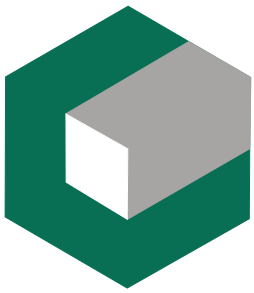

Universidad Tecnológica de La Habana "José Antonio Echeverría" CUJAE

Facultad de Informática

Ánalisis de Texto para comentarios de servicios hoteleros

Proyecto de Curso de TTC-IA

Autores:

Johny A. Pedraza Romero

Fernando José Bernal Suaréz

Gabriela Lucía Rosado León

4 de julio de 2025

Resumen

Palabras claves: ...

Abstract

Keywords: ...

Introducción

La hotelería, como rubro económico, ha evolucionado desde sus orígenes en las posadas medievales hasta convertirse en un sector estratégico en la economía global, particularmente en destinos turísticos como Cuba; donde su desarrollo histórico refleja la adaptación a demandas cambiantes, desde la provisión básica de alojamiento hasta la personalización de servicios bajo estándares internacionales. En este contexto, la competitividad en la industria hotelera no solo depende de la calidad de la infraestructura, sino también de la capacidad para anticipar y satisfacer las expectativas intangibles del cliente, como la experiencia emocional o la percepción de valor. Sin embargo, factores económicos como fluctuaciones monetarias, naturales como eventos climáticos extremos y los cambios en el comportamiento del consumidor; impulsados por la digitalización y la sostenibilidad, han intensificado los desafíos para mantener la rentabilidad y la reputación en este mercado tan saturado.

En el caso concreto de Cuba, la hotelería enfrenta dificultades estructurales derivadas de limitaciones tecnológicas, escasez de recursos y una infraestructura envejecida, contrastando con fortalezas como la hospitalidad tradicional y la riqueza cultural de este destino. Estudios recientes destacan que, aunque existen esfuerzos por incorporar innovaciones en la gestión, persisten barreras como la rigidez de los modelos administrativos y la falta de integración con sistemas de inteligencia de mercado. La dependencia de alianzas estratégicas con cadenas internacionales ha permitido cierta modernización, pero no resuelve problemas críticos, como la baja eficiencia en la respuesta a las necesidades emergentes de los huéspedes. El impacto de esta realidad se evidencia en como las tensiones se agravan en un entorno donde la percepción del cliente, mediatizada por plataformas digitales como TripAdvisor, define gran parte del éxito comercial.

Detectar las variables que impactan en la experiencia del cliente representa un reto central para la dirección hotelera. Las reseñas en línea, aunque ricas en información cualitativa, suelen ser analizadas de manera superficial, limitando su utilidad para decisiones estratégicas. Los métodos tradicionales de evaluación, como encuestas estructuradas, omiten matices contextuales presentes en comentarios textuales, mientras que, por el contrario, técnicas avanzadas de minería de textos permiten sistematizar estas percepciones. Ejemplo de esto son los modelos probabilísticos, que han demostrado eficacia en la identificación de patrones temáticos recurrentes, vinculándolos a dimensiones críticas de la cadena de valor hotelera. De igual forma resulta relevante integrar estas herramientas en flujos analíticos escalables a fin de ofrecer una solución para transformar datos no estructurados en indicadores operativos, facilitando así intervenciones proactivas.

Con esta base, se hace evidente que la fragmentación entre la recolección de re-

señas y su análisis sistemático impide que las organizaciones hoteleras identifiquen oportunidades de mejora con precisión y rapidez. A pesar de la disponibilidad de técnicas de procesamiento del lenguaje natural (PLN), su aplicación en contextos como el cubano es limitada, generando brechas en la correlación entre percepciones del cliente y ajustes operativos.

Es por ello que el presente trabajo pretende abordar el siguiente problema: ¿Cómo pueden las técnicas de minería de textos optimizar la extracción de conocimiento desde reseñas en plataformas digitales para identificar áreas de intervención en la gestión hotelera?

