PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

CARRERA:

INGENIERIA EN SISTEMAS DE LA INFORMACIÓN

TRABAJO DE TITULACIÓN

TEMA:

APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING PARA EL ANÁLISIS DE SENTIMIENTO Y DETECCIÓN DE TEMAS PÚBLICOS DE REDES SOCIALES SOBRE EMPRESAS DEL SECTOR FAMACÉUTICO.

ESTUDIANTE:

ADRIÁN SEBASTIÁN GUANOLUISA RAMOS

TUTOR:

ING. EDISON MORA

QUITO DM, ENERO DE 2026

Índice

Índice de Tablas.

1. CAPITULO I: INTRODUCCIÓN

1.1. Tema

Aplicación de técnicas de *Machine Learning* para el análisis de sentimiento y detección de temas en publicaciones de redes sociales sobre empresas del sector farmacéutico.

1.2. Justificación

El crecimiento del uso de redes sociales ha transformado la forma en que los usuarios expresan sus opiniones, experiencias y percepciones respecto a productos y servicios, especialmente dentro del sector farmacéutico.

Estas plataformas generan una gran cantidad de información textual que refleja la reputación digital de las empresas, sus productos y reacciones del público ante distintos eventos o lanzamientos.

Sin embargo, la gran cantidad de información generada en redes sociales supera la capacidad de análisis manual, lo que dificulta su estudio por métodos tradicionales. Por esa razón, el uso de técnicas automatizadas se vuelve necesario para procesar y analizar.

En este contexto, *Machine Learning* y el Procesamiento de Lenguaje Natural permiten identificar sentimientos y temas recurrentes presentes en las conversaciones digitales. En esta investigación utiliza estas técnicas para detectar patrones de opinión pública y evaluar la reputación digital de las empresas del sector farmacéutico.

Finalmente, el estudio posee relevancia académica al aplicar métodos de inteligencia artificial en el análisis de fenómenos sociales. También, presenta relevancia práctica al ofrecer una herramienta analítica que puede apoyar procesos de monitoreo reputacional y toma de decisiones estratégicas.

1.2.1. Justificación Social

El sector farmacéutico desempeña un papel esencial en la salud pública, por eso la percepción ciudadana sobre su funcionamiento influye en la confianza social. Las redes sociales se han consolidado espacios donde los usuarios comparten reclamos, denuncias y experiencias relacionadas con el acceso y calidad de los medicamentos.

El análisis de estas opciones ayuda a identificar preocupaciones reales de la población y comprender el impacto en la reputación digital del sector farmacéutico.

1.2.2. Justificación Académica

Desde una perspectiva académica, esta investigación aporta al estudio del análisis de sentimiento y detección de temas mediante técnicas de *Machine Learning* aplicadas a un contexto social específico.

El trabajo amplía el uso de inteligencia artificial en el análisis de fenómenos sociales vinculados en el sector farmacéutico del país. Asimismo, el estudio contribuye a cubrir la limitada literatura existente sobre la aplicación práctica de estas técnicas.

1.2.3. Justificación Tecnológica

En el sector tecnológico, el estudio evidencia la aplicabilidad de técnicas de *Machine Learning* y Procesamiento de Lenguaje Natural para el análisis automatizado de grandes volúmenes de texto.

Dichas herramientas ayudan a procesar información de manera eficiente, superando las limitaciones del análisis manual. También, los resultados obtenidos demuestran que estos enfoques generan información escalable, reproducible y útil para la toma de decisiones.

1.3. Planteamiento de Problema

Las empresas farmacéuticas operan en un entorno comunicacional dinámico, donde las redes sociales influyen en la construcción de su reputación digital. Publicaciones, comentarios y debates sobre medicamentos y servicios de salud pueden modificar la percepción del público.

A pesar de la cantidad de información disponible en plataformas digitales, las organizaciones carecen de mecanismos sistemáticos que les permitan realizar adecuadamente las opiniones de los usuarios.

Esta limitación reduce su capacidad de respuesta frente a posibles crisis reputacionales o procesos de desinformación.

Por ello, nace la necesidad de aplicar un enfoque analítico basado en técnicas de *Machine Learning* que permita clasificar el sentimiento de las publicaciones y detectar temas más recurrentes en la conversación digital.

En el contexto ecuatoriano, esta necesidad se intensifica debido a los recurrentes cuestionamientos sobre abastecimiento, precios y calidad del servicio farmacéutico.

Estas situaciones generan una alta participación ciudadana en plataformas digitales, donde se comparten opiniones y experiencias de manera masiva.

A pesar de esta gran cantidad de información disponible, las empresas y organizaciones del sector carecen de herramientas automatizadas que les permitan analizar de forma sistemática la percepción pública.

La ausencia de estos mecanismos dificulta la identificación temprana de problemas reputacionales y limita la capacidad de respuesta institucional.

El problema se puede resumir en la siguiente pregunta. ¿Cómo aplicar técnicas de *Machine Learning* para analizar sentimientos y detectar temas en publicaciones de redes sociales, con el fin de evaluar la reputación digital de las empresas del sector farmacéutico?

1.4. Objetivos

1.4.1. General:

Aplicar técnicas de *Machine Learning* para el análisis de sentimiento y detección de temas en publicaciones de redes sociales, con el propósito de evaluar la reputación digital de empresas del sector farmacéutico.

1.4.2. Específicos:

- Recolectar y preparar un conjunto de datos de publicaciones en redes sociales relacionadas con empresas farmacéuticas, mediante técnicas de extracción de datos y limpieza textual.
- Aplicar algoritmos *de Machine Learning* supervisados y no supervisados para realizar análisis de sentimiento y detección de temas.
- Analizar resultados obtenidos para identificar tendencias, percepciones y temas predominantes en torno a las empresas estudiadas.
- Evaluar la utilidad del análisis de datos en la comprensión y gestión de la reputación digital del sector farmacéutico.

1.5. Alcance

El presente trabajo tiene un alcance descriptivo y analítico, centrado en el estudio de la reputación digital del sector farmacéutico mediante el uso de técnicas de *Machine Learning* aplicadas al análisis de publicaciones digitales.

El objetivo es identificar patrones de sentimiento y temas predominantes en la conversación pública sobre empresas farmacéuticas, con el fin de interpretar la percepción pública hacia este sector.

La investigación se desarrollará en el contexto geográfico de la provincia de Pichincha con énfasis en la ciudad de Quito, por ser el principal centro de actividad farmacéutica sanitaria y comunicacional del país. Esta delimitación permite obtener una muestra representativa del entorno digital y garantizar la viabilidad del estudio.

En el ámbito temporal, el estudio considera publicaciones generadas entre enero y junio de 2025, obtenidas desde plataformas como YouTube, reseñas de aplicaciones móviles y portales de noticias.

La recolección de datos se realizó mediante técnicas automatizadas, respetando los principios éticos y también con las normativas de privacidad.

El análisis se limita a contenido textual en idioma español, excluyendo imágenes, videos y mensajes privados. Los textos recopilados fueron procesados mediante técnicas de limpieza, normalización, tokenización y vectorización para el análisis.

Para el desarrollo del modelo se emplearán algoritmos de *Machine Learning supervisado,* como *Naive Bayes* y *SVM,* como modelos de aprendizaje profundo basados en *BERT,* complementados con técnicas de modelado de temas (LDA).

El estudio no tiene como objetivo verificar la veracidad de las publicaciones, ni realizar comparaciones directas entre empresas específicas, solo identificar patrones discursivos y emocionales, los cuales caracterizan la reputación digital.

1.6. Hipótesis

Hipótesis General:

El uso de técnicas de *Machine Learning* aplicadas al análisis de sentimiento y al modelado de temas en publicaciones de redes sociales posibilita la identificación de tendencias de opinión que representan la percepción digital asociada a las empresas del sector farmacéutico en la provincia de Pichincha.

Hipótesis Específicas:

- La aplicación de algoritmos supervisados de *Machine Learning* permite categorizar las opiniones expresadas por los usuarios en redes sociales en sentimientos positivos, negativos y neutros, dentro del contexto digital del sector farmacéutico.
- La aplicación de técnicas no supervisadas de modelado de temas, basadas en *Latent Dirichlet Allocation* (LDA), posibilita la identificación de los principales temas de conversación asociados a las empresas farmacéuticas en las publicaciones digitales analizadas.
- La combinación de los resultados obtenidos del análisis de sentimiento y del modelado de temas contribuye a una compresión más integral de la reputación digital de las empresas farmacéuticas, al evidenciar factores relacionados con la confianza y percepción ciudadana en el entrono digital local.

2. CAPITULO II: MARCO TEÓRICO Y CONCEPTUAL

2.1. Antecedentes

En la última década, la aplicación de técnicas de inteligencia artificial al análisis de opiniones expresadas en redes sociales ha adquirido mayor relevancia dentro de los estudios de reputación digital.

Varias investigaciones demuestran que el análisis automatizado del lenguaje permite interpretar grandes volúmenes de información generada por los usuarios en diferentes entornos digitales.

En este contexto, López y Ruiz (2022) desarrollaron un estudio centrado en publicaciones de X relacionadas con la empresa farmacéutica Pfizer, utilizando análisis de sentimiento para examinar la percepción social sobre las vacunas contra el virus.

Los autores evidenciaron que la polaridad de los mensajes publicados cae de manera significativa en los niveles de confianza del público.

De forma similar, Torres y Medina (2021) evaluaron la reputación digital de empresas ecuatorianas mediante algoritmos de *Machine Learning*, identificando relación directa entre la presencia de sentimientos positivos en redes sociales y fortalecimiento de la intención de compra y fidelización de los usuarios.

