UNIVERSIDAD LAICA “ELOY ALFARO” DE MANABÍ

FACULTAD CIENCIAS DE LA VIDA Y TECNOLOGÍA

CARRERA TECNOLOGÍAS DE LA INFORMACIÓN

TRABAJO BAJO LA MODALIDAD DE PROYECTO DE INVESTIGACIÓN

TEMA:

Redes Neuronales Artificiales para valorar la percepción de los pacientes respecto de la efectividad de medicamentos.

AUTOR:

Chávez Moreira Elvis Joel

TUTOR DE TRABAJO DE TITULACIÓN

Ing. Jorge Iván Pincay Ponce, PHD.

Manta - Manabí - Ecuador

2025(2)

CERTIFICACIÓN

En calidad de docente tutor de la Facultad de Ciencias de la Vida y Tecnología de la Universidad Laica "Eloy Alfaro" de Manabí, CERTIFICO:

Haber dirigido, revisado y aprobado preliminarmente el Trabajo de Integración Curricular bajo la autoría del estudiante **Chávez Moreira Elvis Joel**, legalmente matriculado en la carrera de Tecnologías de la Información, período académico 2025 (2), cumpliendo el total de 384 horas, cuyo tema del proyecto es ***"Redes Neuronales Artificiales para valorar la percepción de los pacientes respecto de la efectividad de medicamentos"***.

La presente investigación ha sido desarrollada en apego al cumplimiento de los requisitos académicos exigidos por el Reglamento de Régimen Académico y en concordancia con los lineamientos internos de la opción de titulación en mención, reuniendo y cumpliendo con los méritos académicos, científicos y formales, y la originalidad de este, requisitos suficientes para ser sometida a la evaluación del tribunal de titulación que designe la autoridad competente.

Particular que certifico para los fines consiguientes, salvo disposición de Ley en contrario.

Manta, 29 de enero de 2026.

Lo certifico,

Ing. Pincay Ponce Jorge Ivan, PhD.
Docente Tutor
Facultad de Ciencias de la Vida y Tecnología
Carrera Tecnologías de la Información

DECLARACIÓN EXPRESA DE AUTORÍA

Yo, Elvis Joel Chávez Moreira, con número de cédula 131661133-2, estudiante de la carrera de Tecnologías de la Información en la Universidad Laica Eloy Alfaro de Manabí, declaro que el presente trabajo de tesis titulado "**Redes Neuronales Artificiales para valorar la percepción de los pacientes respecto de la efectividad de medicamentos**" ha sido realizado por mi propia autoría bajo la supervisión de mi tutor, Ing. Jorge Iván Pincay Ponce, PhD.

Mediante la presente, asumo la plena responsabilidad intelectual sobre el contenido de este documento y hago constar que esta obra es inédita, producto de mi ejecución personal; por tanto, carece de reproducciones no autorizadas o apropiaciones indebidas de fuentes externas. Toda idea o concepto ajeno incorporado al texto ha sido debidamente acreditado mediante el sistema de referencias correspondiente. Este Proyecto Integrador se somete íntegramente a los códigos de honestidad y rigor académico estipulados por la Universidad Laica Eloy Alfaro de Manabí.

En constancia de lo expresado, firmo la presente declaración.

Elvis Joel Chávez Moreira

Febrero del 2026

Dedicatoria

Dedico este trabajo a mis padres por ser el pilar fundamental de mi vida, por ser mi apoyo y motivación, por alentarme a seguir mis sueños y siempre apoyarme en todas mis metas y objetivos. A mis hermanos por su apoyo inquebrantable y su fe en mí. Sus palabras de aliento y su ejemplo me motivaron a seguir adelante.

A mis amigos, quienes he tenido la dicha de conocer en diferentes etapas y con objetivos distintos: aquellos del colegio, que compartieron conmigo los primeros sueños; los de la universidad, que estuvieron en las largas jornadas de esfuerzo y aprendizaje. Cada uno de ustedes ha sido una parte importante de este camino y ha hecho de esta experiencia algo inolvidable.

Gracias por ser mi inspiración, mi apoyo y mi motivación. Este logro es para ustedes.

Elvis Joel Chávez Moreira

Agradecimiento

Quiero expresar mi más sincero agradecimiento al Ing. Jorge Pincay por ser un excelente tutor, por compartir conmigo sus valiosos conocimientos y por su constante guía en cada etapa de este proceso. Su conocimiento, dedicación y apoyo constante fueron fundamentales para que este trabajo llegara a buen término. Gracias por creer en este proyecto y por compartir toda su experiencia de manera tan generosa conmigo.

Extiendo también mi gratitud a la facultad y a todos los profesores, cuyo apoyo y enseñanzas me han acompañado a lo largo de esta trayectoria académica.

Elvis Joel Chávez Moreira

CAPÍTULO I - INTRODUCCIÓN 15

1. Introducción 15

1.1. Presentación del tema 16

1.2. Ubicación y Contextualización de la Problemática 16

1.3. Planteamiento de problema 17

1.4. Objetivos 18

1.4.1. Objetivo General 18

1.4.2. Objetivos Específicos 18

1.5. Justificación 18

CAPÍTULO II - MARCO TEÓRICO 20

2. Antecedentes de investigaciones 20

2.1. Farmacovigilancia Digital 20

2.1.1. Minería de opiniones en salud 21

2.1.2. Brecha entre Eficacia y Efectividad: Diferencia Teórica Fundamental 22

2.2. Metodología de Ciencia de Datos 23

2.2.1. Metodología CRISP-DM 23

2.3. Fundamentos de Procesamiento de Lenguaje Natural 27

2.3.1. Preprocesamiento de Texto 27

2.3.2. Tokenización y Vectorización 28

2.3.3. Word Embeddings 29

2.4. Ingeniería de Características 30

2.4.1. Taxonomías y Categorización Médica 30

2.4.2. Análisis Temporal en Salud: Estacionalidad y Patrones en Adherencia y Efectividad 31

2.5. Redes Neuronales Artificiales (Deep Learning) 33

2.5.1. Redes Neuronales Recurrentes (RNN) 33

2.5.2. El Problema del Desvanecimiento del Gradiente 33

2.5.3. Arquitectura LSTM 35

2.5.4. Arquitectura GRU 36

2.6. Dinámica del Entrenamiento y Optimización 37

2.6.1. Función de Pérdida 37

2.6.2. Optimizador Adam 38

2.6.3. Sobreajuste y Generalización 39

2.6.4. Parada Temprana 40

2.7. Métricas de Evaluación de Modelos 41

2.7.1. Matriz de Confusión 41

2.7.2. Accuracy (Exactitud) vs. F1-Score 42

2.8. Principales tipos de medicamentos empleados en el Dataset 43

2.8.1. Medicamentos con mayor frecuencia de reseñas 44

2.8.2. Distribución de medicamentos por categoría médica 44

2.8.3. Distribución general de sentimientos 45

2.9. Conclusiones relacionadas al Marco Teórico 46

CAPÍTULO III - MARCO INVESTIGATIVO 48

3. Introducción 48

3.1. Tipo de investigación 48

3.2. Metodología CRISP-DM 49

3.2.1. Relación con los objetivos del proyecto 49
3.3. Limitaciones del Estudio 50
3.4. Ética y Consideraciones de Privacidad 51
CAPÍTULO IV - MARCO PROPOSITIVO **52**
4. Introducción 52
4.1. Descripción de la Propuesta 52
4.2. Determinación de Recursos 53
4.2.1. Recursos Humanos 53
4.2.2. Recursos Tecnológicos 54
4.2.3. Recursos Económicos 54
4.3. Aplicación de la Metodología CRISP-DM 55
4.4. Fases de la Metodología CRISP-DM 55
4.4.1. Comprensión del Negocio (Business Understanding) 55
4.4.2. Comprensión de los Datos (Data Understanding) 57
4.4.3. Preparación de los Datos (Data Preparation) 58
4.4.4. Modelado (Modeling) 62
4.4.5. Evaluación (Evaluation) 66
4.4.6. Despliegue (Deployment) 67
CAPÍTULO V - EVALUACIÓN DE RESULTADOS **76**
5. Análisis del entrenamiento y convergencia 76
5.1. Análisis del comportamiento del entrenamiento 76
5.1.1. Comportamiento de las curvas de aprendizaje 76
5.2. Resultados comparativos de las arquitecturas 80
5.2.1. Métricas de rendimiento (Eficacia) 80

5.2.2. Análisis de tiempos y costo computacional 81

5.3. Análisis detallado por clase (Sentimiento) 82

5.4. Análisis de la matriz de confusión 83

5.5. Despliegue tecnológico y herramienta de validación 85

5.6. Discusión de resultados 86

CAPÍTULO VI - CONCLUSIONES Y RECOMENDACIONES 87

6. Conclusiones 87

6.1. Recomendaciones 90

Trabajos futuros 92

Referencias 93

Anexos 97

Índice de Ilustraciones

Ilustración 1 Arquitectura LSTM 64
Ilustración 2 Arquitectura GRU 65
Ilustración 3 Módulo de inicio y métricas globales 69
Ilustración 4 Módulo de comparación de modelos 70
Ilustración 5 Módulo de visualización de curvas de entrenamiento 71
Ilustración 6 Matriz de confusión y métricas por clase 72
Ilustración 7 Análisis por categoría médica 73
Ilustración 8 Módulo de validación manual 74
Ilustración 9 Curvas completas LSTM 77
Ilustración 10 Curvas Óptimas LSTM 78
Ilustración 11 Curvas completas GRU 79
Ilustración 12 Curvas óptimas GRU 79
Ilustración 13 Matriz de confusión del modelo GRU 84

Índice de Tablas

Tabla 1 Comparativa de Accuracy VS F1 Score..... 43
Tabla 2 Tabla de presupuesto estimado para el desarrollo del proyecto..... 55
Tabla 3 Requisitos Funcionales y Criterios de Aceptación..... 56
Tabla 4 Requisitos No Funcionales..... 57
Tabla 5 Comparativa de rendimiento técnico entre ambos modelos. 80
Tabla 6 Comparativa de tiempos totales de los entrenamientos 81
Tabla 7 Reporte de Clasificación por Sentimiento (Modelo GRU) 82

RESUMEN

Contexto: La presente investigación se enmarca en el proceso de transformación de la farmacovigilancia hacia entornos digitales, en los cuales el volumen creciente de información no estructurada generada por los pacientes supera ampliamente la capacidad de los métodos tradicionales de análisis, por lo que, pese a los avances computacionales, persiste una brecha significativa entre la eficacia clínica reportada en ensayos controlados y la efectividad percibida en la práctica real, lo que limita su aprovechamiento para la toma de decisiones en salud pública

Objetivo: Este proyecto aborda el desafío de ingeniería que supone automatizar la detección de polaridad (sentimiento) en reseñas sobre medicamentos. Sin embargo, la investigación va más allá de una clasificación básica: se planteó como una confrontación técnica entre dos arquitecturas de memoria profunda, LSTM y GRU. La prioridad central fue determinar, de manera empírica, cuál de estos modelos logra mantener una alta precisión predictiva sin consumir excesivos recursos computacionales. En definitiva, se buscaba ese equilibrio ideal que permitiera un despliegue viable en entornos reales.

Metodología: La hoja de ruta del proyecto siguió la disciplina del estándar CRISP-DM para gestionar la complejidad inherente al "Drug Review Dataset" (UCI). El tratamiento de los datos exigió un flujo de saneamiento bastante riguroso, que transformó más de 160.000 registros crudos en secuencias matemáticas estandarizadas de 200 tokens. Esto conduce a la fase de entrenamiento, donde se explotó la capacidad GPU de Kaggle mediante una configuración de alto rendimiento. Es importante destacar que se forzaron lotes masivos de 2048 muestras a lo largo de 300 épocas, una táctica diseñada específicamente para acelerar la convergencia del gradiente y, en consecuencia, reducir drásticamente los tiempos de espera.

Resultados: Aunque ambas arquitecturas alcanzaron un empate técnico en exactitud (cercana al 83.3%), la red GRU marcó la diferencia en eficiencia operativa al reducir el tiempo de entrenamiento en casi una hora (8.3 vs 9.25 horas) y requerir una menor carga paramétrica. Mientras que la detección de sentimientos polarizados fue robusta (F1-Score > 0.90), la clase neutral se perfiló como el mayor desafío debido a su ambigüedad semántica. El proyecto

culminó integrando estos modelos en un dashboard de Streamlit, facilitando así una auditoría manual que trasciende las métricas puras

Conclusiones: Los datos confirman que la arquitectura GRU bidireccional constituye la alternativa técnica superior para esta tarea, pues replica el rendimiento de modelos más densos consumiendo significativamente menos recursos de igual forma, se demostró que la implementación de interfaces visuales es un requisito ineludible para dotar de interpretabilidad a los modelos de "caja negra" y generar la confianza necesaria para su uso en triaje sanitario.

Trabajos futuros: Se sugiere la incorporación de modelos basados en mecanismos de atención (Transformers) con el propósito de mejorar la capacidad discriminativa frente a los matices presentes en textos médicos complejos.

Palabras clave: Análisis de sentimientos, Redes Neuronales Recurrentes, LSTM, GRU, farmacovigilancia digital, procesamiento de Lenguaje Natural.

Abstract

Context: This research is framed within the digital transformation of pharmacovigilance, where the growing volume of unstructured patient-generated information exceeds the analytical capacity of traditional methods. Despite computational advances, a significant gap remains between clinical efficacy reported in controlled trials and perceived effectiveness in real-world settings, limiting its value for public health decision-making.

Objective: This project addresses the engineering challenge of automating sentiment polarity detection in drug reviews. Beyond basic classification, the study was conceived as a technical comparison between two deep memory architectures: LSTM and GRU. The primary objective was to empirically determine which model achieves high predictive accuracy while minimizing computational resource consumption, thereby ensuring feasible deployment in real-world environments**.**

Methodology: The project followed the CRISP-DM standard to manage the complexity of the UCI Drug Review Dataset. A rigorous preprocessing pipeline transformed more than

160,000 raw records into standardized sequences of 200 tokens. The training phase leveraged Kaggle's GPU environment under a high-performance configuration, using large batch sizes of 2048 samples over 300 epochs to accelerate gradient convergence and significantly reduce training time**.**

Results: Both architectures achieved similar accuracy (approximately 83.3%). However, the GRU network demonstrated superior operational efficiency, reducing training time by nearly one hour (8.3 vs. 9.25 hours) and requiring fewer parameters. While polarized sentiment detection proved robust (F1-score > 0.90), the neutral class represented the greatest challenge due to semantic ambiguity. The project concluded with the integration of the models into a Streamlit dashboard, enabling manual auditing beyond purely quantitative metrics.

Conclusions: The bidirectional GRU architecture constitutes the technically superior alternative for this task, as it replicates the performance of denser models while consuming significantly fewer computational resources. Furthermore, the implementation of visual interfaces is essential to enhance interpretability and build trust in black-box models for healthcare triage applications**.**

Future Work: It is recommended to incorporate attention-based models, such as Transformers, to improve discriminative capacity when handling nuanced and complex medical texts.

Keywords: Sentiment Analysis, Recurrent Neural Networks, LSTM, GRU, Digital Pharmacovigilance, Natural Language Processing**.**

CAPÍTULO I - INTRODUCCIÓN

1. Introducción

La industria farmacéutica y la práctica médica han dependido tradicionalmente de ensayos clínicos controlados para determinar la seguridad y eficacia de los medicamentos (Gedyt, 2023). Sin embargo, en la era digital actual, la interacción de los pacientes con las plataformas de salud ha generado un cambio de paradigma hacia la "Farmacovigilancia 2.0", tecnologías digitales como inteligencia artificial, web scraping y análisis de redes sociales para mejorar la detección precoz de eventos adversos (Martínez López, Ascuntar Marcillo, Cárdenas Guaranán, Muñoz Chana, & Urbano Urbano, 2025).

A inicios de 2026, el 72.1% de los usuarios que buscan activamente información de salud, suelen compartir información incluyendo experiencias con tratamientos y efectos secundarios, en redes sociales como Instagram, TikTok y X. Aquello genera gran volumen de datos no estructurados como parte de reseñas, comentarios y discusiones espontáneas que representan una fuente valiosa de percepciones reales de pacientes sobre la efectividad de medicamentos (Carton-Erlandsson, y otros, 2024).

En este trabajo de titulación universitaria de nombre "Redes Neuronales Artificiales para valorar la percepción de los pacientes respecto de la efectividad de medicamentos", se aborda este desafío tecnológico mediante la implementación de técnicas de Procesamiento de Lenguaje Natural (PLN). Específicamente, se centra en la capacidad resolutiva de las Redes Neuronales Recurrentes (RNN), a través de arquitecturas como LSTM y GRU, para analizar secuencias de texto y clasificar sentimientos médicos complejos.

En este trabajo se describe el desarrollo de un modelo orientado al análisis e interpretación de reseñas escritas por pacientes, cuyo propósito es transformar valoraciones de naturaleza subjetiva en información cuantificable acerca de la efectividad percibida de diversos fármacos. Este enfoque permite acercar los resultados obtenidos mediante técnicas analíticas a la experiencia clínica del día a día, lo que ayuda a reducir la brecha existente entre los ensayos controlados y el uso real de los medicamentos en la práctica. De esta manera, se aporta una herramienta de apoyo

que puede complementar los procesos de evaluación y toma de decisiones en contextos sanitarios.

1.1. Presentación del tema

REDES NEURONALES ARTIFICIALES PARA VALORAR LA PERCEPCIÓN DE LOS PACIENTES RESPECTO DE LA EFECTIVIDAD DE MEDICAMENTOS

1.2. Ubicación y Contextualización de la Problemática

La investigación se ubica en la intersección de las Ciencias de la Computación y las Ciencias de la Salud. En el contexto global, el crecimiento exponencial de los datos generados por usuarios en internet ha superado la capacidad humana de análisis manual. Plataformas como Drugs disponible en siguiente enlace https://www.drugs.com/ o incluso también la plataforma de WebMD que se encuentra disponible en el siguiente enlace https://www.webmd.com/, dichas plataformas se encargan de acumular millones de testimonios que aportan información clave sobre la adherencia a diversos tratamientos y a su vez sobre posibles reacciones adversas, las cuales con mucha frecuencia no quedan ni registradas en los reportes oficiales.

El escenario regional presenta una desconexión tecnológica preocupante: no contamos con motores de procesamiento capaces de interpretar la voz del paciente de forma directa, sin la distorsión que introducen los intermediarios. Esta carencia nos mantiene dependientes de modelos estadísticos que han quedado obsoletos y que fracasan notablemente cuando enfrentan la ambigüedad propia del lenguaje narrativo. En este sentido, tal como advierten (Ahmad & Veeramanickam, 2024), la estadística clásica colapsa ante enunciados de naturaleza dual como "el dolor cesó, pero el mareo es insoportable". En lugar de identificar y alertar sobre el riesgo, el algoritmo tradicional simplemente promedia los vectores opuestos, lo que termina anulando el matiz semántico y provoca una pérdida de información que resulta inaceptable cuando hablamos de seguridad farmacológica. Entonces, existe la necesidad urgente de aprovechar de manera sistemática la evidencia generada en entornos donde los sistemas de salud avanzan hacia

modelos de atención más preventivos, personalizados y centrados en el paciente. Gran parte de la información sobre el impacto de los medicamentos en la vida cotidiana de los usuarios permanece dispersa en textos no estructurados. Esta falta de estructuración impide que diversos profesionales médicos, instituciones sanitarias y entidades farmacéuticas obtengan una visión integral del ciclo de vida del fármaco tras su comercialización, limitando la capacidad para identificar tendencias, efectos adversos no documentados y variaciones en la efectividad percibida por los pacientes.

