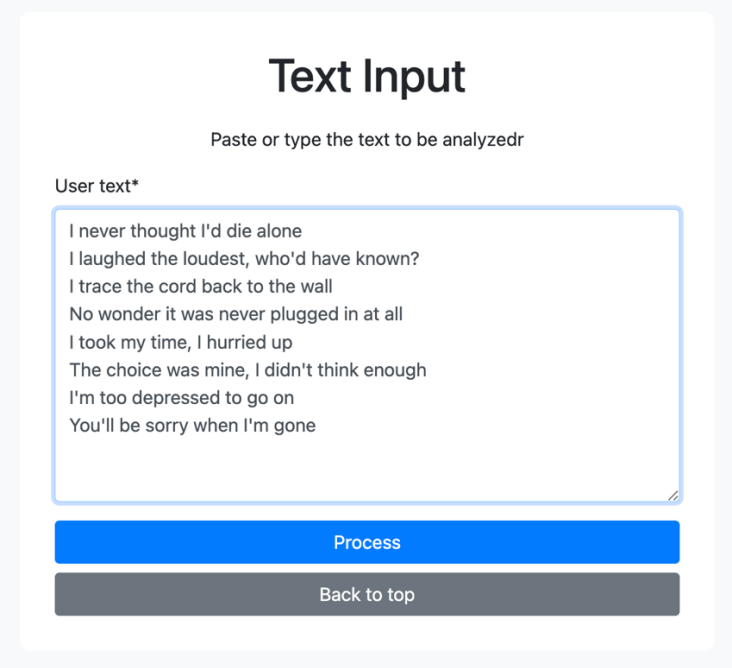

Ilustración 16 Ejemplo de entrada de texto.

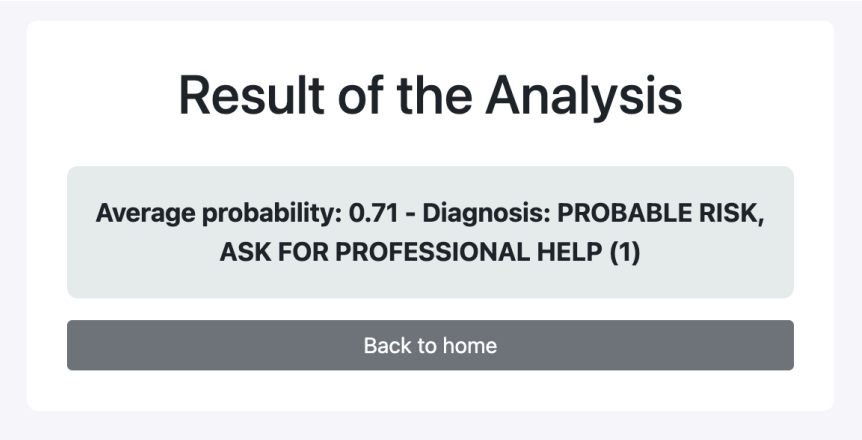

Ilustración 17 Ejemplo de diagnóstico con riesgo probable de suicidio.

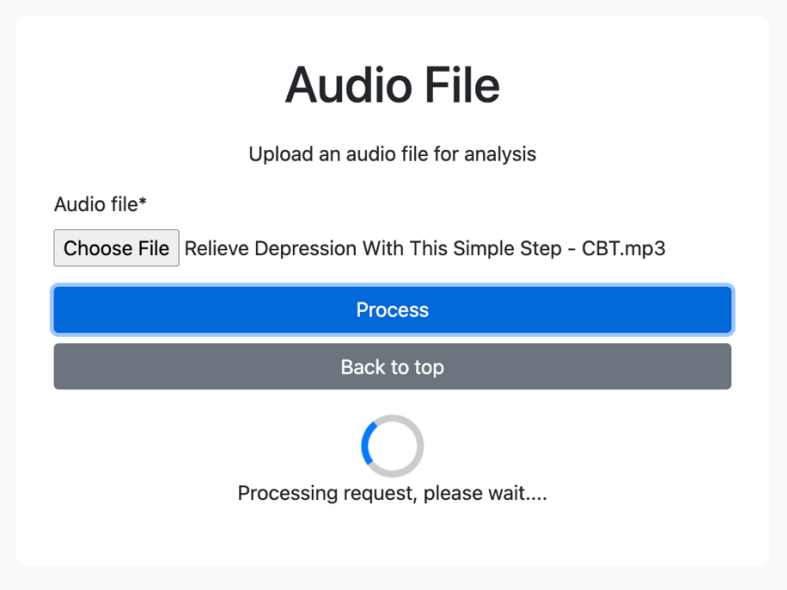

Ilustración 18 Ejemplo de cargue de entrada de audio.