Por otra parte, Díaz y Herrera (2021) señalaron que la integración del análisis de sentimiento con técnicas de modelado de temas permite obtener una visión ampliada de los factores que influyen en la reputación digital.

A pesar de los avances descritos, la aplicación de técnicas de *Machine Learning* para el análisis de reputación digital en el sector farmacéutico sigue siendo limitada. Esto justifica la pertinencia y novedad del presente estudio, que busca aportar evidencia y metodologías aplicables al contexto nacional.

2.1.1. Antecedentes internacionales sobre análisis de sentimiento.

Diversos estudios internacionales han demostrado la eficiencia del análisis de sentimiento para evaluar la percepción pública en redes sociales. Devlin et al. (2019) introdujeron el modelo BERT, el cual representó un avance significativo.

En el procesamiento de lenguaje natural al incorporar un enfoque bidireccional capaz de comprender el contexto completo de las palabras dentro de una oración, superando el rendimiento de modelos tradicionales de *Machine Learning*.

A nivel internacional, existen varios estudios que han resaltado el potencial de los modelos avanzados de procesamiento del lenguaje natural para el análisis de contenido digital.

Wolf et al. (2020) destacaron que los modelos basados en arquitecturas *Transformers* ofrecen una mayor capacidad para capturar el significado contextual del texto, esto resulta especialmente relevante en plataformas digitales que se caracterizan por su lenguaje informal, ambigüedad semántica e ironía.

En el ámbito de la salud, Zhang, Wang y Liu (2021) aplicaron técnicas de análisis de sentimiento para examinar la percepción ciudadana sobre los servicios sanitarios. Sus resultados evidenciaron que una alta concentración de sentimientos negativos en redes sociales es relacionada con experiencias desfavorables en el tema de salud.

2.1.2. Antecedentes en contextos latinoamericanos.

En el contexto latinoamericano, el análisis de sentimiento ha sido empleado para estudiar la opinión pública en sectores considerados críticos, tales como la salud y los servicios públicos.

Asimismo, Ramírez et al. (2022) utilizaron modelos de *Machine Learning* para examinar la reputación digital de organizaciones en distintos países. Los autores concluyeron que la combinación del análisis de sentimiento con técnicas de detección de temas ayuda a identificar factores críticos que influyen en la percepción pública e imagen institucional.

Estos antecedentes evidencian la aplicabilidad de técnicas de *Machine Learning* en contextos sociales complejos y refuerzan la necesidad de investigaciones localizadas que consideren las particularidades lingüísticas y culturales propias de cada país.

2.1.3. Síntesis de antecedentes y aporte del estudio

A partir de la revisión de los antecedentes, se observa que, si bien existen múltiples investigaciones sobre análisis de sentimiento y reputación digital, son limitados los estudios enfocaos específicamente en el sector farmacéutico ecuatoriano.

Además, pocos trabajos integran de manera conjunta modelos tradicionales de *Machine Learning* con enfoques avanzados basados en transformers.

En este sentido, el presente estudio aporta una aproximación metodológica multifuente y contextualizada, aplicando técnicas de análisis de sentimiento y detección de temas para evaluar la reputación digital del sector farmacéutico, contribuyendo tanto al conocimiento académico como a la práctica institucional.

2.2. Fundamentos Teóricos.

2.2.1. Reputación digital en redes sociales.

La reputación digital puede entenderse como la valoración colectiva que los usuarios construyen sobre una organización a partir de sus interacciones, comportamientos y mensajes difundidos en medios digitales (García & Calderón, 2021).

En este marco, las redes sociales desempeñaron un rol central, ya que funcionan como espacios donde los usuarios expresan opiniones, experiencias y reacciones ante la comunicación corporativa de las marcas (Rodríguez, 2020).

En el sector farmacéutico, la reputación digital adquiere una importancia particular debido a la relación directa entre los productos ofrecidos y la salud de la población. Una opinión negativa o publicación viral puede influir significativamente en la percepción pública y afectar la credibilidad de la empresa farmacéutica (Kim, Lee & Park, 2019).

2.2.1.1. Gestión y construcción de la reputación digital.

La reputación digital no surge de forma inmediata, sino que se consolida a través de un proceso continuo de interacción entre las organizaciones y usuarios en los entornos digitales. La coherencia comunicacional, transparencia y respuesta frente a las opiniones públicas, influyen en este proceso.

Fombrun (2012) sostiene que la reputación constituye un activo intangible de carácter estratégico, esto incide en la confianza, credibilidad y legitimidad de una organización ante sus diferentes del público.

En el contexto de las redes sociales, este activo se ve condicionado por la velocidad de difusión de la información y por la capacidad de los usuarios para amplificar contenidos positivos o negativos.

Kaplan y Haenlein (2010) destacan que las plataformas sociales facilitan una comunicación bidireccional que reduce el control absoluto de las organizaciones sobre su imagen, aumentando la relevancia de la percepción colectiva.

En este sentido, la gestión de la reputación digital requiere un monitoreo constante de las opiniones expresadas por los usuarios.

El análisis automatizado de contenido textual se transforma en una herramienta fundamental para interpretar volúmenes de información, detectar variaciones en la percepción pública y anticipar riesgos reputacionales (Aula, 2010).

2.2.1.2. Reputación digital en sectores sensibles como la salud.

En sectores farmacéuticos, la reputación digital adquiere relevancia mayor debido a la relación directa entre los productos que ofrecen y la salud de las personas que la compran.

La confianza del consumidor en tratamientos y medicamentos depende de la información disponible y de experiencias compartidas en entornos digitales (Edelman, 2021).

Diversas investigaciones muestran que las opiniones negativas difundidas en redes sociales sobre efectos secundarios, precios elevados o problemas de abastecimiento pueden generar pérdida de confianza y destruir la imagen de la empresa (Ventola, 2014).

De igual manera, la propagación de desinformación médica en entornos digitales representa un riesgo adicional para la reputación del sector Chou, Oh y Klein (2018) advierte que la circulación de información no verificada influye negativamente en el público y en las decisiones de los usuarios.

2.2.1.3. Relación entre reputación digital y análisis automatizado.

El análisis automatizado de sentimiento y detección de tema permite transformar volúmenes de texto en indicadores útiles para la gestión de la reputación digital.

De acuerdo con Cambria et al. (2017), estas técnicas facilitan la identificación de tendencias emocionales y temáticas que reflejan el estado de la percepción pública en intervalos de tiempo.

La integración del análisis de sentimiento con el modelado de temas proporciona una visión más completa de la conversación digital, porque permite conocer no solo la orientación emocional de las opiniones, también los aspectos que los usuarios expresan sus percepciones.

Este enfoque resulta especialmente pertinente para el sector farmacéutico, donde más factores influyen simultáneamente en la reputación digital. En consecuencia, el uso de técnicas de *Machine Learning* aplicadas al análisis de publicaciones en redes sociales.

2.2.2. Análisis de Sentimiento (*Sentiment Analysis*).

El análisis de sentimiento o minería de opiniones es una técnica del Procesamiento de Lenguaje Natural cuyo propósito es identificar la orientación emocional de un

texto, clasificándolo generalmente como positivo, negativo o neutro (Medhat, Hassan & Kprashy, 2014).

2.2.2.1. Enfoques y metodologías

- **Enfoque léxico-basado:** Usa diccionarios de palabras con connotación positiva/negativa para poder clasificar.
- **Enfoque de aprendizaje automático (*Machine Learning*):** Entrena modelos como *Naive Bayes, SVM* para clasificar textos según el sentimiento.
- **Enfoque de aprendizaje profundo (*Deep Learning*):** Redes neuronales, *BERT* y *Transformers* que capturan contexto y dependencias más complejas.

2.2.2.2. Importancia en reputación digital.

La aplicación del análisis de sentimiento permite a las organizaciones monitorear la percepción pública expresada en plataformas digitales, identificar señales tempranas de crisis reputacionales y ajustar sus estrategias de comunicación.

En un entorno donde gran parte de la interacción con la marca se produce en redes sociales, este tipo de análisis es fundamental para poder gestionar la reputación y que sea efectiva (Zhang, Wang & Liu, 2021).

2.2.2.3. Modelos supervisados en el análisis de sentimiento.

Dentro del enfoque de aprendizaje automático, los modelos supervisados muestran un desempeño sólido en tareas de clasificación textual. *Naive Bayes* es uno de los algoritmos utilizados por si simplicidad, eficiencia computacional y buenos resultados con datos grandes (Manning, Raghavan & Schütze, 2008).

Por otro lado, *Support Vector Machine* es aplicada en análisis de sentimiento gracias a sus capacidades para definir fronteras de decisión óptimas entre clases, incluso en espacios de alta dimensionalidad (Cortes & Vapnik, 1995).

Diversos estudios indican que *SVM* suele presentar un rendimiento equilibrado que *Naive Bayes* en escenarios multiclase (Joachims, 2002).

A pesar del avance de técnicas más complejas, estos modelos tradicionales continúan siendo relevantes como línea base para la comparación con enfoques de aprendizaje profundo.

2.2.2.4. Aprendizaje profundo y modelos basados en transformers.

En los últimos años, los enfoques de aprendizaje profundo han superado los modelos tradicionales en múltiples tareas de procesamiento del lenguaje natural. Específicamente, los modelos basados en transformers introdujeron mecanismos de atención que permiten analizar el contexto semántico completo de los textos (Vaswani et al., 2017).

BERT (Bidirectional Encoder Representations from Transformers), propuesto por Devlin et al. (2019), constituye uno de los avances más relevantes en este ámbito, permite un análisis bidireccional del lenguaje y capturar relaciones complejas.