1.3. Planteamiento de problema

El problema central radica en la dificultad técnica para procesar y valorar objetivamente la percepción de efectividad de los medicamentos contenida en grandes volúmenes de reseñas no estructuradas. Actualmente, existe una discrepancia significativa, conocida como la "brecha eficacia-efectividad": un medicamento puede ser eficaz en un ensayo clínico controlado, pero su efectividad percibida en el mundo real puede variar drásticamente debido a factores como la adherencia, efectos secundarios no documentados o interacciones con otros fármacos.

Las reseñas que son emitidas por los pacientes constituyen una fuente valiosa de información; sin embargo, se presentan en forma de texto libre, frecuentemente acompañadas de jerga, errores gramaticales y ambigüedades semánticas que los métodos tradicionales de análisis resultan incapaces de procesar adecuadamente. En ausencia de modelos computacionales avanzados capaces de capturar dependencias contextuales de largo alcance en el lenguaje como las arquitecturas LSTM o GRU gran parte de esta información termina desaprovechada, lo cual limita significativamente las capacidades de vigilancia farmacológica y la comprensión precisa del impacto real de los tratamientos en los usuarios.

Por lo tanto, se formula la siguiente interrogante de investigación:

¿De qué manera las Redes Neuronales Artificiales Recurrentes pueden mejorar la valoración y clasificación de la percepción de los pacientes respecto a la efectividad de los medicamentos a partir del análisis de reseñas no estructuradas?

1.4. Objetivos

1.4.1. Objetivo General

Desarrollar y validar un modelo de Redes Neuronales Artificiales Recurrentes para valorar y clasificar la percepción de los pacientes respecto de la efectividad de medicamentos, a partir del procesamiento de lenguaje natural de reseñas no estructuradas.

1.4.2. Objetivos Específicos

1. Diseñar un proceso de transformación de datos que estandarice las condiciones médicas.
2. Implementar y comparar el rendimiento de dos arquitecturas de redes neuronales (LSTM y GRU) con hiper parámetros ajustados a cada una.
3. Evaluar la capacidad del modelo en la tarea de clasificación de la percepción de los pacientes respecto de la efectividad de medicamentos, con base en métricas cuantitativas.
4. Construir reflexiones cualitativitas respecto de casos ambiguos encontrados.

1.5. Justificación

La presente investigación se justifica desde tres perspectivas fundamentales: teórica, práctica y metodológica.

- **Desde una perspectiva teórica**, este estudio contribuye al estado del arte en el análisis de sentimientos aplicado a la salud (Minería de Opiniones en Salud). Al comparar arquitecturas de aprendizaje profundo como LSTM (Memoria a Corto y Largo Plazo) y GRU (Unidades Recurrentes Cerradas), se genera conocimiento sobre cuál arquitectura es más eficiente computacionalmente y precisa para interpretar el lenguaje médico

informal, abordando problemas clásicos como el desvanecimiento del gradiente en textos largos.

- **Desde la perspectiva práctica y social,** el desarrollo de este modelo tiene un impacto directo en la farmacovigilancia. Permitir una valoración automática de la efectividad percibida empodera a los organismos de salud para detectar patrones de ineficacia o reacciones adversas de manera más rápida que los canales oficiales. Para los pacientes, esto se traduce en una futura atención médica más segura y validada por la experiencia colectiva. Además, ayuda a entender la diferencia entre la eficacia clínica es decir dentro de un laboratorio y la efectividad real que es un paciente en su casa.

- **Desde una perspectiva metodológica**, el estudio aplica de manera rigurosamente la metodología CRISP-DM que proviene de sus siglas en inglés Cross-Industry Standard Process for Data Mining, evidenciando la estructuración adecuada de un proyecto de ciencia de datos desde la fase de comprensión del negocio hasta su despliegue final. Este enfoque metodológico no solo fortalece la coherencia del proceso investigativo, sino que además constituye un referente académico para futuros estudiantes de la carrera de Tecnologías de la Información de la Universidad Laica Eloy Alfaro de Manabí, al demostrar la aplicabilidad práctica de la Inteligencia Artificial en la resolución de problemáticas sociales reales.

CAPÍTULO II - MARCO TEÓRICO

2. Antecedentes de investigaciones

La valoración de la percepción de los pacientes ha transitado un largo camino, evolucionando desde sistemas manuales y pasivos hacia enfoques computacionales de alta sofisticación. Comprender esta trayectoria es fundamental para contextualizar la presente investigación y su contribución al estado del arte.

2.1. Farmacovigilancia Digital

Según (Pitti, 2024) esta evolución hacia la farmacovigilancia ha experimentado una transformación significativa en las últimas décadas, evolucionando desde sistemas tradicionales de notificación espontánea hacia enfoques digitales más sofisticados. El ecosistema digital ha reconfigurado el rol de redes como Twitter o Facebook, que ya no funcionan como simples escaparates sociales. Hoy en día, operan de facto como sensores de vigilancia sanitaria en tiempo real. En estos entornos, la narrativa del paciente fluye sin la asepsia propia del protocolo clínico, volcando síntomas y reacciones adversas de manera espontánea. Este fenómeno convierte el aparente ruido de las redes en un flujo de información vivo, cuya minería permite detectar tendencias de seguridad farmacológica con una latencia considerablemente menor que la de los canales oficiales de reporte. Esta transición hacia la denominada “Farmacovigilancia 2.0” introduce ventajas significativas en primera instancia, posibilita la obtención de retroalimentación inmediata, ya que los pacientes no necesitan recurrir a mecanismos formales para reportar inquietudes, permitiendo que profesionales de la salud y organismos reguladores accedan a información oportuna sobre la eficacia y seguridad de los medicamentos. De igual manera se facilita la detección temprana de reacciones adversas, superando las limitaciones de los esquemas tradicionales de vigilancia post comercialización y de los ensayos clínicos, cuyos datos suelen tardar en acumularse.

Además, en este contexto (Van Stekelenborg, y otros, 2019) demostró en su estudio muy significativo demostró que los datos de redes sociales fueron más informativos que los reportes espontáneos de reacciones adversas y la literatura publicada respecto al uso no médico de

sustancias no controladas como antidepresivos, donde los datos habían sido difíciles de obtener mediante métodos tradicionales. La integración de inteligencia artificial con procesamiento de lenguaje natural ha permitido desarrollar sistemas de monitoreo que procesan datos desde la extracción de publicaciones en foros web hasta la generación de indicadores y alertas dentro de entornos visuales e interactivos.

2.1.1. Minería de opiniones en salud

La minería de opiniones en el contexto sanitario ha emergido como un campo de investigación activo con aplicaciones directas en la toma de decisiones clínicas. El análisis de sentimientos de reseñas de medicamentos se ha convertido en una metodología crucial para determinar si el comportamiento del paciente hacia un medicamento, producto o tratamiento es positivo, negativo o neutral utilizando técnicas de procesamiento de lenguaje natural (Panda, Panigrahi, & Pati, 2022).

Uno de los conjuntos de datos más empleados en investigaciones relacionadas con el análisis de sentimientos en el ámbito farmacológico es el UCI ML Drug Review Dataset, desarrollado por la Universidad de California en Irvine. Este recurso compila reseñas emitidas por pacientes sobre medicamentos específicos, acompañadas de la condición médica asociada y una calificación en una escala de diez puntos que refleja su nivel de satisfacción general. Gracias a su riqueza semántica y gran volumen de registros, dicho dataset se ha convertido en un insumo fundamental para múltiples estudios que implementan técnicas de aprendizaje automático y modelos basados en arquitecturas de transformadores orientados a la clasificación y análisis de sentimientos en reseñas médicas (Li, 2018).

Entre los modelos más significativos del estado del arte se destacan principalmente:

Los modelos basados en transformadores, particularmente BERT, han demostrado superar consistentemente los enfoques convencionales, alcanzando precisiones del 96% gracias a su capacidad para interpretar patrones complejos del lenguaje y señales contextuales indetectables por modelos más simples. Un estudio sobre Inhibidores Selectivos de la Recaptación de Serotonina (ISRS) e Inhibidores de la Recaptación de Serotonina-Norepinefrina

(IRSN) utilizados para tratar la depresión encontró que el sentimiento promedio fue significativamente mayor en ISRS comparado con IRSN (0.065 vs. 0.005, $p < 0.001$), y también mayor en reseñas con calificaciones altas que en reseñas con calificaciones bajas (Ahmad & Veeramanickam, 2024).

Y también los modelos de Memoria a Largo-Corto Plazo (LSTM) han alcanzado precisiones superiores al 97% en la clasificación de sentimientos de análisis de medicamentos en categorías como positivo, neutral y negativo, superando enfoques como RNN tradicionales. Adicionalmente, modelos híbridos como BioBERT han demostrado rendimiento superior con una precisión general del 87% en la clasificación de sentimientos de reseñas médicas (Bansal & Bansal, 2024).

2.1.2. Brecha entre Eficacia y Efectividad: Diferencia Teórica Fundamental

La distinción entre eficacia y efectividad constituye un concepto central en la investigación farmacológica y clínica que justifica el análisis de la percepción del paciente como complemento a los ensayos clínicos controlados.

La eficacia se refiere se refiere a los resultados de un tratamiento en condiciones ideales de ensayos clínicos controlados, donde los pacientes son cuidadosamente seleccionados y las condiciones son estrictamente monitoreadas. Los ensayos regulatorios prueban rigurosamente la capacidad de un tratamiento para impactar un resultado conocido, como la frecuencia de convulsiones, medido en estudios bien controlados (Boissel, Cogny, Marko, & Boissel, 2019).

Por contraposición, la efectividad mide el desempeño del fármaco una vez que abandona la burbuja experimental y se enfrenta a la complejidad de la práctica clínica diaria, donde la adherencia suele ser irregular y la biología del paciente resulta impredecible. En este sentido (Wagner, Rieger, & Bedi, 2019) sostienen que la única forma de auditar esta realidad es explotando los datos masivos desde historias clínicas hasta huellas digitales en plataformas de salud. Estos registros capturan lo que podríamos llamar la "verdad sucia" del tratamiento y revelan patrones de seguridad que los ensayos controlados, por su propia naturaleza aséptica, tienden a pasar por alto.

La brecha eficacia-efectividad representa las discrepancias entre los resultados de terapias e intervenciones obtenidos en ensayos clínicos frente a lo que ocurre en la práctica del mundo real, diferencias que surgen debido a variaciones en las poblaciones de pacientes y en los niveles de adherencia. Como señala la literatura científica, "la mayoría de los ensayos clínicos no incluyen individuos de la población a la que se aplicarán los hallazgos de eficacia en el mundo real, por lo que la efectividad de los tratamientos debe 'traducirse' a estas poblaciones" (Reyes, Malik, Sahouri, & Knezevic, 2025).

Esta brecha justifica, precisamente, el estudio de la percepción del paciente sobre la efectividad de los medicamentos. Mientras los ensayos clínicos proporcionan la base científica, es la efectividad la que asegura que los tratamientos entreguen beneficios reales en la práctica cotidiana. Cuando los ensayos de efectividad y eficacia producen resultados discordantes, significa que la terapia funciona en algunas situaciones, pero no en otras, lo que requiere juicio clínico para tomar decisiones de tratamiento adecuadas.

2.2. Metodología de Ciencia de Datos

2.2.1. Metodología CRISP-DM

CRISP-DM significa Proceso Estándar Intersectorial para Minería de Datos. Es un marco ampliamente adoptado que describe los pasos de un proyecto de análisis o ciencia de datos Este modelo establece un marco metodológico estructurado que orienta la definición, planificación y control de cada fase del proceso analítico. La adopción del estándar CRISP-DM obedece a la necesidad de blindar el flujo de trabajo contra la incertidumbre inherente a los proyectos de minería de datos. Más que una simple guía este marco actúa como un mecanismo de control que sincroniza la ejecución técnica con los imperativos de la investigación, mitigando el riesgo de invertir recursos computacionales en modelos que no resuelven el problema de fondo su arquitectura se despliega en un ciclo de seis estadios iterativos desde la disección inicial del problema hasta la validación operativa, imponiendo una disciplina metodológica que garantiza la replicabilidad y la coherencia científica del experimento en cualquier contexto de aplicación. Aunque el proceso se presenta secuencialmente, es iterativo y permite revisión constante entre fases.

Fase 1: Comprensión del negocio: sentar las bases

Esta fase inicial crítica sienta las bases para todo el proyecto de análisis o ciencia de datos. Ya que no se trata solo de comprender los datos, sino de comprender a fondo el porqué del proyecto: el problema o la oportunidad de negocio principal que debe abordarse. Esta fase garantiza que el esfuerzo basado en datos esté alineado con los objetivos estratégicos, evitando el desperdicio de recursos y garantizando resultados viables. Aquí, vamos más allá de la comprensión superficial para definir objetivos precisos y establecer criterios de éxito claros.

Actividades clave:

La definición del alcance exige trascender las declaraciones de intenciones genéricas para operacionalizar el problema en términos estrictamente cuantitativos. En lugar de perseguir mejoras abstractas, el método requiere traducir los objetivos de negocio en metas matemáticas que el algoritmo pueda optimizar como establecer un umbral porcentual específico de mejora en los tiempos de respuesta paralelamente, se ejecuta una auditoría de viabilidad técnica que confronta la disponibilidad de recursos (infraestructura, datos y presupuesto) con las restricciones del entorno, evaluando desde el inicio los límites éticos, regulatorios y de privacidad para asegurar que el proyecto sea ejecutable dentro del marco legal y operativo existente.

Fase 2: Comprensión de los datos: Conozca sus datos

En esta etapa, el enfoque se desplaza desde la planificación estratégica hacia la inspección empírica de la información bruta más que una simple recolección, este proceso implica una exploración estadística profunda diseñada para revelar la topología real del dataset antes de cualquier intervención. El analista debe verificar la calidad intrínseca de los registros, identificando patrones distribucionales, valores anómalos y posibles sesgos de origen; esta validación actúa como un filtro de seguridad indispensable que garantiza que la materia prima posea la integridad suficiente para soportar las arquitecturas de modelado sin introducir errores silenciosos en las predicciones futuras.

Actividades clave:

- Identificar y acceder a todas las fuentes de datos relevantes, tanto internas como externas.
- Utilice técnicas de visualización de datos (por ejemplo, histogramas, gráficos de dispersión, gráficos de caja) para identificar patrones, tendencias y relaciones.
- Aplicar técnicas estadísticas (por ejemplo, análisis de correlación, prueba de hipótesis) para descubrir conocimientos más profundos.
- Identifique valores atípicos y anomalías que puedan requerir mayor investigación.

P3: Preparación de datos: limpieza y transformación de sus datos

La fase de preparación de los datos implica la conversión de información bruta en conjuntos de datos estructurados, coherentes y aptos para el análisis. Aunque representa una de las etapas más demandantes del proceso, su correcta ejecución resulta determinante para la calidad de los resultados obtenidos. Las diversas actividades como la de limpieza, transformación y normalización influyen directamente en la confiabilidad de la información generada, estableciendo una base sólida para el desarrollo de modelos analíticos precisos.

Actividades clave:

El proceso comienza aislando los subconjuntos de datos que son estrictamente relevantes para el análisis, descartando información redundante que pueda introducir ruido. Una vez seleccionada la muestra, se procede a corregir la calidad de los datos: esto implica imputar los valores faltantes mediante métodos estadísticos como la media o la moda, gestionar los valores atípicos que puedan sesgar el modelo y unificar formatos inconsistentes en las etiquetas o unidades de medida. Posteriormente, se realiza la ingeniería de características, creando nuevas variables mediante transformaciones matemáticas y aplicando técnicas de normalización o escalado para homogeneizar los rangos numéricos. Finalmente, se codifican las variables categóricas (por ejemplo, con one-hot encoding) y, de ser necesario, se reduce la dimensionalidad del conjunto de datos para optimizar el procesamiento.

Paso 4: Modelado: construcción del marco predictivo o perspicaz

Esta fase marca la transición de la preparación de datos al núcleo del análisis, donde se utilizan diversas técnicas de modelado para construir marcos predictivos o descriptivos. Se trata de convertir los datos preparados en información significativa o predicciones prácticas. Aquí es donde se aplica la experiencia analítica y algorítmica, seleccionando y aplicando las herramientas adecuadas para alcanzar los objetivos del proyecto.

Actividades clave:

La ejecución de esta etapa inicia seleccionando los algoritmos más adecuados según la naturaleza del problema (clasificación, regresión o agrupamiento), buscando siempre un equilibrio entre la complejidad computacional y la interpretabilidad de los resultados. Se exploran diversas arquitecturas, desde modelos estadísticos tradicionales hasta aprendizaje profundo, pero bajo un diseño de pruebas riguroso establecido previamente. Para asegurar la robustez del modelo, se definen métricas de evaluación específicas como la precisión, recuperación o RMSE, y se aplican técnicas de validación cruzada que previenen el sobreajuste, garantizando que el modelo funcione bien con datos nuevos. Todo este comportamiento se analiza mediante herramientas de visualización para detectar patrones de error antes de avanzar.

Paso 5: Evaluación: Evaluación rigurosa y alineación con los objetivos del negocio

El análisis va más allá de una validación métrica aislada, pues somete al sistema a una auditoría de utilidad operativa. No basta con que el algoritmo maximice su precisión estadística; resulta fundamental verificar que las inferencias generadas resuelvan efectivamente la problemática planteada al inicio del proyecto. Este paso funciona como un filtro de calidad final, diseñado para detectar cualquier disonancia entre un modelo que sea matemáticamente correcto pero que constituya una solución inaplicable en el contexto real. De esta manera, se asegura que el despliegue técnico tenga una traducción directa y funcional para la toma de decisiones.

Actividades clave:

En esta instancia, el análisis se centra en verificar si los resultados obtenidos responden directamente a las preguntas de negocio planteadas al inicio, más allá de la simple precisión técnica. Es fundamental cuantificar el impacto real que tendrán estos hallazgos en los indicadores

clave de desempeño (KPI) de la organización y utilizar visualizaciones claras para comunicar estos descubrimientos a las partes interesadas. Basándose en esta evaluación integral, se toma la decisión estratégica de desplegar el modelo, implementar los hallazgos estadísticos o, en caso de resultados insuficientes, iterar el proceso buscando enfoques alternativos.

Paso 6: Implementación: traducir los conocimientos en resultados prácticos

En esta fase final, la información obtenida del análisis ya sea de un modelo sofisticado o de un estudio estadístico sencillo, se traduce en resultados tangibles para los usuarios finales y la organización. No se trata solo de presentar un informe, sino de integrar la información en los procesos operativos y garantizar que genere un impacto real.

Actividades clave:

Para finalizar, se debe desarrollar una estrategia de implementación técnica que defina cómo se integrará el modelo o el conocimiento generado dentro de los flujos de trabajo y sistemas operativos actuales. Una vez en producción, se establece un sistema de vigilancia continua para monitorear el desempeño de las métricas clave y detectar fenómenos como la deriva de datos (data drift), lo que permite adaptar el modelo ante cambios en el entorno. Este ciclo se cierra recopilando activamente la retroalimentación de los usuarios finales para identificar incidencias prácticas y oportunidades de mejora futura (Sharma, 2025).

2.3. Fundamentos de Procesamiento de Lenguaje Natural

2.3.1. Preprocesamiento de Texto

Antes de poder alimentar datos de texto a un modelo de aprendizaje automático, es necesario limpiar y estandarizar el texto. Este proceso, conocido como preprocesamiento de texto, transforma texto crudo y desordenado en un formato predecible.