Descripción del Dataset

0.1. Origen del conjunto de datos

El conjunto de datos analizado surge de la concatenación de dos documentos obtenidos mediante la extracción de información procedente de la plataforma TripAdvisor, específicamente de reseñas asociadas a establecimientos hoteleros en Cuba. La recolección incluye registros publicados entre 2022 y 2025, abarcando un periodo que permite observar tendencias y variaciones en la percepción de los usuarios respecto a aspectos como servicios, infraestructura y atención. La naturaleza de los datos refleja una perspectiva heterogénea, ya que integra tanto valoraciones positivas como negativas, junto con respuestas formales de los hoteles a las críticas o elogios recibidos.

0.2. Características del conjunto de datos

La estructura del dataset combina variables cualitativas y cuantitativas, presentando un formato multimodal que incluye texto libre (reseñas, respuestas institucionales), fechas codificadas en formato numérico, calificaciones en escala ordinal (1 a 5 estrellas) y metadatos relacionados con los usuarios (identificadores anónimos, imágenes adjuntas). Cada registro contiene información sobre la experiencia del huésped, con descripciones detalladas de aspectos como condiciones higiénicas, calidad del servicio, infraestructura física, gestión de quejas y características específicas de las habitaciones o zonas comunes. Además, se observa la presencia de datos incompletos o inconsistentes en algunos campos, como enlaces rotos a imágenes o respuestas genéricas de los hoteles, además de la existencia de reseñas en idiomas distintos al español, lo que implica la necesidad de un proceso de limpieza previo a su análisis.

0.3. Tamaño y dimensionalidad del conjunto de datos

El dataset consta de 1000 filas, resultado de la unión de dos archivos con 500 registros cada uno. Cada fila representa una reseña individual, aunque en algunos casos se registran múltiples comentarios asociados a la misma estanciacon un conjunto de atributos que permiten análisis multidimensionales. Entre las columnas esenciales se destacan: Texto de la Reseña (variable textual con longitud variable, promedio de 250 palabras), Calificación Numérica (escala ordinal del 1 al 5), Fecha de Publicación (formato ISO 8601), Identificador del Autor (anónimo), Presencia de Imágenes (indicador binario) y Respuesta del Hotel (texto libre o vacío). La dimensionalidad total incluye al menos 15 atributos explícitos, aunque algunos presentan valores faltantes o

inconsistencias, como enlaces rotos a imágenes o respuestas genéricas de los establecimientos. La diversidad léxica y semántica de las reseñas permite inferir patrones de satisfacción e identificar factores recurrentes de insatisfacción, útil en las tareas asociadas a la minería de textos.

1. Metodología

1.1. Herramienta Utilizada: KNIME

KNIME (Konstanz Information Miner) es una plataforma de código abierto diseñada para el análisis de datos, la minería de datos y el procesamiento de información mediante un entorno visual intuitivo. Esta herramienta permite construir flujos de trabajo (*workflows*) de manera gráfica mediante la conexión de nodos predefinidos que realizan funciones específicas, lo cual facilita la integración, transformación y modelado de datos sin necesidad de escribir grandes bloques de código.

Una de las características más destacadas de KNIME es su arquitectura modular, que permite al usuario combinar múltiples técnicas analíticas, desde operaciones básicas de limpieza de datos hasta modelos avanzados de aprendizaje automático e inteligencia artificial. Además, KNIME ofrece compatibilidad con lenguajes de programación como Python y R, lo que amplía considerablemente su versatilidad y funcionalidad para usuarios avanzados.

En el contexto de este proyecto, KNIME se utiliza como herramienta central para el desarrollo de un flujo de trabajo de minería de texto, debido a su capacidad para integrar diversas tareas —como preprocesamiento, modelado de temas, minería de redes, análisis de sentimientos y muchos más— en un solo entorno visual. Esto no solo mejora la legibilidad del proceso, sino que también facilita la depuración, reutilización y documentación del flujo de trabajo.

Entre las ventajas clave de utilizar KNIME se encuentran:

Interfaz visual intuitiva: Permite diseñar flujos de trabajo mediante arrastrar y soltar, lo que reduce la curva de aprendizaje, especialmente para usuarios no expertos en programación.

Amplia biblioteca de nodos: Cuenta con una gran cantidad de nodos preconfigurados para tareas comunes en minería de texto, aprendizaje automático y visualización de datos.