Esta capacidad resulta demasiada útil en redes sociales, donde el lenguaje suele ser informal, ambiguo y cargado de significados implícitos.

Recientes estudios demuestran que *BERT* mejora de manera significativa la precisión del análisis de sentimiento en comparación con modelos tradicionales, en mayor cantidad en textos breces y con variabilidad lingüística (Sun, Huang & Qiu, 2019).

2.2.2.5. Comparación de enfoques y pertinencia del estudio.

La literatura especializada destaca la importancia de comparar distintos enfoques de análisis de sentimiento, esto para ver su desempeño en diferentes contextos. Según Liu (2020), la combinación de modelos tradicionales y técnicas de aprendizaje profundo ayuda a tener una compresión más ampliada del comportamiento emocional del lenguaje.

En este sentido, el presente estudio adopta un enfoque comparativo al aplicar modelos de *Naive Bayes*, *SVM* y *BERT* para el análisis de sentimiento en publicaciones relacionadas con el sector farmacéutico.

2.2.3. Detección de Temas (*Topic Modeling*)

El *Topic Modeling* es una técnica de minería de texto no supervisado que busca descubrir los temas latentes en grandes volúmenes de información textual (Blei, Ng & Jordan, 2003).

Uno de los modelos más utilizados es *Latent Dirichlet Allocation* (LDA), el cual representa cada documento como una combinación de temas, y cada tema como una distribución de palabras. Este modelo permite identificar los temas más discutidos por los usuarios en redes sociales, tales como "eficacia del medicamento", "efectos secundarios" o "precio" (Wang, Zhang & Li, 2020).

De acuerdo con Röder, Both y Hinneburg (2015), el uso de *Topic Modeling* complementa el análisis de sentimiento, ya que permite no solo conocer cómo se sienten los usuarios, sino también sobre qué temas están opinando.

2.2.3.1. Importancia del modelado de temas en grandes volúmenes de texto.

El aumento sostenido de contenido textual generado en redes sociales ha hecho inviable su análisis mediante métodos manuales, especialmente cuando se trata de grandes volúmenes de información.

Ante este escenario, se vuelve necesario aplicar técnicas automáticas que permitan explorar y sintetizar el contenido de manera eficiente.

El modelado de temas se ha consolidado como una estrategia adecuada para identificar patrones temáticos presentes en colecciones extensas de documentos, ayudando revelar estructuras latentes en textos sin intervención humana (Blei, 2012).

A diferencia de los enfoques supervisados, estas técnicas no dependen de etiquetas previas.

Esta característica resulta particularmente valiosa en estudios de opinión pública, en donde los temas de discusión nacen de forma espontánea a partir de la interacción de usuarios en los entornos digitales, sin ninguna estructura antes (Griffiths & Steyvers, 2004).

2.2.3.2. Latent Dirichlet Allocation (LDA) y su aplicación.

Es uno de los modelos probabilísticos más utilizados para la detección automática de temas en textos. Este enfoque parte del supuesto de que cada documento puede representarse como una combinación de varios temas, y cada tema se caracteriza por la distribución específica de palabras, esto facilita identificar patrones semánticos (Blei, Ng & Jordan, 2003).

El uso de *LDA* ha sido documentado en investigaciones relacionadas con redes sociales y análisis de reputación digital. De acuerdo con Wang, Zhang y Li (2020), este modelo identifica preocupaciones recurrentes de los usuarios, como la calidad del servicio, precios o disponibilidad de productos.

También, Röder, Both y Hinneburg (2015) proponen métricas de coherencia temática que permiten evaluar la calidad de tópicos generados, contribuyendo a fortalecer la validez y confiablidad de resultados obtenidos mediante este modelo.

2.2.3.3. Relación entre *Topic Modeling* y reputación digital

La detección de temas complementa el análisis de sentimiento al proporcionar un contexto semántico que permite interpretar con mayor precisión las emociones expresadas por los usuarios.

Mientras el análisis de sentimiento responde a la orientación emocional de las opiniones, el modelado de temas permite identificar los aspectos específicos sobre los cuales se está opinando (Cambria et al., 2017).

En el sector farmacéutico, esta integración resulta especialmente relevante, ya que la reputación digital puede verse influenciada simultáneamente por múltiples factores, como la eficiencia de los medicamentos, los efectos secundarios, precios o calidad.

La identificación de estos temas ayuda a las organizaciones comprender las causas subyacentes de los sentimientos positivos o negativos expresados en redes sociales.

En consecuencia, la combinación de ambas técnicas ofrece una visión integral de la reputación digital y facilita una interpretación más precisa de las percepciones ciudadanas.

2.2.4. Técnicas de *Machine Learning* aplicadas al análisis de publicaciones sociales.

Machine Learning o aprendizaje automático es una rama de la inteligencia artificial que permite a los sistemas identificar patrones en los datos y realizar predicciones o clasificaciones sin utilizar programación para cada caso (Mitchell, 1997).

En el análisis de texto, la aplicación de *Machine Learning* requiere un proceso previo de preparación de los datos, que incluye etapas de *tokenización*, eliminación de palabras vacías, lematización y vectorización (Manning & Schütze, 1999).

Entre los modelos utilizados en el análisis de publicaciones sociales se encuentran *Naive Bayes, Support Vector Machine (SVM)* y modelos basados en aprendizaje profundo como *BERT,* cada uno con su nivel de complejidad.

2.2.4.1. Proceso de análisis textual mediante Machine Learning

El análisis de publicaciones en redes sociales mediante *Machine Learning* sigue un proceso estructurado que permite convertir texto no estructurado en información útil. Este proceso inicia con la recolección de datos y continúa con una fase de preprocesamiento orientada a mejorar la calidad del contenido textual.

Durante esta etapa se aplican técnicas como normalización, eliminación de caracteres especiales, tokenización, eliminación de palabras vacías y lematización, con el fin de reducir ruido y facilitar el aprendizaje (Manning & Schütze, 1999).

Posteriormente, los textos son transformados en representaciones numéricas mediante métodos de vectorización. Entre las más utilizadas se encuentra *Term Frequency-Inverse Document Frequency* (TF-IDF).

La cual permite ponderar la relevancia de las palabras dentro del corpus y facilita la clasificación por medio de algoritmos supervisados (Salton & Buckley, 1988).

2.2.4.2. Modelos tradicionales de *Machine Learning.*

Los modelos tradicionales de *Machine Learning* continúan siendo empleadas en tareas de análisis de sentimiento debido a su eficiencia computacional y facilidad de interpretación.

Naive Bayes es un clasificador probabilístico que asume independencia condicional entre las características del texto, lo que lo convierte en una opción rápida y efectiva para grandes volúmenes de datos (Manning et al., 2008).

Por su parte, SVM ha demostrado un alto desempeño en la clasificación de textos al identificar hiperplanos óptimos que separan las distintas clases de sentimiento en espacios de alta dimensionalidad.

Según Cortes y Vapnik (1995), este modelo es especialmente robusto frente a datos ruidosos y desbalanceados.

La regresión logística también es utilizada como modelo base en análisis de sentimiento, ya que permite estimar probabilidades de pertenencia a cada clase y facilita la interpretación de resultados, siendo útil como referencia comparativa frente a enfoques complejos (Hosmer, Lemeshow & Sturdivant, 2013).

2.2.4.3. Modelos de aprendizaje profundo y *BERT*.

El aprendizaje profundo ha permitido superar varias limitaciones de los modelos tradicionales al capturar relaciones semánticas complejas en el lenguaje. Dentro de este enfoque, los modelos basados transformers han demostrado un buen rendimiento en tareas de clasificación textual (Vaswani et al., 2017).

BERT (Bidirectional Encoder Representations from Transformers) introduce una arquitectura que analiza el contexto de las palabras de forma bidireccional, permitiendo una comprensión más precisa del significado de los textos.

Diversos estudios han evidenciados que *BERT* resulta especialmente eficaz en textos cortos y con estructuras lingüísticas complejas, características comunes en publicaciones de redes sociales, y ha sido aplicado con éxito en el análisis de opiniones dentro del sector de salud (Sun, Huang & Qiu, 2019).

2.2.4.4. Justificación del uso de múltiples algoritmos.

La literatura especializada recomienda la utilización de múltiples algoritmos para evaluar el desempeño del análisis de sentimiento en distintos contextos. Según Liu (2020), la comparación entre modelos tradicionales y enfoques de aprendizaje profundo permite identificar el método más adecuado según características de aprendizaje profundo permite identificar el método más adecuado dependiendo las características del conjunto de datos.

En este estudio se emplean *Naive Bayes*, Regresión Logística, *SVM* y *BERT* con el propósito de comparar métricas de desempeño como precisión, recall y F1-score. Esta estrategia garantiza una evaluación objetiva y robusta del análisis de sentimiento aplicado a publicaciones relacionadas con el sector farmacéutico.

2.2.5. Aplicación al sector farmacéutico.

En el sector farmacéutico, el análisis de publicaciones en redes sociales constituye una herramienta estratégica para comprender la opinión de los usuarios sobre medicamentos, tratamientos y marcas.

Gómez (2023) señala que la reputación digital de las empresas farmacéuticas depende en gran medida de su capacidad para gestionar información, transparencia y comunicación con el público.

La integración del análisis de sentimiento y detección de temas permite identificar preocupaciones recurrentes y evaluar el nivel de confianza hacia los productos ofrecidos, aportando información valiosa para la toma de decisiones y diseño de estrategias comunicacionales.

En el contexto ecuatoriano, el sector farmacéutico ha sido objeto de atención constante debido a problemáticas asociadas al abastecimiento de medicamentos, precios, calidad del servicio y acceso a la salud.