Normalización

La normalización es el proceso de reducir variaciones de palabras a una forma base común para ayudar al modelo a reconocer que estas palabras son semánticamente similares. La

conversión a minúsculas es fundamental porque las listas de stopwords típicamente están en minúsculas, y si se intenta hacer coincidir 'The' contra 'the', no funcionará. Por lo tanto, se normaliza primero y luego se filtra.

Limpieza de Ruido

El proceso de limpieza incluye la remoción de elementos HTML, caracteres especiales, URLs y otros elementos que añaden ruido, pero no contribuyen al significado semántico del texto. Esta limpieza reduce el ruido y permite que los modelos se enfoquen en las palabras de contenido que realmente importan.

Eliminación de Stopwords

Los stopwords son palabras extremadamente comunes que añaden poco valor semántico a una oración. Palabras como "the", "a", "is", "in" y "on" frecuentemente se eliminan para reducir el ruido y enfocar la atención del modelo en palabras más significativas. Esta eliminación tiene tres beneficios principales: reduce el ruido permitiendo que los modelos se enfoquen en palabras de contenido significativo, disminuye el tiempo de procesamiento (analizar un millón de documentos puede reducir el tiempo de procesamiento entre 30-40% al remover stopwords), y mejora el rendimiento para tareas como similitud de documentos o extracción de palabras clave (LearnCodePro, 2025).

2.3.2. Tokenización y Vectorización

Dado que los modelos de aprendizaje profundo son matemáticamente incompatibles con el flujo de texto continuo, la tokenización se implementó como una etapa de discretización obligatoria. En el contexto de este diseño, el proceso fue más allá de la simple separación mecánica de palabras: funcionó como un mecanismo de segmentación lógica cuyo propósito es descomponer el corpus clínico en unidades semánticas procesables de este modo, se logra reducir la entropía dimensional de la entrada, facilitando su posterior tratamiento por parte de los modelos.

Operativamente, ante una entrada clínica compleja del tipo “Dolor intenso, pero alivio rápido”, el sistema descompone la cadena en vectores independientes ['Dolor', 'intenso', ',', 'pero', 'alivio', 'rápido', '.']. Esta granularidad resulta crítica para la arquitectura neuronal, pues le permite asignar tensores de peso distintos a conceptos contrapuestos la severidad del síntoma frente a la eficacia terapéutica evitando la contaminación semántica que ocurriría si se procesara la frase como un bloque monolítico.

Siguiendo la línea técnica de (Akram, 2024), esta transformación constituye el prerrequisito para la extracción de características; sin esta conversión preliminar a embeddings cuantificables, el modelo carecería de la estructura numérica necesaria para ejecutar cualquier operación de cálculo matricial o clasificación probabilística.

2.3.3. Word Embeddings

Los word embeddings son representaciones de palabras que transforman palabras en vectores en un espacio de menor dimensionalidad donde las dimensiones representan características semánticas. A diferencia de los modelos tradicionales de bolsa de palabras, donde cada palabra única en el corpus recibe una dimensión única, los word embeddings mapean palabras a un espacio de menor dimensionalidad donde las dimensiones representan características semánticas (Swimm Team, 2023).

Ventajas sobre One-Hot Encoding

El One-Hot Encoding representa cada palabra como un vector disperso donde solo una posición tiene valor 1 y el resto son ceros Este enfoque presenta limitaciones importantes: los vectores generados son de gran tamaño pues su dimensionalidad corresponde directamente al total del vocabulario, no permiten capturar relaciones semánticas entre términos y asignan representaciones completamente distintas a palabras con significados similares, como “feliz” y “alegre”.

En contraste, los word embeddings proporcionan representaciones densas en las que cada palabra se ubica como un punto dentro de un espacio multidimensional, permitiendo que términos con

significados afines se agrupen de manera natural en dicho espacio en esta nueva representación numérica, "feliz" podría estar cerca de "alegre", mientras que "manzana" estaría lejos de "perro".

Captura de Significado Semántico

La idea central de los embeddings sigue el principio de que el contexto equivale al significado de representar cada palabra por los contextos en los que aparece. Si se puede codificar las palabras circundantes en un vector, se obtiene una representación numérica que refleja el uso y, por tanto, el significado.

Los modelos como Word2Vec utilizan dos arquitecturas la CBOW, siglas de Continuous Bag of Words, que predice una palabra objetivo basándose en su contexto, y Skip-gram que hace lo opuesto, mejorando la capacidad del modelo para entender contextos y relaciones de palabras. Por ejemplo, la técnica CBOW predice la palabra "perro" si las palabras circundantes incluyen "animal" o "mascota".

Una demostración notable de la capacidad semántica de los embeddings es la aritmética de analogías: el modelo aprende aspectos de relaciones como género en su geometría vectorial, permitiendo operaciones como "rey - hombre + mujer ≈ reina". Esto muestra cómo los word embeddings pueden entender no solo significados, sino también relaciones entre palabras, casi como razonamiento con el lenguaje (Singh, 2025).

2.4. Ingeniería de Características

La agrupación de medicamentos o condiciones en categorías clínicas es una forma poderosa de ingeniería de características basada en el dominio esto nos permite reducir la dimensionalidad del ruido, identificar patrones transversales y facilitar el análisis estratificado y la interpretación de resultados.

2.4.1. Taxonomías y Categorización Médica

La capacidad de estructurar e interpretar datos médicos contextuales es crucial para proporcionar atención al paciente personalizada y eficiente. Las taxonomías médicas y la

categorización de enfermedades en categorías clínicas (como Salud Mental, Dolor, etc.) son fundamentales para el análisis estructurado de datos de salud (Msheik, Mcheick, Hariri, Adda, & Dbouk, 2025).

Los estudios de PLN aplicado a notas clínicas han utilizado consistentemente categorías de enfermedades basadas en la Clasificación Internacional de Enfermedades para estructurar el análisis. Esta categorización permite:

- **Análisis comparativo**: Facilita la comparación de patrones de sentimiento y efectividad percibida entre diferentes categorías de condiciones médicas.
- **Identificación de patrones específicos**: Diferentes categorías de enfermedades pueden tener características únicas en las reseñas de pacientes.
- **Mejora de la generalización del modelo**: Agrupar condiciones relacionadas puede mejorar la capacidad del modelo para generalizar patrones aprendidos.

La investigación ha demostrado que las condiciones neurológicas como enfermedades del sistema circulatorio contienen mucha más información no estructurada y, consecuentemente, han visto un enfoque más fuerte de PLN, mientras que registros médicos para enfermedades metabólicas típicamente contienen muchos más datos estructurados (Sheikhalishahi, y otros, 2019).

2.4.2. Análisis Temporal en Salud: Estacionalidad y Patrones en Adherencia y Efectividad

El análisis de patrones temporales constituye un componente esencial en la evaluación de la adherencia y efectividad de tratamientos médicos. Las diversas investigaciones han identificado que el tiempo influye significativamente en algunas intervenciones mientras que no tiene influencia en otras.

Patrones Temporales de Adherencia a Medicamentos

Los estudios de modelado de trayectorias realizados por (Nadkarni, Ohno-Machado, & Chapman, 2011) han identificado de manera consistente tres patrones principales de adherencia para diferentes tratamientos: "adherentes excelentes", "no adherentes tardíos" y "no adherentes tempranos". Estos patrones resultan más informativos que las métricas tradicionales de adherencia promedio, y esto se debe a que un mismo valor promedio puede representar comportamientos clínicamente muy distintos. Por ejemplo, podría tratarse de una adherencia adecuada en las primeras etapas seguida de una disminución progresiva, o bien de una adherencia moderada que se mantiene constante durante todo el período de análisis.

En cuanto a la trayectoria correspondiente a los "adherentes excelentes", esta evidenció un porcentaje significativamente mayor de días de abstinencia y un menor número de días de consumo excesivo en comparación con las demás trayectorias identificadas. Asimismo, el análisis de importancia de características indicó que las variables históricas constituyen los factores predominantes aportando el 94% de la relevancia total, mientras que el “día de la semana” y la “frecuencia de medicación” emergieron como los siguientes predictores más influyentes.

Estacionalidad en Salud Mental y Física

La investigación de (Ford, Li, Zhao, & Mokdad, 2009) ha evidenciado la existencia de patrones estacionales significativos tanto en la salud física como en la salud mental. En particular, la salud física tiende a deteriorarse durante el invierno y a mejorar en el verano, observándose una diferencia promedio de 0.63 días no saludables entre los meses de enero y julio. Por otro lado, la salud mental presenta variaciones estacionales de menor magnitud, con los valores más desfavorables durante la primavera y el otoño, alcanzando una diferencia aproximada de 0.23 días no saludables entre los períodos extremos del año.

Efectividad de Intervenciones a lo Largo del Tiempo

Las intervenciones que muestran éxito en seguimientos de menos de 10 meses varían a través del tiempo. Las intervenciones técnicas mostraron efectos significativos comparados con el cuidado estándar en el período de 4-6 meses, mientras que las intervenciones actitudinales

fueron más efectivas en seguimientos de 7-9 meses. Las intervenciones multicomponente mostraron los resultados más prometedores en el mantenimiento de la adherencia a medicamentos a largo plazo (Jiang, y otros, 2017).

2.5. Redes Neuronales Artificiales (Deep Learning)

2.5.1. Redes Neuronales Recurrentes (RNN)

Las redes neuronales recurrentes (RNN) son una clase de modelos de aprendizaje profundo que han sido desarrolladas para manejar datos secuenciales, como series temporales, texto o señales de voz. A diferencia de las redes neuronales tradicionales, las RNN incorporan un estado de memoria que permite utilizar entradas anteriores para informar predicciones futuras, lo que las hace especialmente útiles en tareas donde el contexto histórico es relevante (Amazon Web Services, 2026).

2.5.2. El Problema del Desvanecimiento del Gradiente

El problema del desvanecimiento del gradiente surge directamente de la mecánica del pase hacia atrás durante el entrenamiento mediante Backpropagation Through Time (BPTT), particularmente debido a la aplicación repetida de la regla de la cadena sobre muchos pasos temporales.

Explicación Matemática

El gradiente de la pérdida L con respecto al estado oculto en un paso temporal temprano k, denotado $\frac{\partial L}{\partial h_k}$, depende del gradiente en un paso temporal posterior t. La conexión entre ambos se establece a través de la regla de la cadena, involucrando las derivadas de las transiciones del estado oculto entre los pasos *k* y *t*.

Cada Jacobiano intermedio $\frac{\partial h_i}{\partial h_{i-1}}$ depende de la matriz de pesos recurrentes W_{hh} y la derivada de la función de activación usada en la celda RNN. Para una RNN simple con función de activación f, la actualización del estado oculto es:

$$h_i = f(W_{hh}h_{i-1} + W_{xh}x_i + b_h)$$

El problema surge porque las funciones de activación como tanh o sigmoid, que eran comúnmente usadas en las RNN tempranas, tienen derivadas que son estrictamente menores a 1 para la mayoría de las entradas. Para ser más precisos, la derivada de tanh siempre es ≤1, mientras que la derivada de sigmoid siempre es ≤0.25.

Consecuencia del Desvanecimiento Exponencial

Cuando estas pequeñas derivadas se multiplican repetidamente durante el pase hacia atrás a través de muchos pasos temporales, los Jacobianos se reducen exponencialmente:

- $0.9^1 = 0.9$
- $0.9^{10} \approx 0.35$
- $0.9^{50} \approx 0.005$
- $0.9^{100} \approx 0.000027$

La consecuencia práctica es severa: la señal del gradiente que se origina de errores en pasos temporales tardíos se vuelve increíblemente pequeña, o "desvanece", para cuando alcanza las capas de red correspondientes a pasos temporales tempranos. Estos gradientes cercanos a cero significan que los pesos responsables de procesar la secuencia en esos pasos tempranos reciben casi ninguna señal de actualización basada en resultados a largo plazo (ApX Machine Learning, 2025).

Esencialmente, la red se vuelve incapaz de aprender correlaciones entre eventos separados por intervalos largos en la secuencia, justificando por qué no se utiliza una RNN simple para análisis de texto de reseñas de medicamentos.

2.5.3. Arquitectura LSTM

La arquitectura LSTM (Long Short-Term Memory), conocida en español como Memoria a Largo y Corto Plazo, constituye una variante avanzada de las redes neuronales recurrentes (RNN) diseñada para capturar dependencias a largo plazo en datos secuenciales. Se introdujo originalmente en 1997 como solución al problema del desvanecimiento de gradientes que afecta a las RNN estándar.

Componentes Principales

La arquitectura LSTM organiza sus neuronas en bloques de memoria que contienen los siguientes elementos:

1. Celda de Memoria (Memory Cell)

La celda de memoria es el corazón de la arquitectura LSTM. Se caracteriza por una conexión recurrente consigo misma con peso 1.0, creando un "carrusel de error constante" que permite que los gradientes fluyan sin desvanecerse ni explotar. Esta celda almacena información a largo plazo sin restricciones de rango, a diferencia de los estados ocultos convencionales que están acotados entre -1 y 1.

2. Las Tres Compuertas (Gates)

Las compuertas son mecanismos multiplicativos que controlan el flujo de información hacia y desde la celda de memoria. Utilizan funciones sigmoides que producen valores entre 0 y 1:

- **Compuerta de Entrada**: Controla qué nueva información se almacena en la celda. Valores cercanos a 1 permiten que la información fluya, mientras que valores cercanos a 0 la bloquean.

- **Compuerta de Olvido**: Decide qué información antigua se descarta de la celda anterior. Fue añadida posteriormente a la formulación original de LSTM para mejorar el rendimiento.

- **Compuerta de Salida**: Controla qué partes del estado de la celda se transmiten como salida de la red (Blalock, Gonzalez Ortiz, Frankle, & Guttag, 2020).

2.5.4. Arquitectura GRU

La arquitectura GRU (Gated Recurrent Unit o Unidad Recurrente Gestionada) constituye una variante simplificada de las redes LSTM, diseñada para disminuir la complejidad computacional sin comprometer su capacidad para modelar y aprender dependencias de largo alcance presentes en secuencias de datos.

Fue propuesta por Cho et al. en 2014 como una alternativa más eficiente al LSTM.

Componentes Principales

A diferencia del LSTM que utiliza tres compuertas, la arquitectura GRU incorpora solo dos compuertas principales que controlan el flujo de información:

1. Compuerta de Olvido/Reinicio (Reset Gate r_t)

La compuerta de reinicio controla en qué medida la información del estado oculto anterior se combina con la entrada actual. Determina qué partes del estado anterior se deben ignorar para calcular el nuevo contenido del estado. Utiliza una función sigmoide que produce valores entre 0 y 1:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

2. Compuerta de Actualización/Retención (Update Gate z_t)

La compuerta de actualización es conceptualmente similar a la compuerta de entrada del LSTM. Decide cuánta información del estado anterior debe mantenerse y cuánta debe reemplazarse con el nuevo contenido del candidato. También produce valores entre 0 y 1:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

3. Candidato a Estado Oculto ($\tilde{h}_t$)

El candidato al nuevo estado oculto es una función tanh que combina la entrada actual con el estado anterior filtrado por la compuerta de reinicio. Esta activación no lineal genera la información que podría reemplazar el estado existente:

$$\tilde{h}_t = \tanh\left(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h\right)$$

Aquí, la compuerta de reinicio r_t multiplica elemento a elemento con el estado anterior, lo que permite que el modelo controle selectivamente qué información histórica es relevante para el cálculo actual.

Actualización del Estado Oculto

El estado oculto final se obtiene mediante una combinación lineal del candidato y el estado anterior, controlada por la compuerta de actualización:

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1}$$

Esta ecuación muestra cómo la compuerta de actualización realiza una interpolación: cuando $z_t \approx 0$, el nuevo estado adopta principalmente el candidato $\tilde{h}_t$; cuando $z_t \approx 1$, el estado retiene casi completamente su valor anterior h_{t-1} (Jordan, Sokół, & Park, 2021).

2.6. Dinámica del Entrenamiento y Optimización

2.6.1. Función de Pérdida

La Categorical Cross-Entropy es ampliamente utilizada como función de pérdida para medir qué tan bien un modelo predice la clase correcta en problemas de clasificación multiclase. Mide la diferencia entre la distribución de probabilidad predicha y las etiquetas verdaderas codificadas en one-hot, guiando al modelo a asignar probabilidades más altas a la clase correcta.

Características Principales

- Se utiliza cuando hay más de dos clases
- Trabaja con salidas softmax donde las probabilidades suman 1

- Mayor pérdida significa que la predicción está lejos de la clase verdadera; menor pérdida significa que el modelo está funcionando bien

- Comúnmente utilizada en clasificación de imágenes, clasificación de texto y tareas de reconocimiento de voz

Fórmula Matemática

La Categorical Cross-Entropy mide la diferencia entre las etiquetas verdaderas y las probabilidades predichas de un modelo. Penaliza al modelo cuando asigna baja confianza a la clase correcta:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

Donde:

- $L(y, \hat{y})$: Pérdida de Categorical Cross-Entropy

- y_i: Etiqueta verdadera para clase i (codificación one-hot)

- $\hat{y}_i$: Probabilidad predicha para clase i

- C: Número de clases (GeeksforGeeks, 2025).

2.6.2. Optimizador Adam

El Optimizador Adam (Adaptive Moment Estimation) revoluciona los métodos tradicionales de descenso de gradiente introduciendo un mecanismo sofisticado que adapta las tasas de aprendizaje basándose en el primer y segundo momento de los gradientes.

Funcionamiento del Descenso de Gradiente Adaptativo

A diferencia del Descenso de Gradiente Estocástico (SGD) estándar, que emplea una tasa de aprendizaje fija para todos los parámetros, Adam ajusta dinámicamente la tasa de aprendizaje

para cada parámetro individualmente. Esta adaptabilidad se logra manteniendo promedios exponencialmente decrecientes de gradientes pasados y sus cuadrados, permitiendo a Adam navegar paisajes de pérdida complejos con notable eficiencia (LunarTech, 2025).

2.6.3. Sobreajuste y Generalización

El sobreajuste ocurre cuando un modelo se vuelve demasiado especializado en los datos de entrenamiento, perdiendo la capacidad de generalizar a datos nuevos no vistos. En contraste, la generalización es la capacidad del modelo para entender y aplicar patrones aprendidos a datos no vistos.

Memorización vs. Aprendizaje

El sobreajuste puede resultar en alta precisión del modelo en datos de entrenamiento, pero baja precisión en datos nuevos debido a memorización en lugar de generalización. El concepto es análogo a memorizar respuestas para un examen en lugar de entender los conceptos necesarios para obtener las respuestas: si el examen difiere de lo estudiado, habrá dificultades para responder las preguntas.

Indicadores de Sobreajuste

- La precisión de entrenamiento continúa mejorando mientras la precisión de validación se estanca o disminuye.
- La pérdida de entrenamiento disminuye mientras la pérdida de validación aumenta.
- Alta varianza entre rendimiento de entrenamiento y prueba.

Estrategias de Mitigación

Existen varias técnicas para evitar el sobreajuste, incluyendo el uso de regularización L1 (que fomenta dispersión reduciendo algunos coeficientes a cero) y L2 (que reduce el tamaño de todos los coeficientes para hacer el modelo más simple y generalizable). Otras técnicas incluyen

data agumentation (expandir artificialmente los datos de entrenamiento), simplificación del modelo reduciendo el número de parámetros o capas, y early stopping (Mucci, 2024).

2.6.4. Parada Temprana

La Parada Temprana es una técnica de regularización que monitorea el rendimiento del modelo en un conjunto de validación durante el entrenamiento y detiene el proceso cuando el rendimiento en este conjunto deja de mejorar o comienza a empeorar.