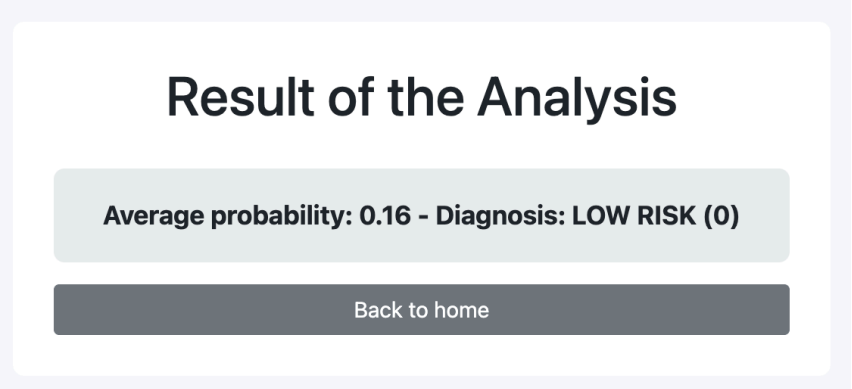

Ilustración 19 Ejemplo resultado de análisis bajo, sin riesgo de suicidio.

Considerando lo sensible del tema y con el fin de garantizar la privacidad el usuario final, se decide no mencionar la palabra suicidio dentro de la interfaz gráfica de la solución.

1. Entorno de ejecución

Para ejecutar esta solución de forma adecuada se recomienda utilizar un entorno virtual con Python 3.8 o superior, instalar las dependencias especificadas en el archivo de requisitos (por ejemplo, mediante pip install -r requirements.txt), verificar que se disponga de los modelos "Local" y "Unificado" en la ruta designada (con formato .pkl para joblib), así como contar con las credenciales de la API de OpenAI (establecidas en una variable de entorno) para habilitar el componente de ChatGPT; además, es preciso asegurarse de que las librerías de reconocimiento de voz (como whisper) estén correctamente instaladas, lo cual puede requerir configuraciones específicas de GPU o CUDA si se busca un mayor rendimiento; finalmente, se deben configurar las rutas y ajustes de Django, incluyendo la URL de la aplicación y la base de datos, antes de ejecutar las migraciones y levantar el servidor (por ejemplo, con python manage.py runserver), garantizando con ello que la interfaz web, los modelos de clasificación y los módulos de transcripción funcionen de forma integrada y ofrezcan el diagnóstico final en el navegador. Se recomienda el uso de lenguaje inglés para uso del aplicativo considerando el entrenamiento y dataset utilizado en este trabajo.

VIII. CONCLUSIONES Y TRABAJO FUTURO

La solución presentada para la **detección temprana de posibles tendencias suicidas** integra de manera **multimodal** tres fuentes principales de predicción: un **Modelo Local** basado en técnicas estadísticas tradicionales (TF-IDF y Regresión Logística), un **Modelo Unificado** que combina embeddings semánticos (p. ej., BERT, VADER) y rasgos de audio (ej. rms_mean, pitch_mean, tempo, etc.) mediante un Random Forest, y un **componente complementario de ChatGPT** que, gracias a su capacidad de razonamiento contextual, otorga un análisis profundo en textos ambiguos o complejos. A continuación se destacan los principales hallazgos y reflexiones:

Eficiencia vs. Sensibilidad

- El **Modelo Local** se caracteriza por un **equilibrio alto** entre precisión y sensibilidad, brindando un desempeño estable y con menor tasa de falsos positivos que el Modelo Unificado.
- El **Modelo Unificado**, por su parte, opta por **maximizar el recall** para la clase con riesgo suicida, reduciendo drásticamente los falsos negativos pero incrementando los falsos positivos. Esta aproximación resulta muy valiosa en entornos donde no se puede permitir la omisión de casos críticos.
- **ChatGPT**, al incorporar una comprensión semántica avanzada, permite complementar y confirmar resultados, especialmente en casos con alto nivel de ambigüedad lingüística, beneficiándose de su entrenamiento a gran escala.