De acuerdo con Edelman (2021), en sectores vinculados a la salud, la confianza del público se construye a partir de la transparencia, comunicación efectiva y la capacidad de respuesta ante las inquietudes ciudadanas.

En este sentido, el análisis de publicaciones y expectativas del público, constituyéndose en una herramienta estratégica para la gestión de la reputación digital y la toma de decisiones informadas.

2.2.6. Marco Conceptual

Tabla 1. Definiciones conceptuales del estudio.

Concepto	Definición	Fuente
Reputación digital	Imagen o percepción colectiva de una organización en entornos en línea, influenciada por la interacción de los usuarios.	García & Calderón (2021)
Análisis de sentimiento	Técnica del NLP que identifica la polaridad emocional (positiva, negativa o neutra) de un texto.	Medhat et al. (2014)
Topic Modeling	Método de minería de texto que identifica temas ocultos en documentos o publicaciones.	Blei, Ng & Jordan (2003)
Machine Learning	Rama de la inteligencia artificial que permite a los sistemas aprender patrones y realizar predicciones a partir de datos.	Mitchell (1997)
Reputación farmacéutica	Imagen pública de una empresa del sector salud, construida por las percepciones expresadas en medios digitales	Gómez (2023)

Fuente: Elaboración propia a partir de autores citados.

3. CAPÍTULO III: MARCO METODOLÓGICO

3.1. Tipo y enfoque de investigación

El estudio emplea un enfoque cuantitativo, mediante técnicas computacionales de análisis textual. Este enfoque permite procesar grandes volúmenes de datos proveniente de plataformas digitales y evaluar la percepción ciudadana.

El tipo de investigación es descriptivo y exploratorio, ya que busca caracterizar sentimientos y temas asociados a la experiencia de los usuarios con servicios de salud y acceso a medicamentos, sin manipular variables ni intervenir en el entorno digital.

3.2. Diseño de investigación (no experimental, transversal)

El diseño es no experimental, debido a que se analiza datos tal como fueron generados por los usuarios en sus interacciones digitales, sin modificación por parte del investigador.

El corte temporal es transversal, ya que la recolección de información se realizó en un único periodo determinado, lo que permite obtener una visión puntual del estado de la reputación digital durante la etapa analizada.

Este diseño es adecuado para estudios de reputación digital, ya que permite analizar percepciones ciudadanas sin alterar el comportamiento natural de los usuarios en entornos digitales.

3.3. Población y muestra (criterios de selección y tamaño)

La población está conformada por mensajes, comentarios y reseñas relacionados con servicios de salud, farmacias, instituciones públicas y aplicaciones del sector farmacéutico ecuatoriano.

La muestra fue obtenida de YouTube, Google Play Store y medios digitales, aplicando criterios de inclusión basados en palabras clave relacionadas con acceso a medicamentos, atención sanitaria y percepción ciudadana.

Después del proceso de limpieza y filtrado, la muestra total estuvo compuesta por 12.376 registros textuales. Para el entrenamiento y evaluación de los modelos de supervisados de análisis de sentimiento, se utilizó un subconjunto balanceado de 2.476 registros.

Esto permitió garantizar comparabilidad entre modelos y evitar sesgos derivados del desbalance de clases.

3.4. Técnicas e instrumentos de recolección de datos (APIs, snscrape, filtro de palabras clave)

Los datos se recopilaron mediante herramientas de extracción basados en Python. Para YouTube se desarrollaron scripts personalizados que permitieron obtener comentarios relevantes utilizando consultas temáticas.

Las reseñas de aplicación móviles se obtuvieron mediante la librería "Google-play-scraper", enfocada en aplicaciones como IESS App, SaludEc MSP, Fybeca, Cruz Azul y otras relacionadas con el sector.

En los medios digitales se emplearon scrapers adaptados a cada portal, junto con filtros de palabras clave para identificar contenido relacionado con disponibilidad de medicamentos, quejas, costos y experiencia de atención.

No se emplearon APIs oficiales de redes sociales debido a restricciones de acceso, optándose por técnicas de scraping ético y herramientas de recolección automatizada.

3.5. Procedimiento de análisis de datos (preprocesamiento, vectorización, modelado)

El procedimiento se desarrolló en tres fases consecutivas. La primera fase correspondió al preprocesamiento, donde los textos fueron normalizados, limpiados de caracteres especiales, depuración de duplicados y filtrados por longitud mínima para garantizar relevancia.

La segunda fase consistió en la vectorización del texto mediante *TF-IDF*, lo que permitió representar cada documento como un vector numérico adecuado para modelos de aprendizaje automático como *Naive Bayes* y *SVM*.

La fase final correspondió al modelado del análisis de sentimiento. Se entrenaron modelos supervisados tradicionales como *Naive Bayes*, Regresión Logística y Máquinas de Soporte Vectorial (*SVM*), utilizando representaciones *TF-IDF*.

Adicionalmente, se implementó un modelo de aprendizaje profundo basado en *BETO*, un transformer preentrenado para el idioma español, con el fin de comparar su desempeño frente a los modelos tradicionales.

3.5.1. Variable de pronóstico.

En la presente investigación, la variable de pronóstico corresponde al sentimiento expresado en las publicaciones digitales, clasificado en tres categorías: positivo, negativo y neutro.

Estas categorías constituyen la variable dependiente de tipo categórica, cuyo valor es pronosticado por los modelos *Machine Learning* a partir del contenido textual de las publicaciones.

El texto preprocesado de cada publicación constituye la variable independiente del estudio, ya que representa la fuente principal de información utilizada por los modelos de análisis. Este texto se transforma en representaciones numéricas por medio de técnicas de vectorización y embeddings.

A partir de estas representaciones, los algoritmos entrenados aprenden patrones lingüísticos asociados a cada tipo de sentimiento.

La variable de pronóstico corresponde a la categoría de sentimiento asignada a cada publicación. La capacidad predictiva de los modelos evalúa a través de métricas de desempeño como precisión, recall y F1-score, para identificar el mejor algoritmo de predicción del sentimiento.

De esta manera, la variable de pronóstico se vincula con la utilidad del modelo para interpretar la reputación digital del sector farmacéutico, al reflejar la orientación emocional predominante en las publicaciones analizadas.

3.6. Validación y métricas (precisión, recall, F1-score; coherencia de tópicos)

La validación del desempeño de los modelos se realizó mediante métricas estándar de clasificación binaria y multiclase, específicamente precisión, recall y F1-score. Estas métricas permiten evaluar la capacidad de los algoritmos para identificar patrones de sentimiento.

Dichas métricas fueron calculadas para cada uno de los modelos entrenados, incluye *Naive Bayes, SVM* y el modelo profundo *BETO*, con el fin de comparar su desempeño y analizar el equilibrio entre predicciones correctas y errores.

En el tema del modelado, la validación se llevó a cabo mediante el análisis de la coherencia semántica de los tópicos generados y la revisión manual de los términos predominantes en cada uno.

Esto permitió verificar que los temas fueran interpretables, consistentes y representativos del contenido del corpus analizado.

3.7. Procedimiento Metodológico

El procedimiento metodológico define de manera estructurada las acciones necesarias para cumplir los objetivos de la investigación y contrastar las hipótesis planteadas. El estudio adopta un enfoque cuantitativo sustentando en técnicas de NLP y *Machine Learning* aplicados al análisis textual.

El proceso se desarrolla de forma secuencial, iniciando con la recolección de datos y continua con etapas de limpieza, transformación y análisis del contenido textual. Cada una de estas fases responde a criterios metodológicos que garantizan validez y consistencia.

Las técnicas aplicadas permiten integrar los resultados del análisis de sentimiento y la detección de temas, proporcionando una visión integral de la reputación del sector farmacéutico y de salud de Ecuador.

3.7.1. Etapas del procedimiento

1. Definición de empresas, palabras clave y alcance del análisis.
2. Extracción de datos mediante scrapers y herramientas de recolección automatizada.
3. Limpieza, normalización y lematización del texto.
4. Aplicación de modelos de análisis de sentimiento (Naive Bayes, SVM y BETO).
5. Modelado de temas mediante LDA.
6. Integración e interpretación de resultados.
7. Redacción del análisis final y elaboración de visualizaciones.

3.8. Operacionalización de la Investigación

La operacionalización permite convertir los conceptos centrales de la investigación en elementos medibles y observables. Esto facilita analizar la reputación digital mediante técnicas de procesamiento de lenguaje natural.

Las variables del estudio se orientan al sentimiento digital y a los temas predominantes expresados en plataformas digitales. Estas dimensiones permiten identificar percepciones públicas sobre servicios, farmacias y aplicaciones asociadas al sector.

La operacionalización integra el marco teórico con el análisis empírico mediante técnicas cuantitativas basadas en texto. Esto permite revelar patrones de opinión, tendencias y emociones presentes en la conversación digital. Su aplicación asegura resultados coherentes, verificables y alineados con la hipótesis planteada.