Justificación Teórica

Detenerse muy temprano puede resultar en subajuste, mientras que continuar por demasiadas épocas frecuentemente lleva a sobreajuste. La idea central del *early stopping* es detenerse cuando el modelo ha aprendido las características más relevantes, evitando tanto el sobreentrenamiento como el subentrenamiento.

Mecanismo de Funcionamiento

Durante el ciclo de entrenamiento, después de cada época, se evalúa el rendimiento del modelo no solo en los datos de entrenamiento sino también en el conjunto de validación. Los componentes importantes incluyen:

- **Métrica de Monitoreo**: Elegir una métrica para rastrear en el conjunto de validación (pérdida de validación, precisión de validación).
- **Paciencia (Patience)**: Definir cuántas épocas esperar antes de detener si no hay mejora.
- **Criterio de parada**: Detener el entrenamiento si la métrica de validación monitoreada no ha mejorado durante el número especificado de épocas de paciencia.
- **Restauración de mejores pesos**: Después de que el entrenamiento se detiene, descartar los pesos del modelo de la época final y cargar los pesos guardados que correspondían al mejor rendimiento de validación observado durante el entrenamiento (ApX Machine Learning, 2025).

Ejemplo Práctico

Experimentos han demostrado que, sin configurar paciencia adecuada, el entrenamiento puede detenerse prematuramente (por ejemplo, en época 200 cuando el óptimo está alrededor de época 800). Configurando una paciencia de 200 épocas, el entrenamiento continúa permitiendo encontrar mejoras adicionales, resultando en mejor rendimiento en el conjunto de prueba que sin early stopping.

La razón principal para monitorear pérdida de validación en lugar de precisión para early stopping es que la precisión es una medida gruesa del rendimiento del modelo durante el entrenamiento, mientras que la pérdida proporciona más matiz cuando se usa early stopping con problemas de clasificación (Brownlee, 2020).

2.7. Métricas de Evaluación de Modelos

2.7.1. Matriz de Confusión

Una matriz de confusión es una tabla que resume el rendimiento de un modelo de clasificación comparando sus etiquetas predichas con las etiquetas verdaderas. Es una herramienta de evaluación de rendimiento fundamental que permite un análisis más detallado que simplemente observar la proporción de clasificaciones correctas.

Componentes de la Matriz de Confusión

La matriz de confusión se divide en cuatro cuadrantes:

- **Verdaderos Positivos (TP)**: Instancias donde el modelo predice correctamente la clase positiva cuando realmente es positiva.
- **Verdaderos Negativos (TN)**: Instancias donde el modelo predice correctamente la clase negativa cuando realmente es negativa.
- **Falsos Positivos (FP)**: Instancias donde el modelo predice incorrectamente la clase positiva cuando debería ser negativa (Error Tipo I).

- **Falsos Negativos (FN)**: Instancias donde el modelo predice incorrectamente la clase negativa cuando debería ser positiva (Error Tipo II).

Errores Tipo I y Tipo II

Los errores Tipo I (Falsos Positivos) afectan la precisión del modelo, que mide la exactitud de las predicciones positivas. Los errores Tipo II (Falsos Negativos) afectan el recall del modelo.

Ejemplo Médico: En una prueba diagnóstica para detectar una enfermedad:

- Error Tipo I (Falso Positivo): La prueba predice que el paciente tiene la enfermedad cuando en realidad está sano.
- Error Tipo II (Falso Negativo): La prueba predice que el paciente está sano cuando en realidad tiene la enfermedad (GeeksforGeeks, 2026).

2.7.2. Accuracy (Exactitud) vs. F1-Score

El accuracy (exactitud) mide la proporción de predicciones correctas totales de todas las predicciones hechas por el modelo. Sin embargo, el accuracy produce resultados engañosos si el conjunto de datos está desbalanceado, es decir, cuando los números de observaciones en diferentes clases varían considerablemente.

F1-Score como alternativa superior

El F1-Score es la media armónica de precisión y recall. Varía de 0 a 1, siendo 1 perfecto. Es especialmente útil para conjuntos de datos desbalanceados, que son comunes en análisis de sentimientos:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

El F1-Score ayuda a equilibrar precisión y recall, siendo particularmente útil cuando se trabaja con clases desbalanceadas. Proporciona una visión balanceada de la precisión de clasificación tanto positiva como negativa en un solo número (Focal, 2024).

Tabla 1

Comparativa de Accuracy VS F1 Score

Métrica	Ventajas	Desventajas
Accuracy	Fácil de entender	Puede engañar con datos desbalanceados
F1-Score	Equilibra precisión y recall	Más complejo de explicar

Nota: Explicación breve de las ventajas y desventajas de las métricas.

2.8. Principales tipos de medicamentos empleados en el Dataset

El conjunto de datos utilizado en esta investigación se caracteriza por una amplia diversidad farmacológica, lo que permite analizar el sentimiento de los pacientes en distintos contextos clínicos y terapéuticos. En total, el dataset está compuesto por 3436 medicamentos únicos, asociados a 161297 reseñas provenientes de usuarios, distribuidas entre 18 categorías médicas diferentes. Esta heterogeneidad constituye un elemento clave a la hora de evaluar la capacidad de generalización de los modelos de análisis de sentimientos aplicados.

La dispersión de los datos exhibe una asimetría notable: mientras un núcleo reducido de fármacos concentra la mayor densidad de registros, prevalece una extensa "cola larga" de tratamientos con escasa retroalimentación. Es importante señalar que este desbalance no representa un defecto de la muestra, sino más bien un indicador de fidelidad estadística que replica las dinámicas reales del entorno sanitario, donde coexisten prescripciones masivas junto a tratamientos de nicho con menor visibilidad.

2.8.1. Medicamentos con mayor frecuencia de reseñas

Al segmentar el corpus por volumen de actividad, la categoría Reproductive/Hormonal se erige como el eje dominante de la distribución. La carga principal de información recae sobre agentes específicos como el Levonorgestrel y el Etonogestrel (incluyendo sus variantes con etinilestradiol), así como en sistemas de liberación prolongada tipo Nexplanon, Mirena e Implanon, los cuales aglutinan la mayor densidad de reportes en el estudio. El perfil de sentimiento en este grupo tiende a ser polarizado; si bien se reconocen los beneficios clínicos, existe una carga significativa de reseñas negativas vinculadas a efectos secundarios hormonales.

Este escenario contrasta con el de la categoría Endocrine/Metabolic, donde medicamentos como la Phentermine alcanzan una aceptación superior al 80%, impulsada por la percepción de eficacia en la pérdida de peso. En el ámbito de la salud mental, antidepresivos comunes como Sertraline, Escitalopram, Citalopram y Bupropion mantienen niveles altos de aprobación, aunque no están exentos de críticas sobre la variabilidad en la respuesta terapéutica. Por el contrario, agentes como el Miconazole presentan una acumulación notable de experiencias negativas, lo que los convierte en casos de estudio relevantes para detectar patrones de baja tolerancia o ineficacia percibida.

2.8.2. Distribución de medicamentos por categoría médica

Desde una perspectiva estructural, el conjunto de datos presenta una concentración marcada en dos áreas principales: Reproductive/Hormonal y Mental Health, las cuales agrupan conjuntamente más del 40% del total de las observaciones. Al profundizar en la granularidad del corpus, se observa que la carga informativa recae desproporcionadamente sobre estos dos ejes terapéuticos. Ambas secciones actúan como los pilares del entrenamiento sumando 277 y 319 fármacos respectivamente, con un caudal conjunto superior a las 65.000 opiniones y comparten un perfil de satisfacción similar, superando los 6 puntos de valoración media. Fuera de este núcleo dominante, el dataset incorpora la terminología específica de áreas como Pain Management o Infectious

Disease, aunque el verdadero reto de ingeniería lo plantea la categoría Other/Uncategorized esta etiqueta, que mezcla más de 1.500 tratamientos dispersos y de clasificación difusa, inyecta un grado de entropía necesario para probar la robustez del modelo frente a datos con alta varianza y baja frecuencia.

2.8.3. Popularidad de los medicamentos y representatividad del corpus

El análisis de densidad de datos revela una asimetría pronunciada típica de los entornos biomédicos reales: apenas un 1,9% de los medicamentos se clasifica como "muy popular" (superando las 500 reseñas), mientras que la vasta mayoría cerca del 60% cuenta con menos de 10 opiniones registradas. Esta distribución de "larga cola" implica un reto metodológico considerable, pues exige que los algoritmos de análisis de sentimientos sean capaces de aprender patrones lingüísticos tanto en contextos de abundancia de datos como en escenarios de escasez de información, poniendo a prueba su capacidad de aprendizaje o generalización.

2.8.3. Distribución general de sentimientos

A nivel general, el dataset presenta un predominio de reseñas con sentimiento positivo (60.4%), seguido de opiniones negativas (24.8%) y neutrales (14.8%). Esta distribución refleja una tendencia moderadamente optimista por parte de los usuarios al evaluar sus tratamientos, aunque mantiene una proporción significativa de experiencias negativas, fundamentales para el análisis en contextos de farmacovigilancia y evaluación de seguridad.

En conjunto, la diversidad de medicamentos, categorías clínicas y expresiones emocionales convierte a este dataset en una fuente robusta y representativa para el estudio del análisis de sentimientos en el ámbito médico, proporcionando un marco adecuado para evaluar el desempeño de modelos basados en redes neuronales recurrentes como LSTM y GRU.

2.9. Conclusiones relacionadas al Marco Teórico

Tras el análisis exhaustivo de la literatura científica, los fundamentos conceptuales y el estado del arte referente al Procesamiento de Lenguaje Natural (PLN) aplicado a la salud, se derivan las siguientes conclusiones teóricas que sustentan la presente investigación:

- **Insuficiencia de los métodos tradicionales:** El marco teórico valida la inviabilidad técnica de los modelos basados en frecuencia pura (como Bag of Words o lexicones rígidos) para operar en entornos de alta densidad semántica. La limitación crítica de estos enfoques reside en su ceguera secuencial: al atomizar el texto y tratarlo como un conjunto desordenado de tokens, terminan destruyendo la estructura sintáctica original. En la narrativa clínica, donde la posición exacta de un adverbio de negación puede definir el éxito o el fracaso terapéutico, esta pérdida de linealidad resulta inaceptable. Esto legitima la transición hacia el Aprendizaje Profundo como el único paradigma capaz de preservar la integridad contextual del mensaje.

- **Idoneidad de las Arquitecturas Recurrentes (RNN):** Desde la perspectiva de la arquitectura de software, se determina que las Redes Neuronales Recurrentes constituyen la topología más adecuada para procesar datos secuenciales como el texto narrativo. En particular, las arquitecturas LSTM y GRU se consideran opciones teóricamente apropiadas para mitigar el problema del desvanecimiento del gradiente, ya que permiten conservar información relevante a lo largo de secuencias extensas, algo que las redes neuronales recurrentes tradicionales no logran de manera efectiva.

- **Equivalencia funcional y eficiencia computacional:** En cuanto a la comparación entre ambas arquitecturas, la literatura señala que, aunque LSTM ha sido históricamente dominante en tareas de modelado secuencial, la variante GRU ha surgido como una alternativa más eficiente. Esto se debe a que simplifica su diseño mediante la unificación de las compuertas de "entrada" y "olvido", lo cual reduce la complejidad computacional sin afectar significativamente el rendimiento. Esta característica justifica la necesidad de realizar un análisis comparativo experimental dentro de la presente investigación, con el

fin de determinar si dicha eficiencia teórica se traduce en ventajas prácticas en el contexto de la farmacovigilancia basada en texto.

- **Valor de la información no estructurada (RWE):** Finalmente, el marco teórico resalta la relevancia de la evidencia generada en entornos reales. Las reseñas de pacientes publicadas en plataformas digitales constituyen una fuente de datos esencial para la denominada Farmacovigilancia 2.0, dado que complementan la información procedente de ensayos clínicos tradicionales y permiten ampliar la comprensión del impacto real de los medicamentos desde la perspectiva del usuario.

CAPÍTULO III - MARCO INVESTIGATIVO

3. Introducción

En el presente capitulo se describe el enfoque metodológico contemplado en el desarrollo del análisis de sentimientos aplicado a las review medicas sobre medicamentos utilizando técnicas de procesamiento de lenguaje natural y modelos de aprendizaje profundo, dicha metodología constituye un componente esencial de esta investigación ya que establece el marco sistemático mediante el cual se estructuraron los datos, se entrenan los modelos y también como se evalúa los resultados obtenidos de los entrenamientos.

El principal objetivo de este capítulo es detallar y justificar tanto los métodos, técnicas y herramientas empleadas a lo largo de este proceso de investigación, para todo esto se expone el tipo de investigación utilizada, la metodología seleccionada para el análisis de datos y también las consideraciones estimadas con las limitaciones del estudio y los aspectos éticos relacionados al uso de información de carácter médico.

3.1. Tipo de investigación

La presente investigación se fundamenta en un enfoque cuantitativo, experimental y de corte transversal esto porque se basa en un análisis numérico de grandes volúmenes de datos textuales y también en la experimentación controlada con distintos modelos de redes neuronales profundas.

- Desde el punto de vistas experimental: este estudio implica la manipulación de manera deliberada de la variable independiente que está representada por las arquitecturas de los modelos de aprendizaje profundo LSTM y GRU esto con el fin de visualizar su impacto sobre las variables dependientes como lo son la precisión, la perdida y el F1-Score estas variables en la tarea de clasificación de sentimientos, este diseño se implementó con el objetivo de comparar de manera objetiva el desempeño de ambas arquitecturas bajo condiciones controladas y lo más importante que se utilizó el mismo dataset.

- Desde el diseño transversal se justifica debido a que los datos analizados corresponden a un periodo de tiempo específico y con esto no se considera una evolución temporal de las reseñas, sino que buscan capturar una fotografía general de la percepción de los pacientes sobre la efectividad de los medicamentos en el dataset.

  También se podría considerar como una investigación aplicada dado que no solo persigue la compresión teórica del problema, sino que también se busca general modelos predictivos que sean funcionales con el objetivo de que en el futuro puedan ser utilizados como una base sólida para un sistema de apoyo a la toma de decisiones en contexto relacionas a la salud.

3.2. Metodología CRISP-DM

El desarrollo metodológico de la investigación se rigió por el estándar CRISP-DM por sus siglas en inglés Cross-Industry Standard Process for Data Mining el cual permitió estructurar el proyecto en seis fases iterativas, desde la comprensión de los datos médicos hasta el despliegue de los modelos predictivos, sin alterar el enfoque experimental del estudio.

3.2.1. Relación con los objetivos del proyecto

- **Objetivo 1: Diseñar un proceso de transformación de datos que estandarice las condiciones médicas**. Este objetivo se articula con las fases de comprensión y preparación de los datos, en las cuales se examina la estructura del dataset y se aplican procesos de normalización y categorización de las condiciones médicas.

- **Objetivo 2: Implementar y comparar el rendimiento de dos arquitecturas de redes neuronales (LSTM y GRU).** Se vincula directamente con la fase de modelado, etapa en la cual se diseñan, entrenan y comparan ambas arquitecturas bajo un mismo conjunto de condiciones experimentales, garantizando así una evaluación objetiva y controlada de su desempeño.

- **Objetivo 3: Evaluar la capacidad del modelo en la tarea de clasificación de la percepción de los pacientes.** Este objetivo somete a los algoritmos a una prueba de estrés estadística, donde se priorizan métricas de sensibilidad crítica como la Función de Pérdida (Loss) y el F1-Score por encima de la simple exactitud global. La intención es cuantificar la solvencia técnica de la red, midiendo su habilidad real para traducir la subjetividad del lenguaje natural en una señal clínica estructurada, sin caer en sesgos de memorización.

- **Objetivo 4: Construir reflexiones cualitativas respecto de casos ambiguos encontrados.** Este hito impone una validación que va más allá de lo aritmético, centrándose en la auditoría forense de las predicciones. La estrategia consiste en analizar manualmente la lógica de decisión del modelo ante textos contradictorios, con el propósito de identificar qué patrones lingüísticos específicos actúan como "puntos ciegos" para el algoritmo. De esta forma, se busca revelar limitaciones de interpretación que suelen permanecer ocultas en los reportes porcentuales.

3.3. Limitaciones del Estudio

A pesar de que se fue muy exigente en la metodología aplicada se presentaron diversas limitaciones que se deben tener en cuenta al momento de la interpretación de resultados dichas limitaciones se detallan a continuación:

- Como primera limitación se tiene que el análisis principal se basa en un dataset de carácter público, por lo cual la veracidad y la calidad de las reseñas depende netamente de la información proporcionada por los usuarios la cual puede contener grandes sesgos subjetivos.

- Como segunda limitación se tiene que el proceso de la categorización de condiciones medicas se realizó mediante el uso de diversas inteligencias artificiales lo cual no sustituye completamente la validación clínica de un especialista y esto podría causar ciertas impresiones semánticas en caso específicos.

- Y como ultima limitación importante es que el carácter transversal del estudio nos impide analizar cambios en la percepción de los pacientes a lo mediad de que transcurre los años.

3.4. Ética y Consideraciones de Privacidad

La investigación se desarrolló respetando principios éticos relacionados con el uso responsable de los datos especialmente en el ámbito de la salud el dataset utilizado es de acceso público y se encuentra debidamente anonimizado garantizando así que no se maneje información que permita identificar a los pacientes.

Los datos fueron utilizados exclusivamente con fines académicos y de investigación, y los resultados se presentan de manera agregada evitando interpretaciones que puedan derivar en conclusiones sesgadas o perjudiciales para determinados grupos o condiciones médicas

CAPÍTULO IV - MARCO PROPOSITIVO

4. Introducción

Este capítulo tiene como principal finalidad describir la propuesta técnica y metodológica que se desarrolló en esta investigación orientada principalmente hacia el análisis de sentimientos de review medicas mediante modelos de aprendizaje profundo, en este apartado se aplica el enfoque metodológico previamente descripto en el capítulo anterior también se describe los recursos necesarios y la debida aplicación de la metodología CRISP-DM dicha estructura nos describe de manera sistemática todo el proceso de análisis de análisis de datos y modelado predictivo.

Este marco propositivo consolida las decisiones técnicas que se han tomado a lo largo de la investigación que se realizó con el objetivo de poder presentar de forma ordenada y coherente la solución que se aplicó tanto como su viabilidad y en un futuro su posible aplicación en contextos médicos reales relacionados con el análisis de la percepción de pacientes sobre la efectividad de medicamentos.

4.1. Descripción de la Propuesta

La propuesta consiste en el desarrollo de un sistema de análisis de sentimientos capaz de clasificar la percepción de los pacientes respecto de la efectividad de medicamentos, a partir de reseñas textuales. El despliegue técnico se estructura como una cadena de procesamiento continua que supervisa la integridad del dato desde la limpieza preanalítica hasta la validación del modelo.

Dentro de este marco experimental, se someten a prueba dos arquitecturas secuenciales rivales las redes LSTM y su evolución GRU con la intención de cuantificar empíricamente cuál optimiza la separación de las clases (negativo, neutral o positivo) en un desafío de clasificación supervisada. Finalmente, el modelo seleccionado fue integrado en una interfaz gráfica interactiva desarrollada en Streamlit, permitiendo demostrar su aplicabilidad práctica como una prueba de concepto funcional.

4.2.Determinación de Recursos

Para el correcto desarrollo de la propuesta fue necesario considerar distintos tipos de recursos, tanto humanos como tecnológicos y económicos, los cuales permitieron llevar a cabo el proceso de análisis, modelado y evaluación de manera eficiente.