Aplicabilidad Clínica

En el contexto de la **detección temprana de tendencias suicidas**, minimizar falsos negativos es vital para no pasar por alto casos potencialmente graves. Sin embargo, un exceso de falsos positivos puede generar sobrecarga de recursos médicos y psicológicos.

La integración de ChatGPT como **componente principal** de diagnóstico (con un peso mayor en la votación) aporta un "criterio humano-símil", al capturar matices complejos que modelos puramente estadísticos o basados en embeddings pueden pasar por alto.

Por lo tanto, la selección y calibración de estos modelos (Local, Unificado, ChatGPT) debe ajustarse a la infraestructura y a la tolerancia a cada tipo de error en el sistema de salud.

1. **Relevancia del Conjunto de Datos**

- El dataset inicial, con aprox. de 23 000 muestras y variables textuales y acústicas, ha permitido entrenar y validar los modelos con **robustez estadística**.
- No obstante, se hace necesaria la **validación clínica** adicional: la confiabilidad del sistema en contextos reales depende de la diversidad y representatividad de los datos.
- Para afinar la detección de riesgos, cada dimensión (texto, audio y sentimiento) deberá seguir ajustándose en función de datos actualizados y garantizando la calidad de la etiqueta "riesgo suicida".

2. **Valor de la Multimodalidad y el Ensamble**

- La combinación de **NLP** (para texto), **análisis de voz** (rasgos acústicos) y **modelos generativos** (ChatGPT) consolida un enfoque integral que captura tanto aspectos lingüísticos explícitos como señales paralingüísticas.
- Esta **fusión** permite un **sistema de alerta temprana** versátil, con capacidad de adaptarse a distintos tipos de entradas (transcripciones de audio, mensajes escritos, contenido de redes sociales) y de proporcionar una salida binaria (riesgo/no riesgo) con bases teóricas y prácticas claras.

En síntesis, la investigación confirma la viabilidad de un **modelo multimodal** que, mediante el uso de **técnicas de aprendizaje automático**, **análisis de texto** y **señales de voz**, junto con la potencia de **ChatGPT**, puede **alertar de forma oportuna** sobre eventuales **intenciones suicidas**, contribuyendo así al **fortalecimiento de las estrategias de prevención** y **reduciendo la dependencia exclusiva** de evaluaciones clínicas subjetivas. No obstante, el **dilema principal** radica en equilibrar la **sensibilidad (recall)** con la **especificidad**, de acuerdo con los recursos clínicos disponibles, al tiempo que se amplía el **alcance y la escalabilidad** de los diagnósticos de riesgo en diversos contextos. Además, este enfoque ofrece un **puente hacia futuras investigaciones**, en las cuales la **calidad de los datos**, la **validación clínica** y la **integración con canales de ayuda** jugarán un papel determinante para **consolidar** y **optimizar** la utilidad práctica de este modelo en entornos reales.

Como trabajo futuro, se contempla:

1. **Validación Clínica con Expertos**

- Involucrar a psiquiatras y psicólogos en la **evaluación sistemática** de los resultados. Contrastando las predicciones del sistema contra diagnósticos

médicos confirmados, se garantiza que la herramienta responda a parámetros de calidad y rigor clínico.

2. **Datos Reales y Ampliados**

- Profundizar en la recolección de **datos validados** de pacientes con experiencias confirmadas. Aumentar la heterogeneidad de muestras (distintas edades, culturas, niveles socioeconómicos) fortalecerá la capacidad de generalización de los modelos.
- Explorar convenios con instituciones de salud mental para lograr un **dataset real** que refleje con mayor fidelidad los casos clínicos.

3. **Integración con Canales de Ayuda**

- Diseñar flujos de **alerta automatizada** que, ante una detección de riesgo elevada, notifiquen directamente a líneas de emergencia psicológica o profesionales de guardia.
- Incluir **módulos de derivación** para que el paciente sea contactado inmediatamente por servicios de atención o familiares cercanos, acortando el tiempo de respuesta ante eventuales crisis.