Tabla 2. Operacionalización

VARIABLE	DIMENSIÓN	INDICADORES	TÉCNICA
Sentimiento digital	Positivo	Palabras o frases con valoración favorable	Análisis de sentimiento (Naive Bayes, SVM y BETO)
Sentimiento digital	Negativo	Quejas, rechazo o emociones adversas	Análisis de sentimiento (Naive Bayes, SVM y BETO)
Sentimiento digital	Neutro	Publicaciones informativas o sin emoción	Análisis de sentimiento (Naive Bayes, SVM y BETO)
Temas predominantes	Eficiencia	Frecuencia de términos sobre	Modelado de temas LDA

		rendimiento del medicamento	
Temas predominantes	Efectos secundarios	Menciones a reacciones adversas	LDA
Temas predominantes	Precio	Opiniones sobre costos o accesibilidad	LDA
Temas predominantes	Servicio	Comentarios sobre disponibilidad o atención	LDA

Fuente: *Elaboración propia.*

4. CAPÍTULO IV – DESARROLLO DEL MODELO Y ANÁLISIS

4.1. Recolección y descripción del dataset

El dataset utilizado en esta investigación fue construido a partir de la integración de datos recolectados desde múltiples plataformas digitales, relacionadas con el sector farmacéutico y de salud. Las fuentes incluyeron comentarios de YouTube, reseñas de aplicaciones móviles y publicaciones de portales de noticias digitales.

La recolección de información se realizó mediante scripts desarrollados en Python, aplicando técnicas de scraping y consumo de fuentes públicas. Todos los datos analizados corresponden a contenido generado por usuarios y disponible de forma abierta en la web.

El dataset consolidado estuvo conformado por 12376 registros, cada uno compuesto por texto original, texto procesado y una etiqueta de sentimiento.

Para el entrenamiento y evaluación de los modelos de análisis de sentimiento, se trabajó con un subconjunto balanceado de 2476 registros, con el fin de garantizar una comparación justa entre modelos y evitar sesgos por desbalance de clases.

La tabla presenta la distribución de sentimientos correspondientes al subconjunto utilizado para la evaluación de los modelos de clasificación.

Tabla 3. Distribución de sentimientos en el dataset.

Sentimiento	Cantidad	Porcentaje
Negativo	1051	42.5%
Neutro	1054	42.6%
Positivo	371	14.9%
Total	2476	100%

Fuente: *Elaboración propia*

4.2. Preprocesamiento y limpieza textual

El preprocesamiento de los datos textuales constituyó una fase clave para asegurar la calidad del análisis de sentimiento y la confiablidad de los modelos. En la etapa inicial, se eliminaron registros duplicados, valores nulos y publicaciones no útiles.

Posteriormente, se aplicaron procedimientos de normalización de texto, tales como conversión a minúsculas, la eliminación de signos de puntuación, caracteres especiales y espacios innecesarios.

Como resultado de este proceso, se generó un conjunto de textos depurados y estandarizados, almacenados en una columna específica del dataset final. Estos textos sirvieron como entrada directa para los modelos de *Machine Learning* y aprendizaje profundo utilizados en el estudio.

4.3. Entrenamiento y evaluación de modelos de análisis de sentimiento

Para el análisis de sentimiento se entrenaron cuatro modelos de clasificación: *Naive Bayes*, Regresión Logística, Máquinas de Soporte Vectorial (*SVM*) y un modelo basado en *BETO*, una variante de *BERT* entrenada para el idioma español.

Todos los modelos fueron evaluados bajo un mismo esquema de partición de datos, con el fin de garantizar condiciones comparables en el análisis de desempeño.

Los modelos tradicionales emplearon representaciones vectoriales basados en TF-IDF, mientras que el modelo *BETO* utilizó embeddings contextualizados obtenidos mediante un proceso de *fine-tuning* ligero ejecutado en entorno CPU.

La evaluación del rendimiento se realizó utilizando métricas estándar de clasificación, tales como precisión, recall, F1-score y exatitud.

Los resultados evidencian diferencias claras en el rendimiento de los modelos, destacando un mejor desempeño del enfoque basado en Transformers frente a los modelos tradicionales.

Tabla 4. Comparación del desempeño de los modelos de análisis de sentimiento.

Modelo	Accuracy	Precision (weighted)	Recall (weighted)	F1-score (weighted)
Naive Bayes	0.70	0.72	0.70	0.69
Regresión Logística	0.73	0.73	0.73	0.73
SVM	0.74	0.74	0.74	0.74
BERT (CPU)	0.76	0.76	0.76	0.76

Fuente: *Elaboración propia.*

Tabla 5. Interpretación comparativa del desempeño de los modelos de análisis de sentimiento.

Modelo	Fortaleza Principal	Limitación identificada
Naive Bayes	Alta rapidez y bajo costo computacional.	Confusión en clases minoritarias.
Regresión Logística	Buen equilibrio entre precisión y recall.	Sensible al desbalance de clases.
SVM	Mejor separación entre clases.	Mayor complejidad computacional.

BETO	Comprensión contextual del lenguaje.	Mayor tiempo de entrenamiento.

Fuente: *Elaboración propia.*

El modelo BERT obtuvo el mejor desempeño global en todas las métricas evaluadas, alcanzando una exactitud del 76%. Este resultado confirma la capacidad de los modelos basados en Transformers para capturar el contexto semántico del lenguaje natural en español.

Los modelos clásicos presentaron un desempeño competitivo, siendo SVM el más destacado entre ellos. No obstante, su rendimiento fue inferior al modelo BERT, especialmente en la clasificación equilibrada de las tres clases de sentimiento.

4.3.1. Análisis de la matriz de confusión

La matriz de confusión permitió analizar en detalle el comportamiento de los modelos de clasificación de sentimiento, evaluando la cantidad de predicciones correctas e incorrectas por cada clase (positivo, negativo y neutro). Esta herramienta facilita identificar patrones de error y fortaleza de cada algoritmo.

El análisis de la matriz de confusión del modelo *Naive Bayes* evidenció un favorable desempeño en la identificación de publicaciones con sentimiento negativo, la cual muestra una alta proporción de clasificaciones correctas en esta categoría.

Sin embargo, se observaron dificultades al diferenciar el sentimiento positivo, el cual fue frecuentemente confundido con la clase neutra.

Esta situación puede atribuirse al menor número de ejemplos positivos presentes en el conjunto de datos, lo que afectó la capacidad del modelo para aprender patrones distintivos asociados a esa clase.

En contraste, el modelo *SVM* presentó un comportamiento más equilibrado entre las tres categorías de sentimiento.

Finalmente, el modelo en *BERT* presentó la matriz de confusión más balanceado con un número de aciertos en todas las clases. Este resultado confirma que los modelos basados en Transformers logran capturar relaciones semánticas más profundas en el lenguaje reduciendo la ambigüedad entre sentimientos similares.

El análisis de las matrices de confusión complementa las métricas cuantitativas y permite una evaluación más detallada del desempeño real de los modelos en escenarios prácticos.

Este comportamiento evidencia que la matriz de confusión no solo permite evaluar exactitud global, sino también analizar la calidad de pronóstico por clase, aspecto clave en aplicaciones reales de monitoreo reputacional.

Tabla 6. Interpretación conceptual de la matriz de confusión aplicada al análisis de sentimiento.

Elemento	Interpretación en el estudio
Verdaderos positivos	Publicaciones correctamente clasificadas.
Falsos Positivos	Publicaciones clasificadas con mayor optimismo.
Falsos Negativos	Opiniones negativas no detectadas.
Confusión neutro-negativo	Ambigüedad semántica en textos informativos.

Fuente: *Elaboración propia.*

4.4. Implementación de *Topic Modeling* y etiquetado de tópicos

Con el objetivo de identificar los temas predominantes en la conversación digital, se aplicó modelado de tópicos mediante el algoritmo *Latent Dirichlet Allocation* (LDA). Este enfoque permitió descubrir patrones temáticos recurrentes dentro del corpus textual analizado.

Previo a la aplicación del modelado de temas por *LDA*, los textos fueron vectorizados y sometidos a un proceso de depuración para eliminar términos irrelevantes o escaso valor semántico. Esto permite mejorar la calidad de los tópicos generados.

Los tópicos identificados estuvieron relacionados principalmente con el desabastecimiento de medicamentos, precios, calidad de servicio y atención médica, reflejando las principales preocupaciones ciudadanas respecto al sector farmacéutico.

El número de tópicos fue definido de manera experimental, evaluando distintas configuraciones y selecciones entre cuatro y seis tópicos en función de la coherencia semántica y la interpretabilidad de los resultados.

4.5. Visualización y análisis temporal

Como complemento al análisis cuantitativo, se elaboraron visualizaciones gráficas con el propósito de facilitar la interpretación de los resultados obtenidos. Estas visualizaciones incluyeron representaciones de la distribución de sentimiento y comparaciones del desempeño entre los modelos.

Adicionalmente, se desarrolló un análisis temporal del sentimiento, lo que permitió observar variaciones en la percepción ciudadana a lo largo del periodo estudiado. Este enfoque aportó una perspectiva dinámica sobre la evaluación de la reputación.

Las visualizaciones generadas cumplen una función descriptiva y de apoyo interpretativo, sin construir pruebas estadísticas inferenciales, pero aportando claridad en la presentación de los hallazgos del estudio.

4.6. Limitaciones del modelo.

A pesar de los resultados obtenidos, el estudio presenta determinadas limitaciones metodológicas que deben ser consideradas. Una de ellas se relaciona con el uso de un

proceso de etiquetado automático inicial para la generación de las clases de sentimiento, esto puede introducir sesgos derivados del modelo de referencia.

Asimismo, los datos analizados provienen exclusivamente de plataformas digitales de acceso público, esto no representa de manera exhaustiva la totalidad de la opinión ciudadana.

Finalmente, el entrenamiento del modelo de *BETO* se realizó bajo restricciones computacionales, específicamente en entorno CPU, lo cual limitó la exploración de configuraciones más complejas. No obstante, los resultados obtenidos fueron consistentes y suficientes para cumplir los objetivos planteados en la investigación.