4.2.1. Recursos Humanos

Los recursos humanos que están involucrados en este proyecto son los siguientes:

1. **Tutor Académico:**

- Ing. Jorge Ivan Pincay Ponce, PhD.
- **Responsabilidades**
    - Supervisión del Proyecto.
    - Asesoramiento técnico y académico
    - Revisión de avances.
    - Validación de resultados

2. Desarrollador:

- Chávez Moreira Elvis Joel

- **Responsabilidades**
    - Realizar la desconstrucción del dataset
    - Implementar el pipeline de procesamiento de texto
    - Diseño y entrenamiento de los modelos neuronales
    - Evaluar los resultados de los modelos

- Documentar el proceso completo

4.2.2. Recursos Tecnológicos

El desarrollo de todo se realizó utilizando un entorno hibrido tanto como de manera local como utilizando la nube en este caso el uso de Kaggle Notebook con el objetivo principal de optimizar el tiempo de los entrenamientos de ambos modelos.

- Entorno local: se utilizó el Visual Studio Code para la ejecución de scripts no tan demandantes como el de categorizar las condiciones médicas.

- Entorno en la nube: este entorno se utilizó principalmente para los entrenamientos de los modelos, para ello se hizo uso del Notebook de Kaggle aprovechando la aceleración por hardware mediante GPUs NVIDIA Tesla P100 lo cual fue de mucha ayuda al momento de procesar millones de parámetros que usan los modelos.

- Librerías utilizadas: las librerías de Python versión 3 que se utilizaron para el entrenamiento de modelos son tensorFlow/Keras (v2.10+) para Deep Learning, NLTK para procesamiento lingüístico, y Scikit-learn para métricas de evaluación

4.2.3. Recursos Económicos

Desde el punto de vista económico la investigación presentó un costo reducid ya que se utilizaron herramientas y plataformas open source, así como el conjunto de datos empleado es de carácter público y las librerías utilizadas también son open source.

El uso de Kaggle Notebooks permitió acceder a recursos computacionales de alto rendimiento sin incurrir en costos adicionales, haciendo viable el desarrollo del proyecto sin necesidad de inversión económica significativa en infraestructura especializada.

Tabla 2

Tabla de presupuesto estimado para el desarrollo del proyecto.

Recurso	Descripción	Costo unitario	Costo total estimado
Horas de investigación	200 horas de desarrollo (programación y tesis) a $10/h	$10	$2000
Servicios básicos	Internet y servicios eléctrico (10 meses)	$40	$400
Total			$2400

Nota. Esta tabla muestra el presupuesto estimado para el desarrollo del proyecto.

4.3. Aplicación de la Metodología CRISP-DM

La metodología CRISP-DM fue aplicada como marco organizativo para estructurar el desarrollo de la propuesta cabe aclarar que cada una de sus fases fue adaptada al contexto del análisis de sentimientos en datos médicos esta metodología permitió mantener un flujo de trabajo ordenado, facilitar la trazabilidad entre los objetivos de investigación y las decisiones técnicas adoptadas y garantizar coherencia entre las distintas etapas de la investigación.

4.4. Fases de la Metodología CRISP-DM

4.4.1. Comprensión del Negocio (Business Understanding)

4.4.1.1.Objetivos del Negocio y Minería de Datos

En esta fase se definió el problema central del estudio, orientado a comprender la percepción de los pacientes sobre la efectividad de los medicamentos a partir de reseñas textuales. Se estableció como objetivo principal evaluar la capacidad de modelos neuronales recurrentes para clasificar dichas percepciones en categorías de sentimiento, aportando una herramienta de apoyo para el análisis de grandes volúmenes de opiniones médicas.

4.4.1.2.Definición de Requisitos y Criterios de Aceptación

Para alinear el desarrollo del modelo con las necesidades de farmacovigilancia, se establecieron los siguientes requisitos técnicos y funcionales que el sistema debe cumplir para ser considerado viable**.**

Tabla 3

Requisitos Funcionales y Criterios de Aceptación

ID	Requisito Funcional	Descripción	Criterio de Aceptación
RF-01	Ingesta de texto libre	El sistema debe permitir la entrada de reseñas médicas en formato de texto no estructurado (inglés).	El modelo acepta párrafos de longitud variable sin truncar información relevante hasta 200 tokens.
RF-02	Clasificación de Sentimiento	El algoritmo debe clasificar la reseña en una de tres categorías: Negativo, Neutral o Positivo.	La clasificación resultante coincide con la etiqueta real en al menos el 80% de los casos (Accuracy > 0.80).
RF-03	Análisis Comparativo	El sistema debe permitir al usuario seleccionar entre los modelos LSTM y GRU para realizar la inferencia.	El dashboard presenta un selector visible que cambia el modelo backend en menos de 2 segundos.
RF-04	Visualización de Métricas	El sistema debe mostrar gráficamente la confianza de la predicción y las métricas de rendimiento del modelo.	Se visualizan tarjetas con F1-Score, Precisión y Matriz de Confusión en la interfaz de usuario.

Nota. Los requisitos funcionales fueron definidos con base en las necesidades de procesamiento de lenguaje natural identificadas en la fase de comprensión del negocio y las limitaciones del dataset de origen**.**

Tabla 4

Requisitos No Funcionales

ID	Requisito No Funcional	Descripción	Criterio de Aceptación
RNF-01	Eficiencia de Entrenamiento	El entrenamiento del modelo no debe exceder las 10 horas en un entorno de GPU estándar (P100).	El tiempo total de entrenamiento registrado es inferior a 36,000 segundos.
RNF-02	Latencia de Inferencia	El tiempo de respuesta para clasificar una nueva reseña debe ser imperceptible para el usuario final.	La predicción se genera en menos de 500 milisegundos por reseña individual.
RNF-03	Escalabilidad del Dataset	El pipeline de datos debe ser capaz de procesar el dataset completo sin desbordamiento de memoria.	El proceso de tokenización maneja las 161k reseñas sin errores de "Out of Memory".

Nota. Los criterios de rendimiento y escalabilidad están calibrados específicamente para el entorno de hardware utilizado (GPU Tesla P100) y el volumen total de registros del repositorio UCI Machine Learning.

4.4.2. Comprensión de los Datos (Data Understanding)

En esta investigación se utilizó un dataset público de Kaggle con el nombre de UCI ML Drug Review que fue utilizado en el hackathon del Club de la Universidad Kaggle en invierno del 2018.

- Población total del dataset: el dataset consta de 161,297 reseñas de pacientes, que van acompañados de diversas variables tales como nombre del medicamento, condición médica, fecha y calificación.

- Muestreo: para el tema del muestreo se utilizó un muestreo aleatorio esto con el fin de garantizar la representatividad de las clases para ellos se utilizó el famoso 20 80 que se describe a continuación:

    - El 80% de las review totales que son 129,037 se utilizaron como entrenamiento de los datos es decir en temas más técnicos para un correcto ajuste de pesos sinápticos.

    - El 20% restante que son un total de 32,260 review se utilizaron para el tema de validaciones de dicho entrenamiento.

4.4.3. Preparación de los Datos (Data Preparation)

4.4.3.1. Fase 1: Ingeniería de Datos y Desconstrucción

El primer paso que se realiza antes de comenzar con cualquier entrenamiento de datos es el de realizar una intervención estructural en el dataset original para corregir su granularidad, mejorando su calidad y capacidad de análisis para ello se hace lo siguiente:

- **Desconstrucción de datos:** Mediante sugerencias de mi tutor del trabajo de titulación me dijo que viera si varios tratamientos y condiciones estaban agrupados en una sola fila lo que cual me iba a dificultad el análisis a nivel individual. Para corregir eso se realizó un hizo un script en Python utilizando diversas librerías de este mismo como Numpy y Panda estas dos las más utilizadas en este script para hacer una correcta desconstrucción del dataset este script generó una fila independiente para cada par único de medicamento-condición.

    Por ejemplo, si un paciente reportaba el uso de 5 fármacos este script creó 5 registros distintos así replicando las demás columnas como la reseña y asegurando que el modelo pudiera asociar el sentimiento y la experiencia a cada medicamento individualmente.

    Todo este proceso se realizó de la siguiente manera;

- **Separación de medicamentos y condiciones:** Se desglosaron las filas que contenían múltiples tratamientos y condiciones gracias a esto se generó un nuevo registro para cada combinación única de medicamento y condición.

- **Replicación de reseñas:** La reseña del paciente se replicó para cada medicamento garantizando así que el análisis de sentimiento pudiera asociarse correctamente a cada fármaco individual.

- **Manejo de múltiples condiciones:** En un caso hipotético de que un paciente reportara más de una condición para un mismo medicamento se realizaba la generación de registros independientes para cada combinación medicamento-condición.

- **Ingeniería de Características**

  - **Categorización Médica:** Se normalizó el texto de la columna condition y se implementó un sistema de categorización automática basado en palabras clave, agrupando diagnósticos clínicos dispersos en macro-categorías médicas, por ejemplo, Cardiovascular, Mental Health, Pain Management, entre otras.
    El diccionario de categorías y términos asociados fue construido con el apoyo de diversos sistemas de inteligencia artificial las cuales permitieron identificar sinónimos variantes léxicas y expresiones clínicas frecuentes de manera eficiente y consistente se adoptó esta estrategia como una alternativa escalable y económicamente viable frente a la validación manual por especialistas médicos la cual resultaba inviable debido al alto número de registros del dataset con esta forma se logró un equilibrio entre precisión semántica y viabilidad operativa reduciendo significativamente la variabilidad léxica y facilitando tanto el análisis estadístico como el entrenamiento de modelos supervisados sin la necesidad de introducir costos prohibitivos asociados a la revisión clínica individual.

- **Análisis temporal:** Gracias a la variable de fecha asociada a cada reseña se pudieron derivar diversas características temporales, entre ellas el año, mes, día, día de la semana, trimestre y semana del año también un indicador de fin de semana y el número de días transcurridos desde el inicio del dataset.

  Estas transformaciones permitieron pasar de una única variable temporal a un conjunto de atributos que capturan mejor el contexto en el que fue emitida cada opinión la inclusión de estas variables tuvo como objetivo analizar si existían patrones temporales o comportamientos estacionales en las valoraciones de los tratamientos así como posibles cambios en la percepción de los usuarios a lo largo del tiempo además si este tipo de información temporal puede ser aprovechada por los modelos de aprendizaje automático para incorporar contexto adicional que no está presente únicamente en el texto de la reseña.
- **Ponderación por utilidad:** Con el fin de incorporar una medida de relevancia social se utilizó la variable usefulCount la cual indica cuántos usuarios consideraron una reseña como útil dentro del dataset a partir de esta variable se generaron diversas columnas adicionales, como categorías cualitativas, percentiles y una versión normalizada, así como una transformación logarítmica para reducir la asimetría de la distribución de esta manera no solo se tuvo en cuenta el contenido de la reseña sino también el nivel de validación por parte de la comunidad permitiendo así diferenciar opiniones aisladas de aquellas que tuvieron mayor impacto esta ponderación aporta información complementaria al modelo y contribuye a una representación más rica de los datos durante el proceso de entrenamiento.

4.4.3.2. Fase 2: Preprocesamiento de Lenguaje Natural

Una vez estructurados y enriquecidos los datos en la fase de ingeniería, se procedió a aplicar un pipeline de preprocesamiento de lenguaje natural (PLN) con el

objetivo principal es el de preparar el texto de las reseñas para su posterior transformación en representaciones numéricas. Este proceso se implementó mediante scripts en Python, centralizados en una función de limpieza cuyo nombre predecible es el de (limpiar_texto) está diseñada para reducir el ruido semántico y mejorar la calidad de la información textual utilizada por los modelos neuronales.

- El primer paso que se realizó fue una limpieza de artefactos de codificación, eliminando secuencias residuales como _x000D_ y otras anomalías derivadas de errores comunes como los de encoding como lo pueden ser las etiquetas HTML presentes en el texto original dichas inconsistencias si no se tratan adecuadamente, pueden introducir tokens irrelevantes que afectan negativamente el aprendizaje del modelo.

- Como segundo paso a seguir en esta fase, se realizó un proceso de normalización del texto, que consistió en la conversión estricta de todo el texto a minúsculas y también la eliminación de caracteres no alfabéticos, como números y signos de puntuación, con todos estos pasos descritos anteriormente se logró unificar la representación de las palabras y reducir la dimensionalidad efectiva del vocabulario evitando que variantes superficiales de un mismo término fueran interpretadas como entidades distintas.

- Y por último paso se llevó a cabo un filtrado de stopwords utilizando el corpus en inglés provisto por la librería NLTK con eta librería se eliminaron palabras vacías como artículos, preposiciones y conectores frecuentes las cuales aportan poca o nula información semántica para la tarea de análisis de sentimientos.

De este modo todo el texto resultante conserva principalmente términos con mayor carga informativa con el objetivo de facilitar una representación más eficiente en las etapas de tokenización y vectorización.

4.4.3.3. Fase 3: Tokenización y Vectorización

Con el texto previamente limpio y normalizado, se procedió a su tokenización y vectorización, con el objetivo de transformar las reseñas en tensores numéricos compatibles con las redes neuronales recurrentes implementadas. Para este proceso se construyó un vocabulario restringido a las 10.000 palabras más frecuentes del corpus, lo que permitió capturar la mayor parte de la información lingüística relevante sin provocar un crecimiento excesivo del espacio de características.

Más que una simple reducción de dimensionalidad, este filtro de frecuencia opera como un mecanismo de supresión de ruido: descarta los términos de la "cola larga" estadística para evitar que el modelo desperdicie recursos intentando memorizar palabras raras que poco aportan a la generalización. Por otro lado, la variabilidad de longitud en los textos se neutralizó imponiendo una normalización geométrica fija de 200 tokens. Esta estandarización resulta innegociable para la arquitectura: al truncar los excesos y rellenar los déficits con ceros (post-padding), se obliga a que datos irregulares encajen en una estructura tensorial rígida, garantizando así que la GPU reciba matrices matemáticamente consistentes en cada ciclo.

Este procedimiento garantizó una matriz de entrada uniforme, condición necesaria para el entrenamiento eficiente de modelos neuronales basados en mini lotes.

4.4.4. Modelado (Modeling)

4.4.4.1. Configuración experimental y justificación de hiperparámetros

Previo al diseño de las arquitecturas, se establecieron los hiperparámetros de entrenamiento basándose en la naturaleza del dataset y las capacidades del hardware disponible (GPU Tesla P100). Esta configuración busca garantizar la reproducibilidad y la eficiencia computacional.

- **Longitud de Secuencia:** Se estableció este límite tras analizar el histograma de longitud de las reseñas. Se observó que el 95% de la información semántica relevante se encuentra

dentro de las primeras 180 palabras. Fijar el límite en 200 garantiza que el modelo capture el contexto completo sin desperdiciar memoria en *padding* innecesario.

- **Tamaño del Lote:** A diferencia de configuraciones estándar (32 o 64), se optó por un lote masivo de 2048 muestras. Esta decisión técnica se justifica por la gran dimensionalidad del dataset (161k registros); un lote grande estabiliza la estimación del gradiente y acelera drásticamente el tiempo por época sin degradar la precisión final.
- **Épocas:** Se definió un ciclo máximo de 300 épocas para permitir una convergencia profunda del modelo, controlado siempre por mecanismos de parada temprana para evitar el sobreajuste.

4.4.4.2.Arquitectura de los modelos neuronales

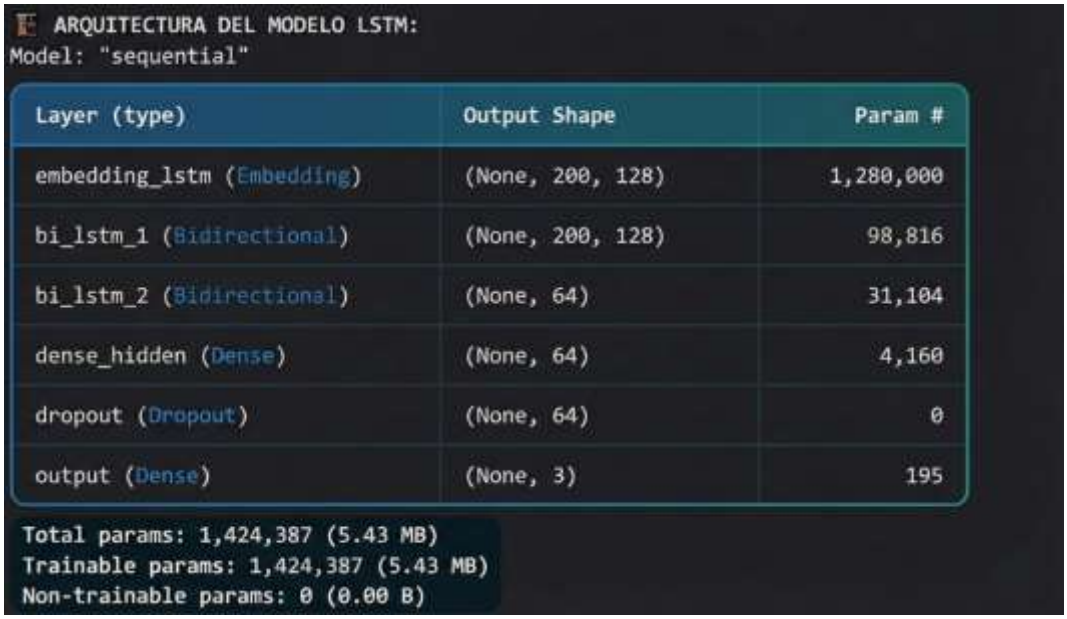
Con las entradas textuales transformadas en representaciones numéricas, se diseñaron e implementaron dos arquitecturas neuronales recurrentes bidireccionales con el objetivo de comparar su desempeño en la tarea de clasificación de sentimientos ambas arquitecturas tienen una estructura general similar solo diferenciándose principalmente en el tipo de celda recurrente que utiliza lo que permite un análisis comparativo controlado

Arquitectura LSTM (Long Short-Term Memory)

El modelo basado en LSTM bidireccional fue diseñado para capturar dependencias de largo plazo presentes en las reseñas médicas, las cuales suelen contener información relevante distribuida a lo largo de todo el texto. La arquitectura inicia con una capa de Embedding de dimensión 128, encargada de proyectar cada token en un espacio vectorial denso donde se preservan relaciones semánticas entre palabras.

Posteriormente, se interpuso un mecanismo denominado como olvido estratégico mediante una capa de Dropout calibrada al 0.3, cuya función es romper las sinergias espurias entre neuronas antes de que la información ingrese al bucle recurrente. La profundidad del análisis recae en una configuración apilada de doble nivel LSTM bidireccional: el primer estrato opera con 64 unidades manteniendo activa la secuencia temporal completa, lo que permite transferir la totalidad del historial cronológico a un segundo bloque de 32 unidades encargado de condensar la abstracción contextual. Previo

a la fase de clasificación densa, se insertó una barrera adicional de regularización con un Dropout más agresivo del 0.5, reforzando la capacidad de generalización del modelo frente al ruido estadístico. Finalmente, una capa totalmente conectada de 32 neuronas con activación ReLU precede a la capa de salida Softmax de 3 neuronas, correspondiente a la clasificación multiclase del sentimiento. El modelo resultante cuenta con un total de 1,424,387 parámetros que se puedan entrenar.