4. **Modelo Exclusivo de Audio**

- Desarrollar un sistema centrado únicamente en el análisis de voz, dando **mayor énfasis a las variables acústicas**. Examinar cómo rasgos como la entonación, el tempo, la variabilidad del pitch o la intensidad pueden, por sí solos, discriminar estados emocionales críticos.
- Comparar este modelo "solo audio" con la versión multimodal, determinando en qué contextos puede ser suficiente o incluso ventajoso.

5. **Optimización del Dataset para Entrenamientos de Audio**

- **Perfeccionar** la calidad de las grabaciones, normalizar niveles de ruido, obtener muestras con diferentes condiciones de ambiente (ruido de fondo, variaciones de micrófono, etc.).
- Incluir etiquetado especializado (fusión de "emociones" y "riesgo suicida") para robustecer el aprendizaje supervisado en audio.

En conclusión, el **desarrollo futuro** se dirige a reforzar la **validez clínica** y **eficiencia operativa** del sistema, enfatizando la incorporación de datos más representativos, la implementación de mecanismos de alerta inmediatos y la especialización en el análisis de audio. De esta forma, se consolida la propuesta de un **modelo multimodal integrado** que sirva como herramienta de **prevención, monitoreo y soporte** en entornos de salud mental y comunidades vulnerables.

IX. ANEXOS

- https://github.com/kristellu/dataset
- https://github.com/kristellu/diagnosis_model
- https://github.com/kristellu/diagnosis_app

X. REFERENCIAS

[1] Organization, World Health, «World Health Organization,» World Health Organization, [En línea]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide.

[2] G. Coppersmith, K. Ngo, R. Leary y A. Wood, «Exploratory Analysis of Social Media Prior to a Suicide Attempt,» *roceedings of the Third Workshop on Computational Linguistics and Clinical Psychology,* nº doi: 10.18653/v1/W16-0311., p. pp. 106–117, 2016.

[3] P. Burnap, G. Colombo y J. Scourfield, «Machine classification and analysis of suicide-related communication on Twitter,,» *Proceedings of the 26th ACM Conference on Hypertext and Social Media,* nº doi: 10.1145/2700171.2791023., 2015.

[4] N. Cummins, S. Amiriparian, S. Ottl, M. Gerczuk, M. Schmitt y B. Schuller, «Multimodal bag-of-words for cross domains sentiment analysis,» *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings,* Vols. %1 de %2vol. 2018-April, 2018., nº doi: 10.1109/ICASSP.2018.8462660, 2018.

[5] Samuel, A. L., «Some studies in machine learning using the game of checkers. ii—recent progress,,» *BM Journal of research and development,* Vols. %1 de %2vol. 11, no. 6, pp. 601–617, 1967.

[6] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus y J. Gonzàlez, «Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis,,» *J. Med. Internet Res.,* Vols. %1 de %2vol. 22, no. 7, nº doi: 10.2196/17758, 2020.

[7] A. Singh, N. Thakur, and A. Sharma, «A review of supervised machine learning algorithms,,» *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE,* p. pp. 1310–1315., 2016.

[8] Carey, R. Gentleman and V. J., «Unsupervised machine learning,» *Bioconductor case studies. Springer,* p. pp. 137–157., 2008.

[9] L. P. Kaelbling, M. L. Littman, and A. W. Moore, «Reinforcement learning: A survey,» *Journal of artificial intelligence research,* vol. vol. 4, p. pp. 237–285, 1996.

[10] Barto, R. S. Sutton and A. G., «Reinforcement learning: An introduction,» *MIT press,* 2018.

[11] S. B. Kotsiantis, I. Zaharakis, P. Pintelas et al., «Supervised machine learning: A review of classification techniques,» *Emerging artificial intelligence applications in computer engineering,* Vols. %1 de %2vol. 160, no. 1, p. pp. 3–24, 2007.

[12] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett y A. Leenaars, «Suicide Note Classification Using Natural Language Processing: A Content Analysis,» *Biomed. Inform. Insights,* vol. vol. 3, nº doi: 10.4137/bii.s4706, 2010.

[13] F. Pedregosa et al., «cikit-learn: Machine Learning in Python,,» *J. Mach. Learn. Res,* vol. vol. 12, p. pp. 2825–2830, 2011.

[14] I. Goodfellow, Y. Bengio, and A. Courville, «Deep learning,» *MIT press,* 2016.