5. CAPÍTULO V – RESULTADOS Y DISCUSIÓN

5.1. Resultados cuantitativos: distribución de sentimientos y métricas de desempeño

El análisis cuantitativo de los datos evidenció una predominancia de sentimientos negativos y neutros en las publicaciones analizadas. Esta distribución refleja una percepción mayormente crítica o informativa por parte de los usuarios respecto al sector farmacéutico y de salud.

Constituyéndose como un indicador relevante de la reputación digital del sector en la provincia de Pichincha.

En relación con el desempeño de los modelos de análisis de sentimiento, se observó una mejora progresiva desde los enfoques tradicionales hacia modelos más avanzados.

El clasificador Naive Bayes presentó el menor rendimiento, mientras que los modelos Regresión Logística y Máquina de Soporte Vectorial *(SVM)* alcanzaron métricas más equilibradas y estables entre las tres clases de sentimiento.

El modelo BERT obtuvo el mejor desempeño global, alcanzando una exactitud del 76% y el mayor F1-score ponderado. Este resultado confirma la eficiencia de los modelos basados en Transformers para tareas de análisis de sentimiento en español.

El análisis de las matrices de confusión mostró que los principales errores de clasificación se produjeron entre las clases neutra y negativa, fenómeno asociado a la ambigüedad semántica presente en publicaciones informativas con carga crítica.

En comparación, el modelo *BERT* presentó una menor tasa de confusión entre clases, confirmando su superioridad en la tarea de pronóstico del sentimiento.

Tabla 7. Síntesis de resultados cuantitativos del análisis de sentimiento.

Aspecto analizado	Resultado principal
Sentimiento predominante	Negativo y neutro
Modelo con mejor desempeño	*BETO*
Métrica más alta	*F1-score* ponderado

Principal dificultad	Ambigüedad entre neutral y negativo

Fuente: *Elaboración propia.*

5.2. Resultados cualitativos: interpretación de tópicos e inferencias sobre reputación

El modelado de tópicos permitió identificar temáticas recurrentes relacionadas con el desabastecimiento de medicamentos, precios elevados, calidad del servicio y atención médica. Estos temas representan los principales factores que influyen en la reputación digital del sector farmacéutica.

La presencia constante de tópicos asociados a quejas y denuncias sugiere un nivel significativo de insatisfacción ciudadana. En contraste, los tópicos de carácter positivo estuvieron vinculados principalmente a experiencias puntuales de buena atención o disponibilidad de medicamentos.

En conjunto, los resultados cualitativos permiten inferir que la reputación digital del sector farmacéutico en Pichincha se encuentra condicionada por problemas estructurales del sistema de salud, los cuales son amplificados en el entorno digital.

5.3. Discusión comparativa con antecedentes

Los resultados obtenidos en esta investigación son consistentes con estudios previos que destacan la superioridad de los modelos basados en Transformers frente a algoritmos clásicos de *Machine Learning* para el análisis de sentimiento.

En particular, BERT demostró una mayor capacidad para capturar el contexto semántico del lenguaje natural.

Asimismo, la predominancia de sentimientos negativos y neutros coinciden con antecedentes en estudios de reputación digital en sectores sensibles como salud y servicios públicos. En el contexto dado, las redes suelen reflejar reclamos, denuncias y experiencias que condicionan la percepción general.

En comparación con investigaciones previas centradas en enfoques generales o en una sola fuente de datos, el presente estudio adopta una estrategia multifuente aplicada específicamente al sector farmacéutico ecuatoriano.

Esta delimitación contextual permite obtener resultados más precisos y directamente relacionados con la realidad digital del entorno analizado.

Asimismo, la integración del análisis de sentimiento con el modelado de temas amplía la compresión del fenómeno reputacional, al combinar la dimensión emocional de las publicaciones con los asuntos concretos que generan opinión pública.

Este enfoque complementario supera los análisis basados únicamente en polaridad, aportando una explicación más profunda del comportamiento discursivo de los usuarios.

Los hallazgos obtenidos respaldan la literatura existente que destaca el valor del análisis automatizado de publicaciones digitales como una herramienta válida para interpretar percepciones colectivas y apoyar procesos de toma de decisiones.

5.4. Implicaciones prácticas de los resultados.

Los resultados alcanzados presentan implicaciones prácticas relevantes para la gestión de la reputación digital en el sector farmacéutico y de la salud. El uso de técnicas automatizadas permite detectar oportunamente percepciones negativas o neutras, y al no ser corregidas, pueden escalar hacia escenarios de crisis reputacional.

De igual manera, la identificación de tópicos recurrentes, como problemas de abastecimiento, costos de medicamentos y calidad de atención, ofrece información concreta que puede ser útil para las instituciones para priorizar y mejorar su servicio.

Finalmente, la aplicación de modelos avanzados como *BETO* demuestra la viabilidad de implementar sistemas de monitoreo reputacional sensibles al contexto lingüísticos del español, estos pueden integrarse en procesos institucionales de análisis de opinión pública orientadas a una toma de decisiones más informada y centrada en las personas.

6. CAPÍTULO VI - CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones generales

A partir del desarrollo de la investigación, se determina que el análisis de la reputación digital del sector farmacéutico es factible mediante la aplicación de *Machine Learning* y Procesamiento de Lenguaje Natural sobre publicaciones de plataformas digitales.

La utilización de fuentes diversas permitió conformar un conjunto de datos amplio y representativo, adecuado para el análisis propuesto.

Los resultados cuantitativos evidenciaron una mayor presencia de sentimientos negativos y neutros en las publicaciones analizadas, lo que indica que la conversación digital se encuentra dominada por reclamos, observaciones críticas e información.

Esta tendencia se ascia principalmente a dificultades relacionadas con el acceso a medicamentos, el abastecimiento y la experiencia de atención en el sistema de salud.

En la evaluación comparativa de modelos, el enfoque basado en *BETO* obtuvo el mejor desempeño general frente a *Naive Bayes*, Regresión Logística y *SVM*, demostrando una mayor capacidad para capturar matices semánticos del idioma español.

De manera complementaria, el modelado de tópicos permitió identificar temas recurrentes vinculados a precios, disponibilidad de medicamentos y calidad del servicio, aportando una compresión más integral de los factores que afectan a la reputación.

6.2. Recomendaciones para la gestión reputacional del sector farmacéutico

Se recomienda a las instituciones del sector farmacéutico y de salud implementar sistemas de monitoreo continuo de reputación digital, apoyados en paneles de control que integren métricas de sentimientos y tópicos.

Esto facilitaría la identificación temprana de cambios en la percepción ciudadana y la detección oportuna de supuestos escenarios críticos en la conversación digital.

Asimismo, es conveniente priorizar acciones de gestión reputacional enfocadas en los temas más recurrentes identificados en el análisis, como la disponibilidad de medicamentos, los costos y la calidad de atención.

En este contexto, las estrategias de comunicación deberían centrarse en la transparencia, la claridad de la información y tiempos de respuesta adecuados frente a los reclamos ciudadanos.

Adicionalmente, se sugiere complementar los resultados del análisis automatizado con revisión humana en casos específicos, especialmente en publicaciones de ato impacto o con ambigüedad semántica.

Finalmente, la consolidación de un repositorio institucional en el cual almacenaría datos históricos y modelos entrenados permitiría replicar el análisis en distintos periodos, asegurando la mejora continua del sistema de monitoreo reputacional.

Por ejemplo, un incremento sostenido de sentimiento negativo asociado a un tópico específico puede activar protocolos de comunicación o revisión operativa antes de que la percepción pública se deteriore de forma significativa.

Del mismo modo, el análisis temático puede apoyar campañas informativas focalizadas, orientadas a aclarar procesos, tiempos de atención o políticas de abastecimiento, reduciendo la desinformación y fortaleciendo la confianza ciudadana.

Tabla 8. Relación entre resultados obtenidos y recomendaciones propuestas

Resultado identificado	Recomendación asociada
Alta presencia de sentimiento negativo	Monitoreo reputacional continuo.
Confusión entre neutral y negativo	Validación humana complementaria.
Tópicos de desabastecimiento	Mejora de comunicación institucional.
Desempeño superior a *BETO*	Uso de modelos contextualizados.

Fuente: *Elaboración propia.*

6.3. Líneas de investigación futuras

Como trabajo de futuro, se propone ampliar la recolección hacia nuevas plataformas digitales y periodos más extensos, con el fin de analizar variaciones estacionales y eventos coyunturales. Esto permitirá robustecer el análisis temporal de reputación.

Otra proyección a futura se propone evaluar otros modelos especializados en el idioma español, así como la aplicación de técnicas de balanceo de clases para mejorar la detección de sentimientos positivos.

También, incorporar análisis de emociones o detección de intención comunicacional, con el fin de diferenciar de manera más precisa entre reclamos, denuncias, consultas y solicitudes de información.

Finalmente, se recomienda explorar el uso de métricas avanzadas de coherencia temática y métodos automáticos para la optimización del número de tópicos en el modelo LDA, esto contribuiría a mejorar la estabilidad, interpretabilidad y aplicabilidad del análisis temático en contextos reales.

Adicionalmente, futuras investigaciones podrían explorar la incorporación técnicas de aprendizaje semi-supervisado o modelos generativos para mejorar la calidad del etiquetado automático y reducir la dependencia de modelos preentrenados.

Esto permitiría adaptar los sistemas de análisis de sentimiento a contextos locales específicos y a variaciones lingüísticas propias del entorno ecuatoriano.