ARQUITECTURA DEL MODELO LSTM:
Model: "sequential"

Layer (type)	Output Shape	Param #
embedding_lstm (Embedding)	(None, 200, 128)	1,280,000
bi_lstm_1 (Bidirectional)	(None, 200, 128)	98,816
bi_lstm_2 (Bidirectional)	(None, 64)	31,104
dense_hidden (Dense)	(None, 64)	4,160
dropout (Dropout)	(None, 64)	0
output (Dense)	(None, 3)	195

Total params: 1,424,387 (5.43 MB)
Trainable params: 1,424,387 (5.43 MB)
Non-trainable params: 0 (0.00 B)

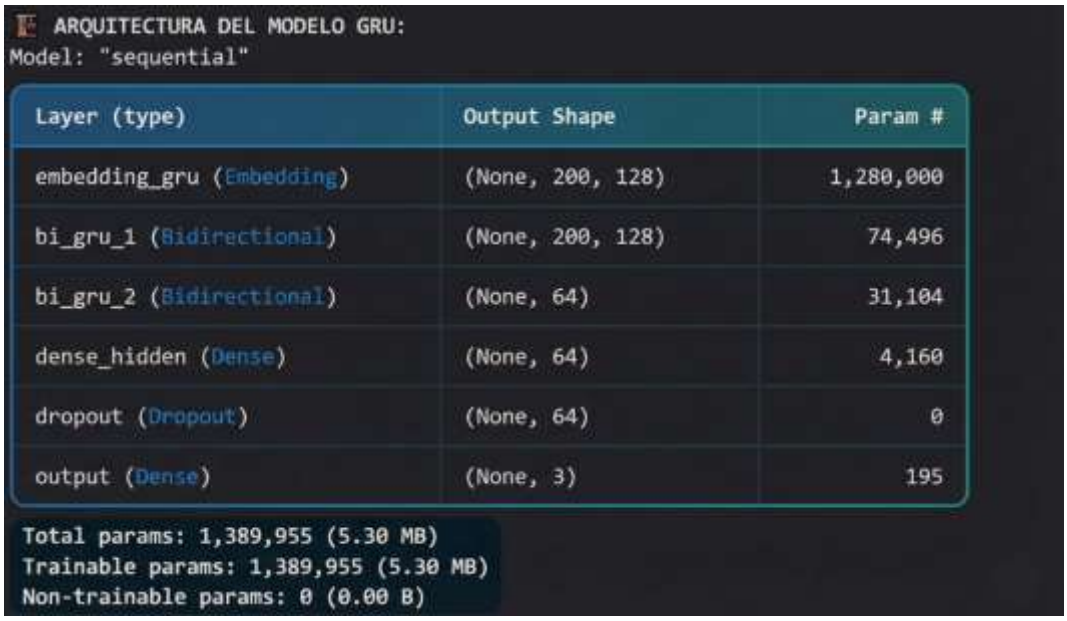
Ilustración 1 Arquitectura LSTM

Arquitectura GRU (Gated Recurrent Unit)

El segundo modelo implementado utiliza celdas GRU bidireccionales, manteniendo una arquitectura prácticamente equivalente a la del modelo LSTM, con el fin de garantizar una comparación justa. La principal diferencia radica en el uso de GRU en lugar de LSTM, lo que reduce la complejidad interna de las celdas recurrentes.

Esta arquitectura conserva la secuencia de capas Embedding → Dropout → GRU bidireccional (64) → GRU bidireccional (32) → Dropout → Dense → Salida, pero presenta un menor número total de parámetros, concretamente 1,389,955, aproximadamente 34,000 menos que el modelo LSTM. Esta reducción teórica de

complejidad permite evaluar si una arquitectura más ligera puede ofrecer un desempeño comparable, con potenciales beneficios en tiempo de entrenamiento y eficiencia computacional.



ARQUITECTURA DEL MODELO GRU:

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding_gru (Embedding)	(None, 200, 128)	1,280,000
bi_gru_1 (Bidirectional)	(None, 200, 128)	74,496
bi_gru_2 (Bidirectional)	(None, 64)	31,104
dense_hidden (Dense)	(None, 64)	4,160
dropout (Dropout)	(None, 64)	0
output (Dense)	(None, 3)	195

Total params: 1,389,955 (5.30 MB)
Trainable params: 1,389,955 (5.30 MB)
Non-trainable params: 0 (0.00 B)

Ilustración 2 Arquitectura GRU

4.4.4.3.Estrategia de Entrenamiento y Validación

Para asegurar la validez interna del estudio, ambos modelos se sometieron a una estrategia de optimización unificada.

Se seleccionó el algoritmo Adam como optimizador debido a su eficiencia en el manejo de gradientes dispersos y su capacidad de adaptación dinámica de la tasa de aprendizaje.

La función de costo seleccionada fue la Categorical Cross-Entropy, métrica estándar para penalizar las divergencias probabilísticas en tareas de clasificación multiclase.

Finalmente, la integridad del modelo frente al sobreajuste (overfitting) se gestionó mediante un esquema de seguridad activo basado en Callbacks:

- **Early Stopping**: Se configuró una interrupción con una paciencia de 10 épocas, deteniendo el entrenamiento si la función de pérdida en validación no presenta mejoras.

- **Model Checkpoint**: Se implementó un mecanismo de guardado selectivo que almacena únicamente los pesos del modelo que lograron el mínimo error de validación histórico, descartando la última época si esta presenta signos de degradación.

4.4.5. Evaluación (Evaluation)

La evaluación del desempeño de los modelos no se limitó exclusivamente a métricas automáticas, sino que se diseñó un protocolo de evaluación mixto, combinando métricas cuantitativas estándar con un proceso de validación cualitativa asistida por intervención humana (human-in-the-loop). Este enfoque permitió analizar no solo el rendimiento estadístico de los modelos, sino también su comportamiento frente a casos reales de lenguaje natural, donde suelen aparecer ambigüedades semánticas difíciles de capturar mediante métricas agregadas.

- Métricas Cuantitativas

  En una primera instancia, se sometió a ambas arquitecturas a un escrutinio técnico sobre el conjunto de prueba completamente aislado. El análisis se fundamentó en el cálculo de la exactitud (accuracy) global y la función de pérdida, complementado por un reporte de clasificación desglosado que detalla la precisión, la sensibilidad (recall) y la puntuación F1 para las tres clases de sentimiento: negativo, neutral y positivo.

  Si bien estos indicadores proporcionaron una base objetiva para contrastar el rendimiento promedio entre LSTM y GRU, se asume metodológicamente que las métricas globales poseen limitaciones para reflejar la interpretación de matices lingüísticos complejos en textos subjetivos y extensos como las reseñas clínicas.

- Validación Cualitativa (Human-in-the-loop)

  La validación humana no se planteó como una revisión aleatoria convencional, sino como una auditoría sistemática apoyada en scripts de muestreo. Para evitar que las clases mayoritarias distorsionaran la percepción de

calidad, se extrajo un subconjunto estratificado de 1.000 instancias que replica fielmente la distribución estadística original del dataset. Cada punto de dato fue reconstruido integrando no solo la triada básica (texto, etiqueta real y predicción), sino añadiendo la métrica de incertidumbre del modelo; esto fue vital para entender con qué seguridad se equivocaba la red.

El análisis se focalizó en diseccionar los fallos del modelo de mejor rendimiento, buscando patrones lingüísticos sutiles como ironías o dobles negaciones que suelen escapar a las métricas globales. El ciclo se cerró exportando la evidencia a registros planos (CSV), lo que transforma una revisión subjetiva en un artefacto de investigación reproducible y auditable por terceros, garantizando la transparencia del hallazgo clínico.

4.4.6. Despliegue (Deployment)

De conformidad con la fase final del ciclo CRISP-DM, la etapa de despliegue se concibió no como una puesta en producción masiva, sino como un mecanismo de transferencia de conocimiento. El objetivo primordial fue garantizar que los hallazgos del estudio fueran accesibles, interpretables y verificables tanto por pares académicos como por usuarios finales sin perfil técnico. Para materializar este propósito, se desarrolló una interfaz gráfica interactiva tipo dashboard analítico, la cual funge como herramienta de validación terminal del sistema de clasificación de sentimientos.

A nivel de implementación, la integración de Streamlit obedeció a un criterio de eficiencia radical: se descartó el desarrollo web convencional para eliminar la deuda técnica asociada al mantenimiento de un frontend complejo. Esta herramienta no funciona como una simple capa de presentación, sino más bien como una consola de depuración en caliente.

Esta arquitectura permite sustituir los reportes estáticos por un entorno de ejecución dinámico, lo que obliga a los modelos a exponer sus inferencias en tiempo real y permite al operador realizar una comparación inmediata entre la respuesta de la GRU y

la de la LSTM. Esta interactividad convierte el análisis de errores filtrando por patología o tipo de sentimiento en una auditoría forense accesible, desmitificando el funcionamiento de la 'caja negra' neuronal ante el usuario final.

4.5.6.1. Arquitectura del sistema

El sistema fue implementado utilizando el framework Streamlit, seleccionado por su ligereza, facilidad de despliegue local y capacidad para integrar modelos de aprendizaje profundo con visualizaciones interactivas.

La arquitectura del aplicativo sigue un enfoque modular, separando claramente:

- **Back-end analítico**, responsable de la carga de modelos entrenados (LSTM y GRU), procesamiento de predicciones y cálculo de métricas.
- **Front-end interactivo**, encargado de la visualización de resultados y la interacción con el usuario.

El sistema permite la gestión dinámica de los modelos, cargando los pesos pre-entrenados que se guardaron en modelo_lstm.h5 y modelo_gru.h5 facilitando así la conmutación entre arquitecturas para comparar predicciones sobre una misma reseña en tiempo real.

4.5.6.2. Módulos funcionales implementados

El dashboard analítico esta desarrollado en Streamlit su estructura es un conjunto de módulos funcionales organizados por páginas, cada una orientada a un propósito específico dentro del proceso de validación, interpretación y auditoría de los modelos de análisis de sentimientos. Esta organización modular permite una navegación intuitiva y una separación clara de responsabilidades analíticas, alineándose con los principios de trazabilidad y explicabilidad del modelo.

Módulo de inicio y métricas globales

Este módulo actúa como punto de entrada al sistema y presenta una síntesis del dataset y del desempeño general de los modelos entrenados. A través de indicadores agregados, se proporciona información sobre el número total de reseñas, medicamentos analizados, categorías médicas y la distribución global de sentimientos. Esta ofrece una visión panorámica de lo que permite contextualizar los resultados antes de profundizar en un análisis más detallado y también facilita una comprensión inicial del alcance del estudio.

Ilustración 3 Módulo de inicio y métricas globales

Módulo de comparación de modelos

Este componente del panel funciona como un árbitro técnico: ejecuta el 'cara a cara' entre las arquitecturas para ir más allá del porcentaje de acierto y revelar el coste real del desempeño. La visualización no es decorativa, sino probatoria; los gráficos exponen la ventaja táctica de la red GRU, demostrando cómo logra comprimir la función de pérdida y aligerar la carga paramétrica. El dato crítico es el tiempo: el sistema valida visualmente el ahorro de casi una hora de procesamiento, un argumento de eficiencia

física que justifica, por sí solo, la elección de la GRU como la arquitectura estándar para el entorno productivo.

Ilustración 4 Módulo de comparación de modelos

Módulo de visualización de curvas de entrenamiento

Este módulo está dedicado a la auditoría forense del proceso de aprendizaje, trazando la evolución histórica de la exactitud y la función de pérdida a lo largo de las 300 épocas programadas. A través de una navegación segmentada por arquitectura, se despliegan las trayectorias de convergencia tanto en su ciclo completo como en el punto óptimo de corte (checkpoint). Dicha visualización resulta fundamental para diagnosticar la estabilidad estocástica del entrenamiento, permitiendo al analista identificar

visualmente el momento exacto de estabilización y descartar fenómenos de sobreajuste (overfitting) en las fases terminales del ajuste.

Ilustración 5 Módulo de visualización de curvas de entrenamiento

Módulo de Análisis Detallado

El sistema incorpora un módulo de análisis profundo orientado a la evaluación explicativa del modelo seleccionado. En este apartado se incluyen matrices de confusión y métricas específicas por clase de sentimiento (Negativo, Neutral y Positivo), permitiendo identificar patrones de acierto, errores sistemáticos y posibles sesgos de

predicción.

Ilustración 6 Matriz de confusión y métricas por clase

También este mismo módulo incluye un análisis segmentado por categoría médica, lo que posibilita observar cómo varía el desempeño del modelo en distintos contextos clínicos. Esta funcionalidad amplía el alcance del análisis más allá de métricas globales, aportando una visión más fina y contextualizada del comportamiento del

clasificador.

Ilustración 7 Análisis por categoría médica

Módulo de validación manual (Auditoría Human-in-the-loop)

Implementa la parte de auditoría human-in-the-loop este módulo carga el archivo CSV con validación consolidada y permite filtrar por aciertos/errores y por sentimiento real. Muestra tablas con reseñas, indica si la predicción fue correcta y ofrece un selector para leer el texto completo de cualquier review. Incluye también métricas filtradas y un gráfico de barras que compara correctos e incorrectos por sentimiento esto es

especialmente útil para inspeccionar casos difíciles y entender dónde falla el modelo en la práctica.

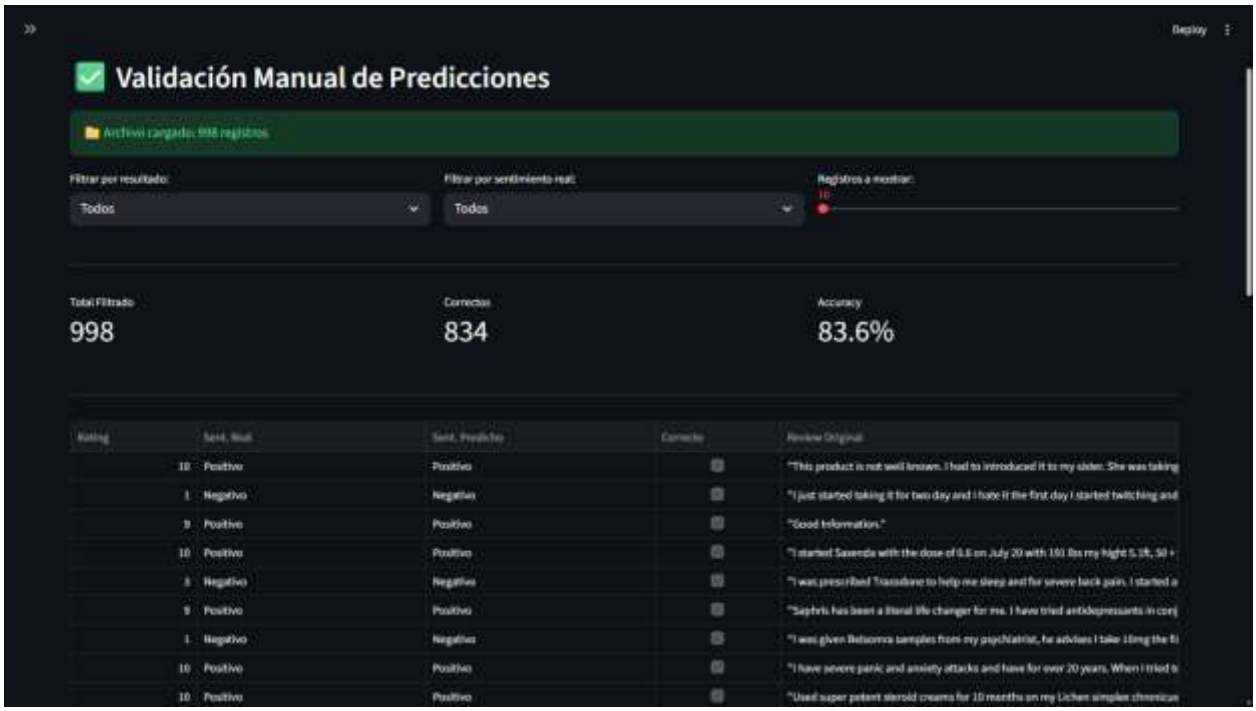


Ilustración 8 Módulo de validación manual

4.5.6.3. Pruebas de ejecución y latencia

Las pruebas de ejecución realizadas en entorno local es decir en el editor de código Visual Studio Code, la aplicación se ejecutó mediante el siguiente comando streamlit run app.py demostraron una alta eficiencia computacional, con tiempos de carga inferiores a cinco segundos para los modelos y datasets de validación.

La interacción con los filtros y visualizaciones es inmediata, validando la viabilidad de utilizar interfaces ligeras para la presentación de resultados complejos en proyectos de ciencia de datos aplicados al ámbito médico.

4.5.6.4. Alcance del despliegue

El despliegue desarrollado cumple un rol fundamental como:

- Herramienta de validación final
- Medio de comunicación de resultados
- Soporte para auditoría humana y análisis explicativo

En el contexto de esta investigación, el despliegue se orienta a apoyar la toma de decisiones académicas y técnicas, dejando abierta la posibilidad de futuras extensiones hacia entornos productivos o clínicos.

CAPÍTULO V - EVALUACIÓN DE RESULTADOS

5. Análisis del entrenamiento y convergencia

5.1. Análisis del comportamiento del entrenamiento

Este proceso de entrenamiento de las arquitecturas neuronales como lo son LSTM y GRU se ejecutó en un entorno controlado de hasta 300 épocas utilizando la plataforma Kaggle Notebook con la ayuda de su aceleración mediante el GPU NVIDIA Tesla P100 con la ayuda de esta aceleración se optimizar el tiempo de entrenamiento de los modelos. Durante el entrenamiento se monitorearon de forma simultánea la función de pérdida categórica (Categorical Cross-Entropy) y la métrica de exactitud (Accuracy), tanto en el conjunto de entrenamiento como en el conjunto de validación.

Mediante un análisis de manera visual a las curvas de aprendizaje se evidenció un comportamiento típico de modelos profundos entrenados sobre texto natural en este caso reseñas, se evidenció una mejora sostenida de la exactitud en los entrenamientos a su vez acompañada de una rápida estabilización y una posterior degradación del desempeño en lo que es la validación.

5.1.1. Comportamiento de las curvas de aprendizaje

El análisis de las curvas de aprendizaje en el entrenamiento de los modelos LSTM y GRU permite evaluar de forma detallada la dinámica del proceso de entrenamiento, la capacidad de generalización y la presencia de fenómenos como el sobreajuste, para ello se examinaron las gráficas de perdida (loss) y precisión (accuracy) tanto en el conjunto de entrenamiento como en el conjunto de validación esto considerando dos escenarios: las primeras épocas optimas y el entrenamiento completo.

Análisis del modelo LSTM

En las curvas correspondientes a las primeras nueve épocas del modelo LSTM se observa una disminución progresiva de la función de pérdida en el conjunto de entrenamiento, acompañada de un incremento sostenido de la precisión este comportamiento indica que la red neuronal logra capturar patrones relevantes del texto desde las primeras iteraciones.

Sin embargo, al analizar la pérdida de validación, se evidencia que a partir de la época 9 el modelo alcanza su punto óptimo de generalización. A partir de este punto, aunque la pérdida de entrenamiento continúa descendiendo, la pérdida de validación comienza a incrementarse de forma sostenida, tal como se observa en las curvas del entrenamiento completo hasta 300 épocas.