[15] C. M. Bishop, «Neural networks and their applications,» *Review of scientific instruments,* Vols. %1 de %2vol. 65, no. 6,, p. pp. 1803–1832, 1994.

[16] D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, D. A. Velazquez, J. M. Gonfaus y J. Gonzàlez, «Detection of suicidal ideation on social media: Multimodal, relational, and behavioral analysis,» *J. Med. Internet Res,* Vols. %1 de %2vol. 22, no. 7, nº 10.2196/17758., 2020.

[17] A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins, «Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets,» *Molecular pharmaceutics,* Vols. %1 de %2vol. 14, no. 12, p. pp. 4462–4475, 2017.

[18] Fawcett, T, «An Introduction to ROC Analysis,» *Pattern Recognition Letters,* vol. 27(8), p. 861–874, 2006.

[19] S. Narkhede, «Understanding auc-roc curve,» *owards Data Science,* Vols. %1 de %2vol. 26, no. 1, p. pp. 220–227, 2018.

[20] scikit-learn documentation, «FeatureUnion" and "GridSearchCV,,» [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

[21] B. McFee et al., «librosa: Audio and Music Signal Analysis in Python,» *Proc. 14th Python in Science Conf.,* 2015.

[22] Martin, D. Jurafsky y J. H., de *Speech and Language Processing,3rd ed*, Prentice Hall, 2020.

[23] B. Pang y L. Lee, «Opinion Mining and Sentiment Analysis,» *Found. Trends Inf. Retr,* Vols. %1 de %2vol. 2, no. 1–2, p. pp. 1–135, 2008.

[24] A. Vaswani et al., «Attention is All you Need,» *Proc. NeurIPS,* p. pp. 5998–6008, 2017.

[25] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» *Proc. NAACL-HLT,* p. pp. 4171–4186, 2019.

[26] T. Mikolov et al., «Efficient Estimation of Word Representations in Vector Space,» *Proc. Int. Conf. Learn. Representations (ICLR),* 2013.

[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf,, «DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,» *arXiv preprint arXiv:1910.01108,* 2019.

[28] N. Reimers and I. Gurevych, «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,,» *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP),* p. pp. 3982–3992, 2019.

[29] C. J. Hutto y E. Gilbert, «VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,» *Proc. Int. Conf. Weblogs Soc. Media,* p. pp. 216–225, 2014.

[30] D. M. Blei, A. Y. Ng y M. I. Jordan, «Latent Dirichlet Allocation,» *J. Mach. Learn. Res.,* vol. vol. 3, p. pp. 993–1022, 2003.

[31] OpenAI, «Whisper: Robust Speech Recognition via Large-Scale Weak Supervision,» [En línea]. Available: https://github.com/openai/whisper..

[32] J. Ramos, «Using TF-IDF to Determine Word Relevance in Document Queries,,» *Proceedings of the First Instructional Conference on Machine Learning,* 2003.

[33] D. W. Hosmer, S. Lemeshow y R. X. Sturdivant, «Applied Logistic Regression,» *3rd ed. Wiley,* 2013.

[34] Z. C. Lipton, J. Berkowitz y C. Elkan, «A Critical Review of Recurrent Neural Networks for Sequence Learning,,» nº arXiv:1506.00019, 2015.

[35] J. Bergstra and Y. Bengio, «Random Search for Hyper-Parameter Optimization,» *Journal of Machine Learning Research,* vol. vol. 13, p. pp. 281–305, 2012.

[36] L. Breiman, «Random Forests,» *Mach. Learn.,* Vols. %1 de %2vol. 45, no. 1, p. pp. 5–32, 2001.

[37] K. Atrey, M. Hossain, A. El Saddik y M. K. Morstatter, «Multimodal fusion for multimedia analysis: a survey,,» *ultimed. Tools Appl,* Vols. %1 de %2 vol. 78, no. 3, p. pp. 243–296, 2013.

[38] M. Mitchell, «Web Scraping with Python: Collecting Data from the Modern Web,» *O'Reilly Media,* 2018.

[39] B. Liu, «Sentiment Analysis and Opinion Mining,» *Synthesis Lectures on Human Language Technologies,* Vols. %1 de %2vol. 5, no. 1, p. pp. 1–167, 2012.

[40] S. Ruder, «An Overview of Transfer Learning,» 2019. [En línea]. Available: https://ruder.io/transfer-learning/.