7. BIBLIOGRAFÍA (formato APA)

- Aula, P. (2010). Social media, reputation risk and ambient publicity management. *Strategy & Leadership, 38*(6), 43–49. https://doi.org/10.1108/10878571011088069
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.
- Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2017). Sentiment analysis: Beyond positive and negative. *IEEE Intelligent Systems, 32*(2), 74–80. https://doi.org/10.1109/MIS.2017.23
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://doi.org/10.18653/v1/N19-1423
- Díaz, L., & Herrera, P. (2020). Integración de técnicas de minería de texto para análisis de reputación digital. *Revista Latinoamericana de Tecnología, 15*(2), 56–67.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

- Joachims, T. (2002). Learning to classify text using support vector machines. Kluwer Academic Publishers.
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003
- Kim, J., Lee, S., & Park, H. (2019). Corporate social media interaction and reputation perception. *Journal of Communication Management, 23*(3), 216–233. https://doi.org/10.1108/JCOM-01-2019-0015
- Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
- López, D., & Ruiz, V. (2022). Aplicación de análisis de sentimiento en redes sociales farmacéuticas. *Revista Ecuatoriana de Tecnología, 9*(1), 25–33.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
- Wang, Y., Zhang, Z., & Li, F. (2020). Topic modeling on social media for brand reputation analysis. *Information Processing & Management, 57*(6), 102115. https://doi.org/10.1016/j.ipm.2020.102115
- Zhang, L., Wang, S., & Liu, B. (2021). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11*(3), e1356. https://doi.org/10.1002/widm.1356

8. ANEXOS (scripts, tablas, evidencias de ejecución y resultados)

8.1. Anexo A: Recolección de datos desde YouTube

En el presente anexo se describen los scripts desarrollados en lenguaje Python para la recolección automatizada de datos desde la plataforma YouTube. El objetivo de estos scripts fue identificar videos y extraer comentarios públicos relacionados con el sector farmacéutico y de salud en Ecuador.

A partir de palabras clave representativas de problemáticas ciudadanas como desabastecimiento de medicamentos, precios y calidad de atención.

La recolección se realizó exclusivamente sobre contenido público, sin acceder a información privada ni requerir autenticación, respetando criterios éticos y de uso responsable de la información disponible en plataformas digitales.

8.1.1. Identificación de videos relevantes en YouTube.

Para la identificación de videos relevantes se utilizó la librería *youtubesearchpython*, la cual permite realizar búsquedas automatizadas a partir de consultas textuales.

Las búsquedas se basaron en un conjunto de palabras clave previamente definidas, relacionadas con el acceso a medicamentos, servicios de salud y percepción ciudadana en el contexto ecuatoriano.

El resultado de este proceso fue un listado de URLs de videos potencialmente relevantes, que posteriormente sirvieron como insumo para la extracción de comentarios.

Palabras clave utilizadas

```
KEYWORDS = [
    "no hay medicinas Ecuador",
    "desabastecimiento medicamentos Ecuador",
    "IESS medicinas",
    "farmacias Ecuador quejas",
    "medicinas caras Ecuador",
    "hospital Ecuador denuncias",
    "precios medicinas Ecuador"
]
```

Fragmento representativo del proceso de búsqueda

```
search = VideosSearch(query, limit=20)
results = search.result()["result"]


rows.append({
    "keyword": q,
    "title": r.get("title"),
    "url": r.get("link")
})
```

Este procedimiento permitió construir un archivo estructurado con los títulos y enlaces de los videos identificados, garantizando que el contenido recolectado estuviera alineado con el objetivo del estudio.

8.1.2. Extracción de comentarios públicos de YouTube.

Una vez identificados los videos relevantes, se procedió a la extracción de comentarios utilizando la herramienta *yt-dlp*, la cual permite acceder a los comentarios públicos asociados a cada video sin necesidad de descargar el contenido audiovisual.

Este proceso permitió recolectar opiniones expresadas por los usuarios, las cuales constituyen una fuente clave para el análisis de sentimiento y la evaluación de la reputación digital del sector farmacéutico.

Configuración general del proceso de extracción.

```
cmd = [
```

```
    "yt-dlp",
    "-J",
    "--write-comments",
    "--skip-download",
    url
]
```

Estructura de los datos recolectados.

```
registros.append({
    "video_url": url,
    "video_title": video_title,
    "author": c.get("author"),
    "text": c.get("text"),
    "like_count": c.get("like_count"),
    "timestamp": c.get("timestamp")
})
```

Los comentarios extraídos fueron almacenados en un archivo CSV, permitiendo su posterior limpieza, normalización y análisis automático mediante técnicas de *Machine Learning* y Procesamiento de Lenguaje Natural.

8.1.3. Recolección de reseñas Google Play Store

Para la recolección de reseñas de aplicaciones móviles se utilizó la librería *google-play-scraper*. Esta herramienta permitió extraer opiniones públicas de usuarios sobre aplicaciones relacionadas con servicios de salud y farmacias, tales como aplicaciones institucionales y privadas del sector farmacéutico.

Las reseñas obtenidas reflejan experiencias directas de los usuarios respecto al acceso a medicamentos, funcionamiento de aplicaciones y calidad del servicio. Cada reseño recolectada incluyó el texto de opinión, la calificación otorgada por el usuario y la fecha de publicación.

Aplicaciones analizadas

- IESS App
- SaludEc MSP
- Fybeca
- Cruz Azul

- SanaSana

Fragmento representativo del proceso de extracción.

```
reviews(
    app_id,
    lang="es",
    country="ec",
    sort=Sort.NEWEST,
    count=200
)
```

8.1.4. Recolección de publicaciones desde medios digitales.

Con el fin de complementar la información generada por usuarios, se recolectaron publicaciones informativas desde portales de noticias digitales ecuatorianos. Para ello se desarrollaron scrapers personalizados en Python, adaptados a la estructura de cada medio.

La recolección se enfocó en noticias relacionadas con disponibilidad de medicamentos, sistemas de salud, precios y denuncias ciudadanas. Las publicaciones obtenidas permitieron incorporar un contexto informativo que complementa la percepción ciudadana expresada en redes sociales y reseñas.

Fragmento representativo del proceso de scraping

```
for url in urls_medios:
    response = requests.get(url, headers=headers)
    soup = BeautifulSoup(response.text, "html.parser")
```

8.1.5. Consideraciones éticas de la recolección

La recolección de datos se limitó exclusivamente a contenido público disponible en YouTube. No se recolectó información privada no se realizó interacción directa con usuarios.

Los datos obtenidos fueron utilizados únicamente con fines académicos y de investigación, respetando principios de anonimización y uso responsable de la información.

8.1.6. Cierre del Anexo A

Los scripts presentados en este anexo permitieron recopilar información relevante desde la plataforma YouTube, constituyendo una base inicial de datos para el estudio. Estos registros fueron posteriormente integrados con datos de otras fuentes digitales, formando parte del conjunto de datos final utilizado en la investigación.

8.2. Anexo B: Preprocesamiento y limpieza de los datos textuales.

En este anexo se describen los procedimientos y herramientas empleadas para el preprocesamiento y la limpieza de los datos textuales obtenidos desde diversas plataformas digitales.

Estas etapas resultaron esenciales para asegurar la calidad del análisis de sentimiento y del modelado de temas, reduciendo ruido y estandarizando la información textual.

El preprocesamiento fue aplicado de forma consistente a todas las fuentes, permitiendo integrar los datos en un conjunto homogéneo y adecuado para su análisis mediante técnicas de *Machine Learning* y Procesamiento de Lenguaje Natural.

8.2.1. Eliminación de registros no válidos y duplicados

Como etapa inicial, se realizó una depuración del conjunto de datos con el propósito de eliminar registros duplicados, valores nulos y textos con longitud insuficiente para el análisis semántico.

Este proceso permitió conservar únicamente publicaciones con contenido relevante, mejorando la consistencia del dataset y reduciendo posibles sesgos derivados de información redundante.

Fragmento representativo del filtrado inicial.

```
df = df.dropna(subset=["texto"])

df = df.drop_duplicates(subset=["texto"])

df = df[df["texto"].str.len() > 3]
```

8.2.2. Normalización y limpieza de texto

Posteriormente, se aplicaron técnicas de normalización textual orientadas a estandarizar el contenido de las publicaciones. Estas incluyeron la conversación del texto a minúsculas, la eliminación de signos de puntuación, caracteres especiales y espacio innecesarios.

La normalización facilitó la reducción de variaciones superficiales del lenguaje, favoreciendo la identificación de patrones semánticos por parte de los modelos.

Fragmento representativo del proceso de limpieza.

```
texto = texto.lower()

texto = re.sub(r"[^\w\s]", "", texto)

texto = re.sub(r"\s+", " ", texto).strip()
```

8.2.3. Eliminación de palabras vacías y tokenización.

Con el objetivo de reducir ruido y mejorar el texto, se procedió a la eliminación de palabras vacías en idioma español. Posteriormente, el texto fue tokenizado para su análisis.

Este paso permitió conservar únicamente términos con valor informativo para el análisis de sentimiento y la detección de temas.

Fragmento representativo de eliminación de stopwords.

```
tokens = [

   word for word in texto.split()

   if word not in stopwords_es

]
```

8.2.4. Generación de la columna de texto limpio

Como resultado del preprocesamiento, se creó una columna específica denominada *texto_limpio,* la cual contiene el texto normalizado y depurado. Esta columna fue utilizada como entrada principal para los modelos de análisis de sentimiento y modelado de temas.

La separación entre texto original y texto limpio permitió mantener trazabilidad entre los datos crudos y los datos procesados.

Ejemplo de asignación del texto limpio.

```
df["texto_limpio"] = df["texto"].apply(limpiar_texto)
```

8.2.5. Resultado del proceso de preprocesamiento

Luego de aplicar todas las etapas de limpieza y normalización, se obtuvo un dataset final compuesto por 12.376 registros válidos. Este conjunto de datos presentó un formato homogéneo y una calidad adecuada para su posterior análisis.