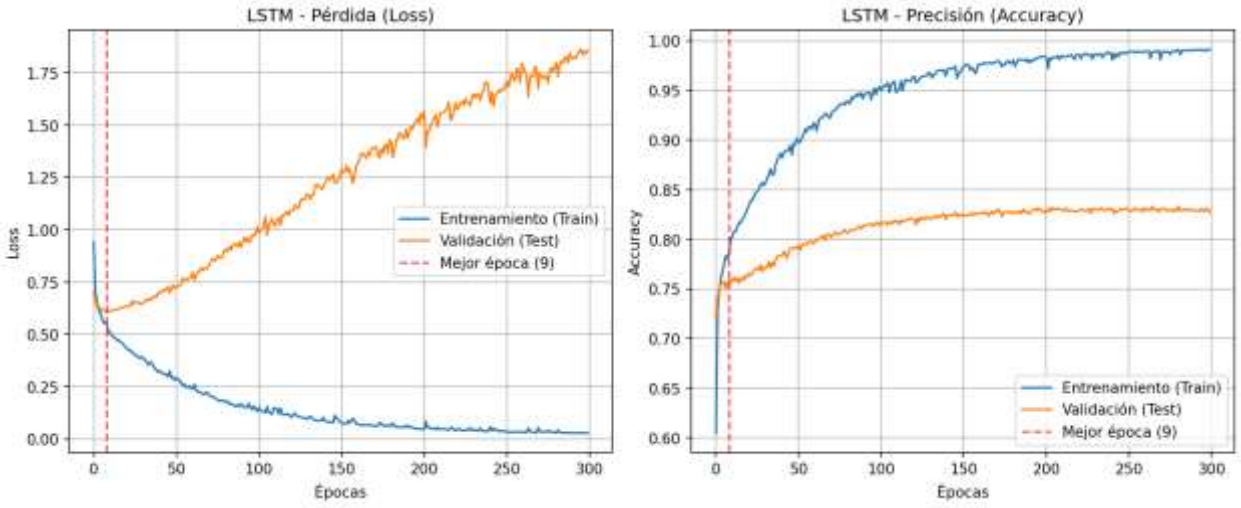


Ilustración 9 Curvas completas LSTM

Este fenómeno es característico del sobreajuste, donde el modelo empieza a memorizar el conjunto de entrenamiento en detrimento de su capacidad para generalizar a datos no vistos.

La gráfica de precisión refuerza este diagnóstico: mientras la exactitud en entrenamiento se aproxima progresivamente al 99%, la precisión de validación se estabiliza alrededor del 83%, sin mejoras significativas posteriores por esta razón, la época 9 fue seleccionada como la mejor iteración, respaldada por el uso de técnicas de Early Stopping y Model Checkpoint.

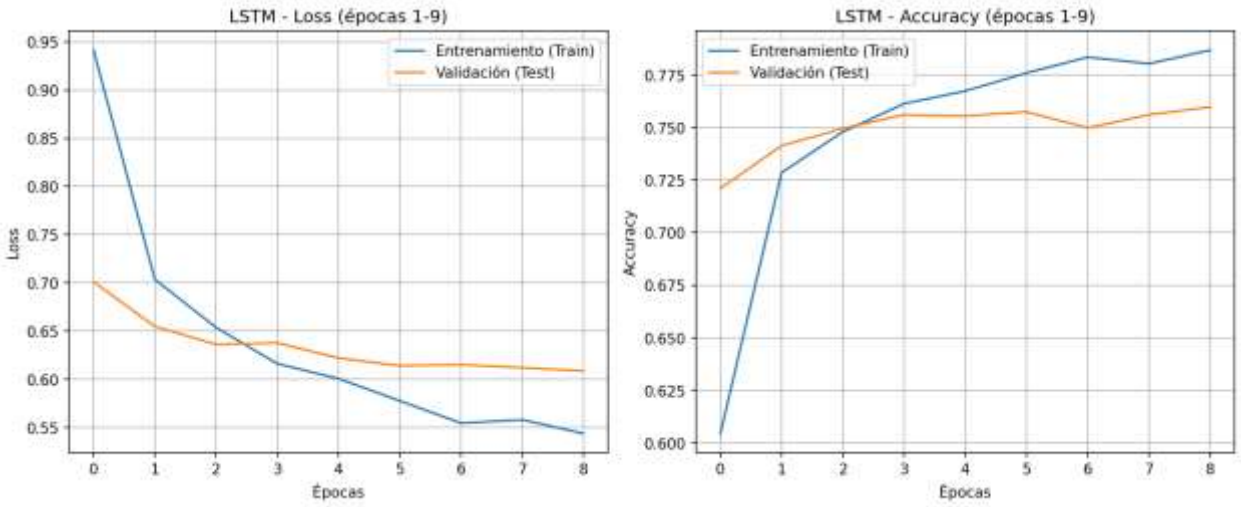


Ilustración 10 Curvas Óptimas LSTM

Análisis del modelo GRU

La aplicación del modelo GRU exhibió curvas de aprendizaje con comportamiento similar a LTSM, pero convergió ligeramente, durante las primeras 8 épocas. La pérdida de entrenamiento disminuyó de manera consistente, y, la precisión aumentó de forma estable, tanto en el conjunto de datos de entrenamiento como en validación.

En la Ilustración 11 se muestra que al igual que en el modelo LSTM, las curvas completas evidencian un incremento continuo de la pérdida de validación a partir de la época 8, confirmando la aparición de sobreajuste cuando el entrenamiento se extiende innecesariamente.

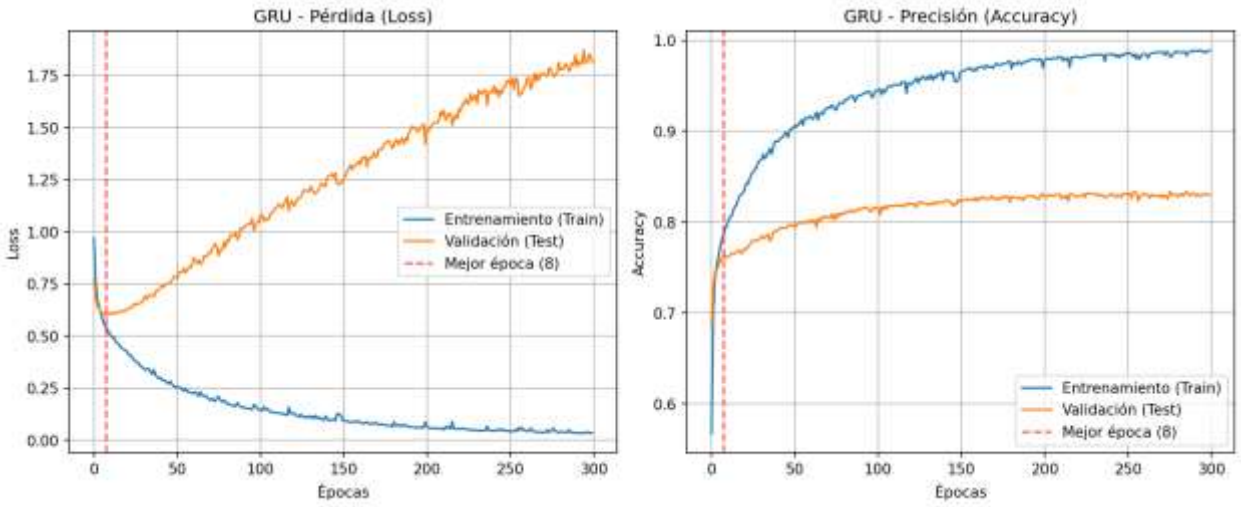


Ilustración 11 Curvas completas GRU

GRU muestra una mejor estabilidad y capacidad de generalización ligeramente superior deducible a partir de la menor brecha entre las curvas de entrenamiento y validación durante las primeras 10 épocas. Este es esperable debido a la arquitectura simplificada de GRU, que opta por reducir la complejidad sin sacrificar desempeño del modelo.

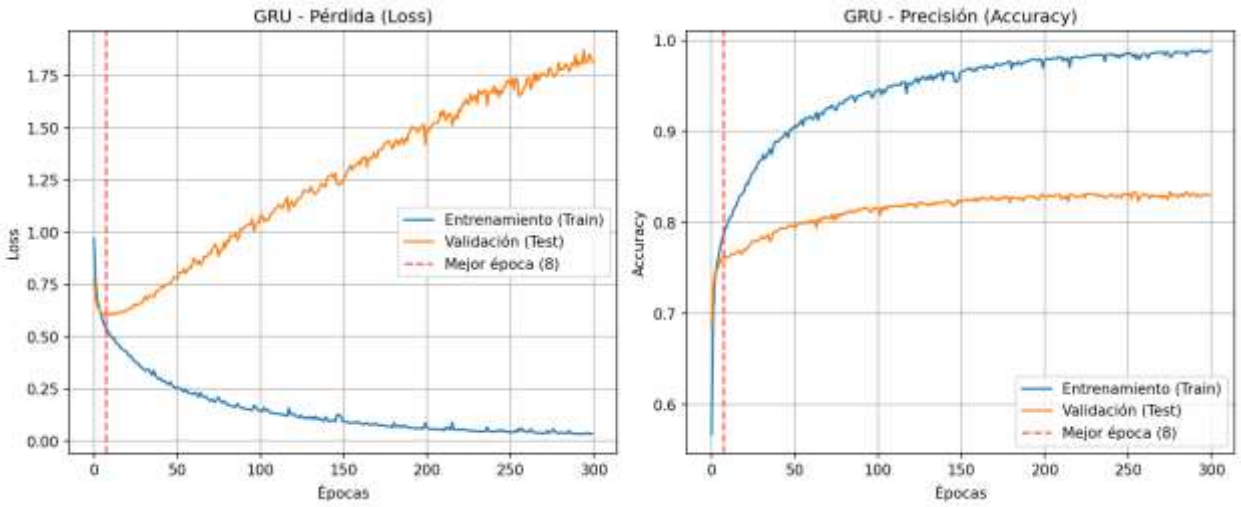


Ilustración 12 Curvas óptimas GRU

Comparación e implicaciones del entrenamiento

Para el análisis de ambas arquitecturas se confirma que el incremento del número de épocas no se traduce en mejoras reales del desempeño, sino que acentúa el sobreajuste. En consecuencia, la aplicación de estrategias de regularización implícita, como Early Stopping, resulta fundamental para preservar la capacidad predictiva de los modelos.

En términos comparativos, el modelo GRU alcanza su punto óptimo una época antes que LSTM y mantiene curvas de validación más estables, lo que anticipa su mejor desempeño en eficiencia computacional y su uso ideal para escenarios con recursos limitados.

5.2. Resultados comparativos de las arquitecturas

Con el objetivo de identificar que arquitectura es la más adecuada para la tarea de clasificación de sentimientos, se realizó una comparación directa entre los modelos LSTM y GRU considerando tres dimensiones fundamentales: eficacia predictiva, eficiencia computacional y complejidad estructural.

5.2.1. Métricas de rendimiento (Eficacia)

Los resultados obtenidos sobre el conjunto de prueba muestran que ambas arquitecturas alcanzaron un desempeño competitivo, superando el 83% de exactitud global en datos no vistos. Sin embargo, el modelo GRU presentó una ligera ventaja tanto en exactitud como en la minimización de la función de pérdida.

Tabla 5

Comparativa de rendimiento técnico entre ambos modelos.

Métrica	Modelo LSTM (Bidireccional)	Modelo GRU (Bidireccional)	Interpretación
Accuracy Global	83.34%	83.38%	GRU presenta una mejora marginal en exactitud.
Loss (Pérdida)	1.7980	1.7516	GRU minimiza mejor el error.

Precision	0.81	0.82	GRU tiene mejor capacidad para no generar falsos positivos.
Recall	0.79	0.80	GRU recupera mejor la información relevante.
F1-Score	0.80	0.81	GRU ofrece un mejor balance general.

Nota. Más allá del empate técnico en la exactitud global donde el diferencial es apenas del 0.04%, la superioridad de la arquitectura GRU se manifiesta en la minimización de la función de pérdida (Loss). Este descenso métrico no es trivial; matemáticamente implica que la red no solo clasifica correctamente, sino que lo hace con una calibración probabilística más ajustada, reduciendo la incertidumbre de la predicción.

Esta firmeza estadística es el factor crítico para su implementación en farmacovigilancia. Asimismo, la eficiencia estructural es concluyente: GRU alcanza esta robustez utilizando una topología más ligera, validando el principio de parsimonia al superar a la LSTM con una menor carga de parámetros computacionales.

5.2.2. Análisis de tiempos y costo computacional

El análisis de eficiencia computacional se realizó utilizando tiempos reales de entrenamiento, medidos directamente durante la ejecución de los experimentos bajo las mismas condiciones de hardware y configuración.

Tabla 6

Comparativa de tiempos totales de los entrenamientos

Factor tiempo	**Modelo LSTM**	**Modelo GRU**	**Diferencia**
Tiempo total (segundos)	33,284 s	29,875 s	-3,409 s
Tiempo total (horas)	9.25 horas	8.30 horas	-0.95 horas (~57 min)
Eficiencia relativa	Estándar	~10.24% más rápido	Ahorro del ~10% en cómputo

Nota. Esto representa una diferencia absoluta de 3,409 segundos, es decir, aproximadamente 56.8 minutos de ahorro a favor de la arquitectura GRU.

En términos relativos, el modelo GRU requirió alrededor de un 10.2% menos de tiempo total de entrenamiento con respecto de LSTM, lo que se atribuye a la menor complejidad interna de la celda GRU, pese a que ambos fueron entrenados bajo idénticos hiper parámetros, estrategia de regularización y entorno de ejecución.

Si bien existe una diferencia aproximada a una hora el impacto de esto es significativo en escenarios reales donde se realizan múltiples reentrenamientos, ajustes de hiper parámetros y consumo de recursos computacionales usualmente limitados… GRU se presenta como una alternativa más eficiente y sostenible.

5.3. Análisis detallado por clase (Sentimiento)

Dado el desbalance inherente del dataset, se realizó un análisis detallado por clase utilizando la métrica F1-Score, la cual permite evaluar de forma equilibrada la precisión y la exhaustividad.

Tabla 7

Reporte de Clasificación por Sentimiento (Modelo GRU)

Clase (Sentimiento)	Precision	Recall	F1Score	Cantidad de Muestras
Negativo (0)	0.81	0.72	0.76	8,969
Neutral (1)	0.70	0.55	0.62	5,348
Positivo (2)	0.87	0.94	0.90	18,743
Promedio Ponderado	0.83	0.83	0.83	33,060

Nota. Las métricas de precisión, recall y F1-score se calcularon a partir del conjunto de prueba el promedio ponderado considera la proporción de muestras de cada clase.

Los resultados muestran un desempeño sobresaliente en la clase Positiva, con un F1-Score del 90%, lo que evidencia la alta capacidad del modelo para identificar experiencias favorables asociadas a los tratamientos. La clase Negativa también presenta resultados sólidos, indicando una adecuada detección de reseñas con efectos adversos o insatisfacción terapéutica.

En contraste, la clase Neutral obtuvo el menor desempeño, con un F1-Score del 62%. Este resultado se atribuye a la ambigüedad semántica presente en muchas reseñas médicas, donde los pacientes suelen describir simultáneamente beneficios y efectos secundarios, dificultando una categorización clara. Este comportamiento representa una limitación inherente al problema más que una falla específica del modelo.

Matemáticamente, este fenómeno explica la discrepancia entre las métricas por clase y el rendimiento global. Mientras que la detección de casos polares (Positivos/Negativos) alcanza niveles de excelencia (F1 > 0.90 y 0.76 respectivamente), el bajo rendimiento en la clase Neutral (0.62) actúa como un factor de penalización que arrastra el promedio general hacia abajo. Por lo tanto, el Accuracy Global del 83.3% es un reflejo ponderado de esta dificultad: el modelo es altamente fiable para confirmar el éxito o fracaso terapéutico, pero conservador ante la incertidumbre.

5.4. Análisis de la matriz de confusión

Con el objetivo de identificar los patrones de acierto y error del modelo seleccionado, se analizó la matriz de confusión correspondiente a la arquitectura GRU aplicada sobre el conjunto de prueba, la matriz permitió evaluar no solo el desempeño global del clasificador, sino también su comportamiento específico en cada una de las clases de sentimiento: Negativo, Neutral y Positivo.

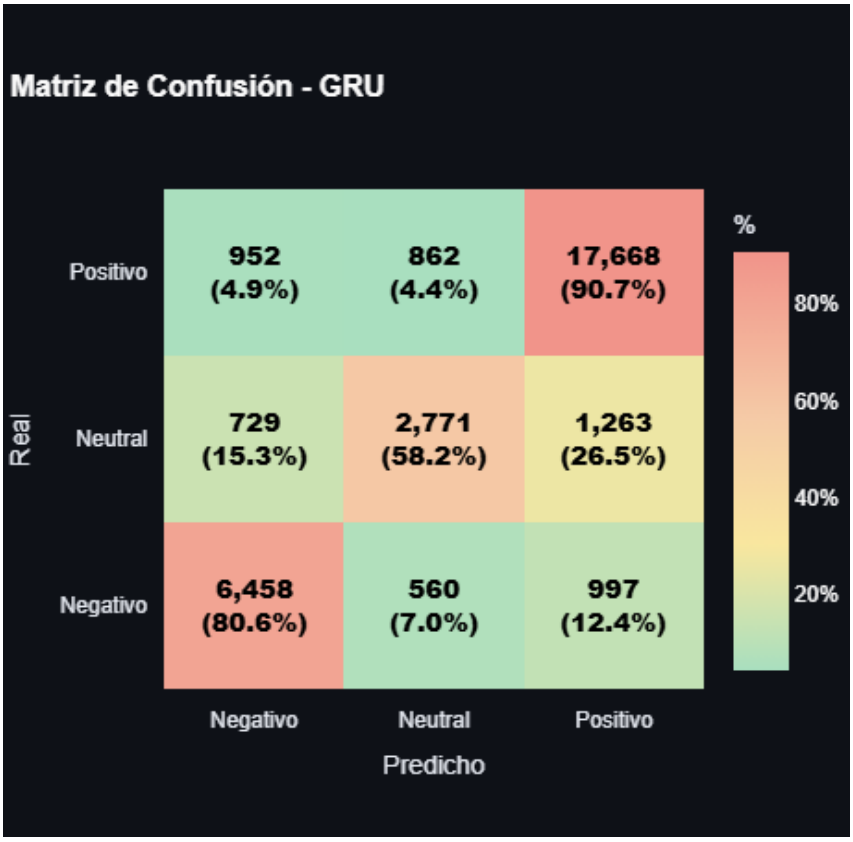


Ilustración 13 Matriz de confusión del modelo GRU

- **Dominio de la polaridad extrema (Positivo y Negativo).** La fortaleza operativa del sistema reside en los extremos del espectro**.**

    - **Clase Positiva:** Constituye el pilar de fiabilidad del modelo de las 19.482 reseñas reales de esta categoría, el algoritmo identificó correctamente 17.668 (90,7%). Los errores aquí son marginales la "fuga" hacia la clasificación negativa fue de apenas un 4,9% (952 casos), lo que demuestra que la red distingue con claridad la satisfacción del paciente y raramente confunde el éxito terapéutico con un efecto adverso.

    - **Clase Negativa:** El desempeño en la detección de insatisfacción es robusto, alcanzando un 80,6% de aciertos (6.458 casos). Desde una perspectiva de farmacovigilancia, este indicador resulta crítico; el hecho de que solo un 12,4% de

las quejas fueran confundidas con elogios sugiere que el modelo es suficientemente seguro para filtrar reportes de efectos secundarios o fallos de medicación.

- **El punto de fractura la Clase Neutral.** La zona de incertidumbre se concentra en la franja central. La clase Neutral presenta la tasa de acierto más baja, con un 58,2% (2.771 casos). El análisis de los errores en esta fila revela un sesgo claro en el clasificador:

  - **Falsos Positivos (26,5%):** El modelo clasificó erróneamente 1.263 reseñas neutrales como Positivas. Esto indica una tendencia "optimista" del algoritmo: ante la ausencia de quejas explícitas o el uso de un lenguaje meramente descriptivo, la red tiende a asumir que la experiencia fue favorable.