[41] S. Giarratano y J. Riley, «Expert Systems: Principles and Programming,» *PWS Publishing,* vol. Expert Systems: Principles and Programming, 2005.

[42] U. M. Fayyad, G. Piatetsky-Shapiro y P. Smyth, «From Data Mining to Knowledge Discovery in Databases,» *AI Mag,* Vols. %1 de %2ol. 17, no. 3,, p. pp. 37–54, 1996.

[43] A. Johnson et al., «Deep Learning for Mental Health Surveillance using Social Media: A Systematic Review,,» *arXiv preprint,* nº arXiv:2204.XXXX, 2022.

[44] K. Müller and S. Scherer, «Investigating Acoustics, Facial Expressions, and Language for Depression Recognition,» *Proceedings of the ACM Multimedia,* p. pp. 121–130, 2022.

[45] S. Lundberg and S. Lee, «A Unified Approach to Interpreting Model Predictions,» *Proc. Advances in Neural Information Processing Systems (NIPS),* p. pp. 4765–4774, 2017.

[46] M. Torous et al., «Growing Pains and Growing Needs: Challenges and Strategies for Data-Driven Mental Health,» *Translational Psychiatry,* Vols. %1 de %2vol. 12, no. 1, pp. p. 347, 2022.

[47] R. Zhao and L. Wang, «Privacy-Preserving Data Mining in Healthcare: Current Trends and Future Directions,» *IEEE Transactions on Big Data,* Vols. %1 de %2vol. 9, no. 2, p. pp. 99–112, 2023.

[48] A. M. Cox, «Legal and Ethical Considerations of AI in Mental Health,» *AI & Society,* Vols. %1 de %2vol. 39, no. 4, p. pp. 1261–1269, 2024.

[49] P. Kumar et al., «Domain Adaptation for Mental Health Classification in Social Media,,» *IEEE Transactions on Neural Networks and Learning Systems,,* Vols. %1 de %2vol. 34, no. 1, p. pp. 138–149, 2023.

[50] K. Urueta and J. Márquez, «Application of Natural Language Processing and Deep Learning Models for the Detection of Mental Health Disorders in Social Networks,» *IEEE LATAM Conference - submitted* , p. pp. 1–7, 2024.

[51] G. Salton and C. Buckley, «Term-weighting approaches in automatic text retrieval,» *Information Processing & Management,* Vols. %1 de %2vol. 24, no. 5, pp. pp. 513-523, 1988.

[52] F. Eyben, K. R. Scherer, B. W. Schuller, «Advances in Audio Emotion Recognition,» *IEEE Transactions on Affective Computing,* Vols. %1 de %2vol. 1, no. 2, pp. pp. 18-31, 2010.

[53] T. Giannakopoulos, «Audio analysis algorithms: a Matlab implementation,» *Wiley,* 2013.

[54] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. Narayanan, «IEMOCAP: Interactive emotional dyadic motion capture database,» *Language Resources and Evaluation,* Vols. %1 de %2vol. 42, no. 4, pp. pp. 335-359, 2008.

[55] D. Parker, «Algorithms for spectral centroid calculation,» *Journal of the Audio Engineering Society,* Vols. %1 de %2vol. 42, no. 6, pp. pp. 413-420, 1994.

[56] B. Logan, «Mel Frequency Cepstral Coefficients for Music Modeling,» *Proceedings of the International Symposium on Music Information Retrieval (ISMIR),* 2000.

[57] M. Pedregosa et al., «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research,,* vol. vol. 12, p. pp. 2825–2830, 2011.

[58] D. Freedman, «Statistical Models: Theory and Practice,» *Cambridge University Press,* 2009.

[59] S. Sanh, L. Debut, J. Chaumond, and T. Wolf, «DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,,» *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing,* 2019.

[60] T. Wolf et al., «Transformers: State-of-the-Art Natural Language Processing,» *Proc. EMNLP,* 2020.

[61] OpenAI, «ChatGPT Model Documentation,» 2023. [En línea]. Available: https://platform.openai.com/docs.

[62] Kohavi, R., «A study of cross-validation and bootstrap for accuracy estimation and model selection,» *International Joint Conference on Artificial Intelligence (IJCAI),* 1995.