El dataset preprocesado fue almacenado en un archivo CSV y utilizado como insumo para el entrenamiento y evaluación de los modelos de análisis de sentimiento de temas.

8.2.6. Cierre del Anexo B.

Los procedimientos descritos en este anexo permitieron transformar datos textuales sin procesar en un conjunto estructurado y analizable. El procesamiento

aplicado garantizó coherencia textual, reducción de ruido y una base sólida para los análisis posteriores desarrollados en la investigación.

8.3. Anexo C: Construcción del dataset final para el análisis de sentimiento y detección de temas.

Este anexo detalla el proceso de integración, consolidación y preparación del dataset final empleado en la investigación.

A partir de los datos recolectados desde múltiples plataformas digitales, se desarrolló un procedimiento sistemático para unificar las fuentes, estandarizar la estructura de los registros.

El contenido de este anexo complementa lo descrito en los Anexo A y B, explicando cómo los datos extraídos y preprocesados fueron integrados en un corpus único y representativo del entorno digital del sector farmacéutico.

8.3.1. Integración de fuentes de datos.

Los datos utilizados en el estudio provinieron de diversas plataformas digitales, incluyendo:

- Comentarios de videos de YouTube relacionados con el sector farmacéutico y salud.
- Reseñas de aplicaciones móviles del sector sanitario.
- Publicaciones y comentarios extraídos de portales de noticias digitales.

Cada fuente generó archivos independientes en formato CSV, los cuales fueron integrados en un solo conjunto de datos mediante scripts desarrollados en Python.

8.3.2. Unificación de estructura y campos

Antes de la consolidación, los datasets presentaban estructuras heterogéneas, por lo que se optó unificar los campos para asegurar consistencia. Se estandarizaron las siguientes columnas principales:

- Texto original *(texto)*
- Texto preprocesado *(texto_limpio)*
- Fuente de origen *(fuente)*
- Fecha o referencia temporal
- Identificador del registro

Este proceso permitió combinar los datos sin pérdida de información relevante.

Fragmento representativo de unificación de columnas.

```
df["texto"] = df["texto"].astype(str)

df["fuente"] = fuente_origen
```

8.3.3. Consolidación del dataset final.

Una vez estandarizadas las estructuras, los datasets individuales fueron concatenados en un único DataFrame. Posteriormente, se aplicaron validaciones finales para eliminar registros duplicados entre fuentes y asegurar la calidad del corpus.

Este procedimiento garantizó que cada publicación fuera única dentro del dataset final.

Fragmento representativo de consolidación.

```
df_final = pd.concat([df_youtube, df_apps, df_noticias], ignore_index=True)

df_final = df_final.drop_duplicates(subset=["texto_limpio"])
```

8.3.4. Generación de etiquetas de sentimiento.

Para la construcción del dataset supervisado, se generaron etiquetas de sentimiento (negativo, neutro y positivo) mediante el uso de un modelo preentrenado en idioma español. Estas etiquetas sirvieron como referencia para el entrenamiento y evolución de los modelos clásicos de *Machine Learning*.

La utilización de etiquetado automático permitió escalar el análisis sin necesidad de anotación manual, manteniendo consistencia en la clasificación.

Ejemplo representativo de etiquetado automático.

```
df_final["sentimiento"] = df_final["prediccion_sentimiento"]
```

8.3.5. Descripción del dataset final.

Como resultado del proceso de integración y depuración, se obtuvo un dataset final compuesto por 12.376 registros válidos, distribuidos entre las tres clases de sentimiento analizado.

Cada registro contiene tanto el texto original como su versión preprocesada, lo que permite asegurar trazabilidad y reproducibilidad del análisis. Este dataset fue almacenado en formato CSV y utilizado como insumo principal para:

- Entrenamiento de modelos de análisis de sentimiento.
- Evaluación comparativa de algoritmos.
- Aplicación de técnicas de *Topic Modeling*.
- Visualización y análisis temporal.

8.3.6. Importancia del dataset construido.

La construcción del dataset final representa un aporte relevante de esta investigación, al integrar información proveniente de múltiples plataformas digitales del contexto ecuatoriano en un único corpus estructurado.

Este enfoque multifuente fortalece la validez de los resultados y permite una interpretación más completa de la reputación digital del sector farmacéutico.

8.3.7. Cierre del Anexo C.

El proceso descrito en este anexo permitió transformar datos disperos y heterogéneos en un conjunto consolidado y confiable. La metodología aplicada asegura la reproducibilidad del estudio y establece una base sólida para futuras investigaciones relacionadas con el análisis de sentimiento y reputación digital en el sector salud.

8.4. Anexo D: Entrenamiento, evaluación y validación de los modelos de análisis de sentimiento.

Este anexo documenta de manera detallada el proceso de entrenamiento, evaluación y validación de los modelos de análisis de sentimiento implementados en la investigación.

Se describen los algoritmos utilizados, las técnicas de representación textual, las métricas de desempeño empleadas y la interpretación de la matriz de confusión.

Con el objetivo de garantizar transparencia, reproducibilidad y rigor metodológico, este anexo complementa los resultados presentados en el Capítulo IV, proporcionado evidencia técnica del funcionamiento de los modelos desarrollados.

8.4.1. Modelos de análisis de sentimiento implementados.

Para el análisis de sentimiento se utilizaron cuatro enfoques:

- *Naive Bayes* (NB).
- Regresión Logística (RL).
- Máquinas de Soporte Vectorial (SVM).
- Modelo basado en Transformers (*BERT/BETO).*

Los modelos tradicionales (NB, RL y SVM) fueron entrenados utilizando representaciones *TF-IDF*, mientras que el modelo *BERT* empleó *embeddings* contextualizados mediante *fine-tuning* en CPU.

8.4.2. Preparación de los datos para el entrenamiento.

El dataset consolidado fue dividido en conjuntos de entrenamiento y prueba, utilizando una proporción del 80% para entrenamiento y 20% para evaluación. Esta partición permitió validar el desempeño de los modelos sobre datos no vistos previamente.

Previo al entrenamiento, los textos fueron vectorizados según el enfoque del modelo:

- *TF-IDF* para NB, RL y SVM.
- Tokenización y *embeddings* contextuales para *BERT*.

8.4.3. Entrenamiento de modelos clásicos.

Los modelos *Naive Bayes,* Regresión Logística y SVM fueron entrenados con parámetros estándar optimizados para clasificación multiclase. La elección de estos modelos se basó en su uso extendido en tareas de análisis de sentimiento y su eficiencia computacional.

Estos modelos permitieron establecer una línea base de comparación frente al enfoque basado en *Deep Learning*.

Fragmento representativo de entrenamiento (modelo clásico).

```
model.fit(X_train_vec, y_train)
```

8.4.4. Entrenamiento del modelo *BERT*.

El modelo *BERT* fue ajustado mediante *fine-tuning* ligero utilizando un conjunto reducido de épocas, debido a limitaciones computacionales. A pesar de ello, el modelo logró capturar relaciones semánticas complejas y contexto bidireccional en el idioma español.

Durante el entrenamiento se monitorearon métricas de pérdida (*los)* y estabilidad del gradiente, garantizando la convergencia del modelo.

8.4.5. Evaluación del desempeño de los modelos.

La evaluación se realizó mediante métricas estándar de clasificación multiclase:

- Precisión *(Precision).*
- Exhaustividad *(Recall).*
- *F1-score.*
- Exactitud *(Accuracy).*

Estas métricas permitieron comparar objetivamente el rendimiento de los distintos modelos.

Los resultados evidenciaron que el modelo *BERT* obtuvo el mejor desempeño global, seguido por SVM y Regresión Logística, mientras que *Naive Bayes* presentó el menor rendimiento.

8.4.6. Interpretación de la matriz de confusión.

La matriz de confusión fue utilizada para el análisis el comportamiento de los modelos en cada clase de sentimiento. Esta herramienta permitió identificar:

- **Verdaderos positivos:** Publicaciones correctamente clasificadas.
- **Falsos Positivo:** Publicaciones clasificadas incorrectamente en una clase.

- **Falsos negativos:** Publicaciones no detectadas correctamente por el modelo.

El análisis mostró que los modelos presentan mayor confusión entre las clases negativa y neutra, lo cual es consistente con la naturaleza del lenguaje en redes sociales, donde muchas publicaciones informativas contienen carga emocional implícita.

El modelo *BERT* redujo significativamente esta confusión, logrando una mejor separación entre clases gracias a su comprensión contextual del texto.

8.4.7. Validación y confiabilidad del modelo.

La consistencia de los resultados obtenidos, junto con el uso de métricas robustas y matrices de confusión, permite afirmar que los modelos entrenados son confiables para el análisis de sentimiento en el contexto del sector farmacéutico.

La validación cruzada implícita mediante separación entrenamiento-prueba asegura que los modelos no están sobreajustados y generalizan adecuadamente sobre nuevos datos.

8.4.8. Relevancia del Anexo D.

Este anexo proporciona evidencia técnica que respalda los resultados presentados en el cuerpo principal del documento. La inclusión de métricas, matrices de confusión e interpretación fortalece la validez científica del estudio y responde a criterios de evaluación académica exigidos en el trabajo.

8.4.9. Cierre del Anexo D.

El proceso de entrenamiento y evaluación descrito en este anexo demuestra que las aplicaciones de técnicas de *Machine Learning* y *Deep Learning* es efectiva para analizar sentimientos en publicaciones digitales.

Los resultados obtenidos validan el cumplimiento de los objetivos planteados y sustentan las conclusiones de la investigación.