  - **Falsos Negativos (15,3%):** En menor medida (729 casos), el modelo confundió la neutralidad con negatividad, probablemente debido a la mención de síntomas que, aunque no eran críticas severas, fueron interpretadas como tales por el sistema.

Conclusión del diagnóstico La matriz evidencia que la arquitectura GRU es altamente eficaz para segregar opiniones polarizadas (éxito vs. fracaso), lo cual valida su utilidad para el triaje automático. Sin embargo, la ambigüedad semántica de la clase Neutral que a menudo carece de adjetivos emocionales fuertes provoca que el modelo "alucine" positividad, un factor que debe considerarse al interpretar las métricas globales de satisfacción.

5.5. Despliegue tecnológico y herramienta de validación

Con el propósito de hacer operativos los resultados obtenidos y optimizar su interpretación, se implementó un Dashboard de Análisis de Sentimientos utilizando el framework Streamlit. Esta herramienta permitió integrar los modelos entrenados dentro de una interfaz

gráfica interactiva, orientada específicamente a los procesos de validación, auditoría y comunicación de resultados

El dashboard incorpora diversos módulos funcionales, entre ellos la comparación entre modelos, el análisis detallado por clase y un sistema de validación manual (human-in-the-loop), lo que posibilita examinar tanto métricas globales como casos individuales de manera exhaustiva. Los ensayos de latencia realizados en entorno local demostraron que el modelo basado en GRU posee la capacidad de ejecutar inferencias prácticamente en tiempo real, alcanzando tiempos de respuesta inferiores a los 200 milisegundos.

5.6. Discusión de resultados

La triangulación de los resultados cuantitativos, el análisis por clase, la evaluación cualitativa y los tiempos reales de entrenamiento permite concluir que la arquitectura GRU Bidireccional constituye la solución más adecuada para el problema planteado.

Aunque la diferencia en exactitud respecto al modelo LSTM es moderada, la superioridad del modelo GRU se manifiesta de manera consistente en términos de menor pérdida, menor complejidad estructural y una reducción aproximada del 10% en el tiempo total de entrenamiento. Estos factores convierten a GRU en una alternativa más eficiente y escalable para tareas de análisis de sentimientos en el dominio médico.

De igual manera la implementación de un dashboard validó que los modelos desarrollados pueden ser traducidos en herramientas de software funcionales, transparentes y auditables.

CAPÍTULO VI - CONCLUSIONES Y RECOMENDACIONES

6. Conclusiones

Respecto al objetivo de diseñar un proceso de transformación de datos que estandarice las condiciones médicas, se concluye que las etapas de ingeniería de datos, preprocesamiento lingüístico y normalización fueron determinantes para el desempeño final de los modelos la desconstrucción del dataset original permitió corregir problemas de granularidad que hubieran afectado negativamente el aprendizaje supervisado, mientras que la categorización médica y la incorporación de variables temporales y de utilidad social enriquecieron la representación del contexto de cada reseña.

Este hallazgo evidencia que, en problemas de PLN aplicados a dominios especializados, la calidad del preprocesamiento tiene un impacto tan relevante como la arquitectura del modelo. Logrando consolidar una base de conocimiento robusta fundamentada principalmente en las cinco categorías médicas con mayor volumen de registros: Reproductive/Hormonal, Mental Health, Other/Uncategorized, Pain Management y Endocrine/Metabolic.

Respecto al objetivo de implementar y comparar el rendimiento de dos arquitecturas de redes neuronales (LSTM y GRU), se concluye que ambas configuraciones presentan comportamientos de aprendizaje estables durante las primeras épocas, pero tienden rápidamente al sobreajuste si no se aplican mecanismos de regularización y control temprano el uso de Early Stopping y Model Checkpoint resultó fundamental para identificar el punto óptimo de generalización, seleccionando 9 épocas para LSTM y la 8 épocas para GRU para atender la necesidad de estrategias de control del entrenamiento cuando se trabaja con grandes volúmenes de texto y modelos con alta capacidad representacional.

Respecto al objetivo de evaluar la capacidad del modelo en la tarea de clasificación de la percepción de los pacientes, se determina que basándonos en los resultados que muestra la arquitectura GRU ofrece un rendimiento ligeramente superior al obtenido con LSTM esta ventaja no solo se refleja en las métricas de exactitud y pérdida, sino también en el comportamiento computacional durante el entrenamiento. El modelo basado en GRU logró mantener un nivel de

desempeño equivalente utilizando alrededor de 34.000 parámetros menos y completando el entrenamiento en un tiempo mucho menor, lo que reduce de forma notable la demanda de recursos. Esta diferencia cobra importancia en contextos donde la disponibilidad de hardware es limitada o cuando se requiere escalar el sistema hacia entornos de mayor carga operativa.

Respecto al objetivo de construir reflexiones cualitativas respecto de casos ambiguos encontrados, se concluye que la clase Neutral representa el principal desafío del sistema la matriz de confusión y el análisis por clase evidenciaron una tendencia sistemática a confundir reseñas neutrales con positivas, lo cual se explica por la coexistencia de descripciones de mejora y efectos adversos en un mismo texto. Este resultado confirma que las métricas globales, por sí solas, no son suficientes para evaluar modelos de PLN en dominios sensibles, y justifica plenamente la incorporación de validación human-in-the-loop.

En el mismo sentido, los resultados de clasificación global por clases fueron: Positiva 60.4% (97,410 reseñas), Negativa 24.8% (40,075 reseñas) y Neutral un 14.8% (23,812 reseñas).

En detalle, las percepciones Positivas se asocian mayoritariamente con medicinas de tipo Salud Mental en un 22.0%, seguidas de la categoría Reproductivo/Hormonal con un 17.4%, Otras categorías con un 16.1%, Analgésicos/Dolor con un 11.1% y Metabólicos con un 7.4%.

En el caso de las percepciones clasificadas como Neutrales, se visualiza una mayor concentración de reseñas en el grupo pertenecientes a Reproductivo/Hormonal, este grupo reúne el 24,8% de estas valoraciones les siguen los fármacos empleados en Salud Mental, con un 19,9%, y aquellos incluidos en la categoría Otras, que alcanzan el 14,3% también los medicamentos asociados al manejo del dolor representan el 8,2%, mientras que los de tipo Metabólico aportan el 6,2%.

En cuanto a las percepciones clasificadas como Negativas, el mayor impacto vuelve a identificarse en la categoría Reproductivo/Hormonal, donde se registra el 27,3% de las opiniones desfavorables, seguido aparece la categoría de Salud Mental, con un 16,5%, posterior por el grupo Otras, que acumula el 14,8% los tratamientos vinculados a Enfermedades Infecciosas representan el 8,0% y los relacionados con el manejo del dolor contribuyen con el 7,4%.

Al analizar los medicamentos de manera individual, se identificó que los cinco con mejor valoración por parte de los pacientes son:

- Soma (91.6%)
- Carisoprodol (90.8%)
- Rizatriptan (90.8%)
- Drysol (90.0%)
- Percocet (89.9%)

Por el contrario, los cinco fármacos con las valoraciones más bajas o con mayor proporción de opiniones negativas son:

- Miconazole (72.6%)
- Belsomra (67.2%)
- Suvorexant (65.2%)
- Dextromethorphan (63.1%)
- Cefdinir (60.9%)

En conclusión, la investigación realizada demuestra que con una integración adecuada entre preparación de datos, modelado neuronal recurrente, evaluación mixta y despliegue se aborda de manera efectiva el análisis de sentimientos en reseñas médicas. La arquitectura GRU se posiciona como una alternativa eficiente y competitiva frente a LSTM, mientras que la incorporación de validación cualitativa y herramientas interactivas aporta transparencia y profundidad al análisis, cerrando de forma sólida el ciclo metodológico planteado.

6.1. Recomendaciones

- Se recomienda que los investigadores y profesionales en ciencia de datos aplicada al ámbito médico fortalezcan y prioricen las etapas de ingeniería de datos, preprocesamiento lingüístico y estandarización semántica antes del entrenamiento de modelos de aprendizaje automático, debido a que la correcta transformación de los datos demostró ser determinante para el desempeño final de los modelos, permitiendo corregir problemas de granularidad y enriquecer el contexto de las reseñas médicas.

- Referente a la estrategia de implementación y ajuste de hiperparámetros, la experiencia empírica dicta que los mecanismos de contención específicamente el Early Stopping y el Model Checkpoint no deben tratarse como configuraciones opcionales, sino como requisitos de diseño no negociables. El comportamiento observado durante el entrenamiento reveló que la alta plasticidad de estas redes es un arma de doble filo: sin una supervisión automatizada que corte el proceso en el punto de inflexión (detectado entre la octava y novena época), el sistema degenera rápidamente hacia la memorización de ruido estadístico, anulando su capacidad de operar con datos reales.

- Simultáneamente, para entornos productivos donde el presupuesto computacional es una restricción, la evidencia señala a la red GRU como la estrategia de escalabilidad por defecto. Al lograr una paridad métrica con la LSTM eliminando 34.000 parámetros de la ecuación, esta arquitectura democratiza el acceso a modelos de alto rendimiento, permitiendo su implementación en infraestructuras ligeras sin sacrificar la precisión del diagnóstico.

- Finalmente, como respuesta a las limitaciones detectadas en el análisis cualitativo, se concluye que la confianza exclusiva en los indicadores matemáticos debe dar paso a un enfoque híbrido en aplicaciones sanitarias. La persistente confusión en la categoría "Neutral" demuestra que existen matices semánticos que escapan a la aritmética del modelo. Por tanto, la integración de un bucle de validación humana (human-in-the-loop) deja de ser un simple complemento para convertirse en un blindaje de seguridad

obligatorio, garantizando así que la ambigüedad inherente al lenguaje del paciente no se traduzca en falsos positivos clínicos.

Trabajos futuros

A partir de los hallazgos y las limitaciones identificadas en el presente documento de carácter investigativo, se proponen las siguientes líneas de trabajo para dar continuidad y profundidad al estudio del análisis de sentimientos en el dominio farmacológico:

1. **Mejora en la discriminación de la clase neutral:** Como se evidenció en la evaluación de resultados, persiste el desafío de mejorar la clasificación de la clase "Neutral", cuya ambigüedad semántica limitó el rendimiento del modelo en comparación con las clases polares. En futuras investigaciones sería pertinente orientar los esfuerzos hacia el desarrollo de enfoques híbridos que integren técnicas de re-etiquetado manual asistido y la incorporación de reglas lingüísticas especializadas. Estas estrategias permitirían desambiguar reseñas en las que coexisten referencias a efectos secundarios negativos junto con valoraciones favorables de eficacia, un patrón frecuente en textos clínicos emitidos por pacientes.
2. **Implementación de arquitecturas basadas en Transformers,** aunque las redes recurrentes (LSTM y GRU) demostraron ser efectivas y eficientes en la presente investigación, el estado del arte en Procesamiento de Lenguaje Natural se ha desplazado hacia modelos basados completamente en mecanismos de atención. Por ello, se recomienda explorar la utilización de modelos pre entrenados como BERT (Bidirectional Encoder Representations from Transformers) o su variante biomédica BioBERT, los cuales poseen una mayor capacidad para captar dependencias contextuales profundas y podrían superar las limitaciones de precisión observadas en reseñas con alta complejidad semántica.
3. **Validación clínica,** aunque la validación estadística respalda la solidez matemática del código, los números carecen de licencia médica la evolución del proyecto exige integrar el juicio de farmacólogos, no como un complemento, sino como una barrera de seguridad necesaria. El algoritmo actual, ciego al contexto clínico, corre el riesgo de equipare una molestia leve con una reacción adversa grave, sin la calibración de un experto que distinga la urgencia real del relato, el sistema podría disparar alertas de seguridad infundadas, magnificando riesgos que, desde la perspectiva sanitaria, son irrelevantes.

Referencias

Ahmad, S. A., & Veeramanickam, M. (16 de septiembre de 2024). Sentiment analysis on drug reviews using hybrid deep learning approaches. *Journal of Data Science and Intelligent Systems, 3*(3). doi:10.47852/bonviewJDSIS42022889

Akram, B. (2024). *LinkedIn.* Obtenido de https://www.linkedin.com/pulse/tokenization-text-preprocessing-nlp-bushra-akram-jmf0f/

Amazon Web Services. (2026). *¿Qué son las RNN?* Obtenido de Amazon Web Services: https://aws.amazon.com/es/what-is/recurrent-neural-network/

ApX Machine Learning. (2025). Obtenido de apxml.com: https://apxml.com/courses/rnns-and-sequence-modeling/chapter-4-rnn-training-challenges/problem-vanishing-gradients

ApX Machine Learning. (2025). Obtenido de ApX Machine Learning: https://apxml.com/courses/introduction-to-neural-networks/chapter-6-improving-network-performance/early-stopping

Bansal, A., & Bansal, R. (2024). ircle Chaotic Mapping Based White Shark Optimization Algorithm for Sentiment Analysis Classification. IEEE. Obtenido de https://ieeexplore.ieee.org/document/10872303

Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., & Guttag, J. (2 de marzo de 2020). What is the state of neural network pruning? *Proceedings of Machine Learning and Systems, 2*, 129-146. Obtenido de https://arxiv.org/pdf/1909.09586

Boissel, J.-P., Cogny, F., Marko, N. F., & Boissel, F.-H. (septiembre de 2019). From Clinical Trial Efficacy to Real-Life Effectiveness: Why Conventional Metrics Do Not Work. *Drugs - Real World Outcomes, 6*(3), 125-132. doi:10.1007/s40801-019-0159-z

Brownlee, J. (25 de agosto de 2020). *Machine Learning Mastery*. Obtenido de https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/

Carton-Erlandsson, L., Martín-Duce, A., De los Reyes-Gragera-Martínez, R., Sanz-Guijo, M., Muriel-García, A., Mirón-González, R., & Gigante-Pérez, C. (mayo de 2024). Use of social networks as a source of information on health and digital health literacy in the

Spanish general population. *Revista Española de Salud Pública*. Obtenido de https://pmc.ncbi.nlm.nih.gov/articles/PMC11566471/

Focal. (24 de diciembre de 2024). Obtenido de https://www.getfocal.co/post/top-7-metrics-to-evaluate-sentiment-analysis-models

Ford, E. S., Li, C., Zhao, G., & Mokdad, A. H. (15 de octubre de 2009). Time trends and seasonal patterns of health-related quality of life among U.S. adults. *Preventing Chronic Disease*, A119. Obtenido de https://pmc.ncbi.nlm.nih.gov/articles/PMC2728661/

Gedyt. (2023). *Ensayos Clínicos*. Obtenido de Gedyt: https://gedyt.com.ar/ensayos-clinicos/

GeeksforGeeks. (2025). Obtenido de GeeksforGeeks: https://www.geeksforgeeks.org/deep-learning/categorical-cross-entropy-in-multi-class-classification/

GeeksforGeeks. (2026). Obtenido de GeeksforGeeks: https://www.geeksforgeeks.org/machine-learning/confusion-matrix-machine-learning/

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., . . . Wang, Y. (21 de diciembre de 2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology, 2*(4), 230-243. doi:10.1136/svn-2017-000101

Jordan, I. D., Sokół, P. A., & Park, I. M. (22 de julio de 2021). Gated recurrent units viewed through the lens of continuous time dynamical systems. *Frontiers in Computational Neuroscience, 15*. doi:10.3389/fncom.2021.678158

LearnCodePro. (2025). *Sequence Models & NLP. Text preprocessing: tokenization, stopwords, stemming/lemmatization*. Obtenido de LearnCodePro: https://www.learncodepro.com/tutorials/data-science-machine-learning-ai/sequence-models-nlp/sequence-models-nlp-text-preprocessing-tokenization-stopwords-stemminglemmatization

Li, J. (noviembre de 2018). *Kaggle*. Obtenido de Kaggle: https://www.kaggle.com/datasets/jessicali9530/kuc-hackathon-winter-2018

LunarTech. (18 de enero de 2025). Obtenido de LunarTech: https://www.lunartech.ai/blog/mastering-the-adam-optimizer-unlocking-superior-deep-learning-performance

Martínez López, J. C., Ascuntar Marcillo, G. D., Cárdenas Guaranán, T. V., Muñoz Chana, S. J., & Urbano Urbano, E. A. (2025). *armacovigilancia inteligente: el impacto de las*

tecnologías digitales en la seguridad de los medicamentos. Repositorio Institucional UNAD. Obtenido de https://repository.unad.edu.co/handle/10596/68359

Msheik, B., Mcheick, H., Hariri, S., Adda, M., & Dbouk, M. (11 de junio de 2025). From data to medical context: the power of categorization in healthcare. *Frontiers in Medicine*. doi:10.3389/fmed.2025.1575195

Msheik, B., Mcheick, H., Hariri, S., Adda, M., & Dbouk, M. (11 de junio de 2025). From data to medical context: the power of categorization in healthcare. *Frontiers in Medicine, 12*. doi:10.3389/fmed.2025.1575195

Mucci, T. (2024). *Overfitting vs. underfitting: Finding the balance*. Obtenido de IBM: https://www.ibm.com/think/topics/overfitting-vs-underfitting

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (1 de septiembre de 2011). Introduction to natural language processing. *Journal of the American Medical Informatics Association, 18*(5), 544-551. doi:10.1136/amiajnl-2011-000464

Panda, B., Panigrahi, C. R., & Pati, B. (2022). Exploratory data analysis and sentiment analysis of drug reviews. *Computación y Sistemas, 26*(3), 1191-1199. doi:10.13053/cys-26-3-4093

Pitti, A. (30 de Julio de 2024). *linkedin*. Obtenido de https://www.linkedin.com/pulse/social-media-pharmacovigilance-monitoring-drug-safety-abhishek-pitti-dq6pe

Reyes, A., Malik, M., Sahouri, M., & Knezevic, N. N. (18 de octubre de 2025). FDA-Regulated Clinical Trials vs. Real-World Data: How to Bridge the Gap in Pain Research. *Brain Sciences, 15*(10), 119. doi:10.3390/brainsci15101119

Sharma, R. (13 de marzo de 2025). Obtenido de Udacity: https://www.udacity.com/blog/2025/03/crisp-dm-explained-a-proven-data-mining-methodology.html

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (27 de abril de 2019). Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics, 7*(2), e12239. doi:10.2196/12239

Singh, M. P. (14 de mayo de 2025). *Vizuara*. Obtenido de https://vizuara.substack.com/p/from-words-to-vectors-understanding

Swimm Team. (2023). *5 Types of Word Embeddings and Example NLP Applications*. Obtenido de Swimm: https://swimm.io/learn/large-language-models/5-types-of-word-embeddings-and-example-nlp-applications

Van Stekelenborg, J., Ellenius, J., Maskell, S., Bergvall, T., Caster, O., Dasgupta, N., . . . Pi. (2019). Recommendations for the Use of Social Media in Pharmacovigilance: Lessons from IMI WEB-RADR. *Drug Safety*, 1393–1407.

Wagner, N., Rieger, V., & Bedi, K. (7 de agosto de 2019). Does Money Buy Happiness? Evidence from an Unconditional Cash Transfer in Zambia. *Frontiers in Psychology, 10*, 1968. doi:10.3389/fpsyg.2019.01968

Walczak, S., & Okuboyejo, S. (01 de 2017). *esearchgate*. Obtenido de https://www.researchgate.net/publication/311670797_An_Artificial_Neural_Network_Classification_of_Prescription_Nonadherence

Anexos

https://github.com/elviisch26/Redes-Neuronaels-Recurrentes