[63] He, H., & Garcia, E. A., «Learning from imbalanced data,» *IEEE Transactions on Knowledge and Data Engineering,* Vols. %1 de %2vol. 21, no. 9, pp. pp. 1263-1284, 2009.

[64] Rabiner, L., & Juang, B. H, «Fundamentals of Speech Recognition. Prentice-Hall,» 1993.

[65] Z.-H. Zhou, «Ensemble Methods: Foundations and Algorithms,» *CRC Press,* 2012.

[66] N. Japkowicz and M. Shah, «Evaluating Learning Algorithms: A Classification Perspective,» *Cambridge University Press,* 2011.

[67] M. Sokolova and G. Lapalme, «A systematic analysis of performance measures for classification tasks,» *Information Processing & Management,* Vols. %1 de %2 vol. 45, no. 4, p. pp. 427–437, 2009.

[68] R. Kohavi and F. Provost, «Glossary of terms,» *Machine Learning,* Vols. %1 de %2ol. 30, no. 2–3, p. pp. 271–274, 1998.

[69] A. Y. Kim, E. H. Jang, S. H. Lee, K. Y. Choi, J. G. Park y H. C. Shin, «Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach,» *Med. Internet Res,* vol. vol. 25, nº doi: 10.2196/34474, 2023.

[70] S. Gupta and V. Kumar, «Suicidal Ideation Detection on Reddit: A Transfer Learning Approach,» *IEEE Access,* vol. vol. 10, p. p. 17255–17266, 2022.

[71] X. Li et al, «A Multimodal Deep Learning Framework for Detecting Suicidal Ideation,» *Deep Learning Methods,* Vols. %1 de %2ol. 35, no. 3,, p. pp. 98–105, 2023.

[72] H. J. Ewell et al., «Physiological Markers in Predicting Suicidal Behavior: A Systematic Review,,» *International Journal of Environmental Research and Public Health,* Vols. %1 de %2vol. 19, no. 15, p. p. 9072, 2022.

[73] L. Smith, «Virtual Suicidal Alert Intelligent System (VSAIL) in Electronic Health Records,» *EPE Journal,* Vols. %1 de %2vol. 54, no. 2, p. p. 45–53, 2024.

[74] M. O'Connor et al., «Predicting Suicide Attempts Using ML Approaches on EHR Data,,» *Journal of Affective Disorders,* vol. vol. 303, p. pp. 477–483, 2022.

[75] L. Perez et al., «Fusion Strategies for Multimodal Depression Detection,» *Proc. of Interspeech,* vol. 2023, p. pp. 3120–3124..

[76] OpenAI, «ChatGPT: Optimizing Language Models for Dialogue,» 2022. [En línea]. Available: https://openai.com/blog/chatgpt.

[77] Frye, J. Wilson and D., «Evaluating ChatGPT for Suicide Risk Assessment: A Clinical Perspective,,» *Journal of Telemedicine and Telecare,* Vols. %1 de %2 vol. 30, no. 1, p. pp. 55–61, 2024.

[78] Geekflare, «Youper: AI Therapy for Emotional Wellbeing,» 2024. [En línea]. Available: https://geekflare.com/.

[79] Morning Dough, «AI Chatbots for Mental Health,» 2024. [En línea]. Available: https://morningdough.com/.

[80] ConSalud, «Aimentia Health: Transformación Digital en Salud Mental,» 2023. [En línea]. Available: https://consalud.es/.

[81] S. Davis, P. Mermelstein, «omparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,» *IEEE Transactions on Acoustics, Speech, and Signal Processing,* Vols. %1 de %2vol. 28, no. 4, pp. pp. 357-366.

[82] E. J. Humphrey, S. Bello, and K. P. W. A. M. Bolton, «Moving Beyond Feature Engineering: Deep Architectures and Automatic Feature Learning in Music Informatics,» *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014),* 2014.

[83] Bishop, C. M., «Pattern Recognition and Machine Learning.,» *Springer,* 2006.

[84] O. Chorowski, D. Bahdanau, K. Cho, Y. Bengio, «End-to-End Continuous Speech Recognition using Attention-based Recurrent NN: First Results,» *Proc. NIPS Workshop,,* 2015.

[85] T. Brown et al., «Language Models are Few-Shot Learners,» *Proc. NeurIPS ,* 2020.