Extensibilidad: Soporta la integración con lenguajes de programación como Python, R o Java, permitiendo personalizar operaciones complejas cuando sea necesario.

Reproducibilidad y colaboración: Los workflows pueden guardarse, compartirse y ejecutarse repetidamente, lo cual es fundamental para garantizar la reproducibilidad científica.

1.2. Workflow de Ánalisis de Texto

El siguiente *workflow* en KNIME fue diseñado para realizar una serie de tareas relacionadas con la minería de texto para el problema presentando, incluyendo preprocesamiento, modelado de temas, minería de redes y análisis de sentimientos. La figura 1 muestra la estructura completa del workflow, dividido en cuatro secciones principales: (1) Lectura y Preprocesamiento, (2) Modelado de Temas, (3) Minería de Redes y (4) Análisis de Sentimientos. A continuación, se describe cada componente y su función en detalle.

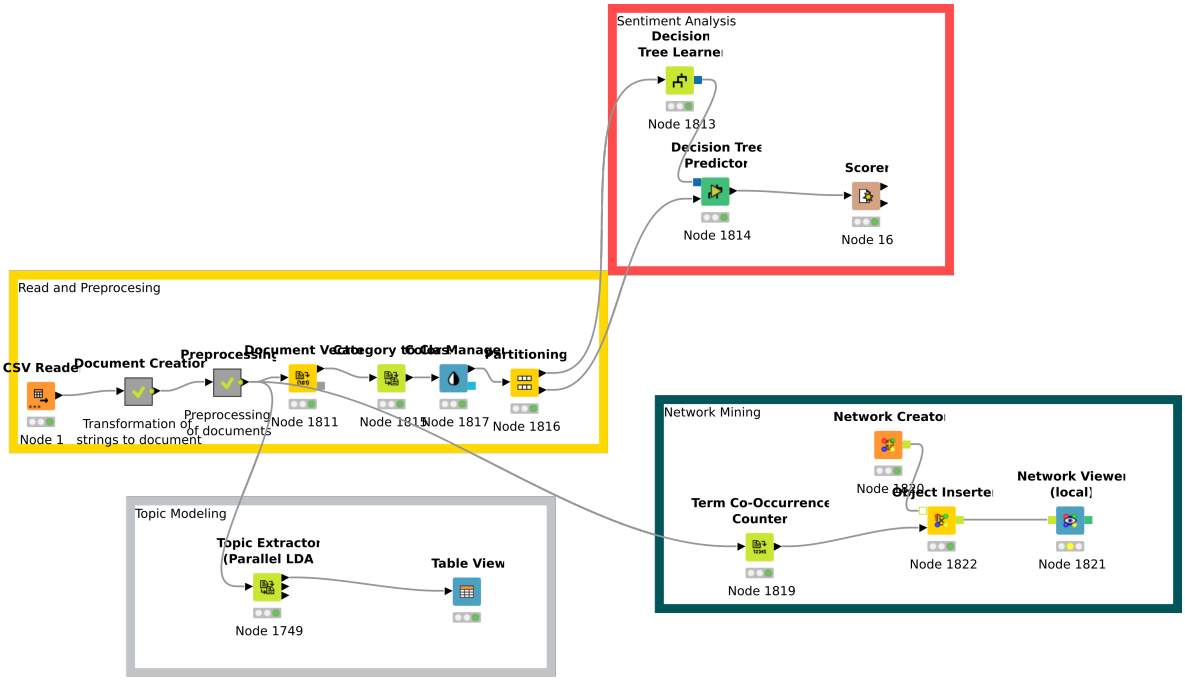

Figura 1: Workflow de KNIME para Minería de Texto

1.2.1. Lectura y Preprocesamiento

La primera etapa del workflow está dedicada a la lectura de datos y el preprocesamiento inicial de los textos. Comienza con el nodo **CSV Reader**, que lee un archivo CSV que contiene los documentos de texto brutos. Este nodo es fundamental para cargar los datos iniciales en el flujo de trabajo.

Una vez cargados los datos, el metanodo **Document Creator** transforma las cadenas de texto brutas en objetos de documento estructurados, lo que facilita su procesamiento posterior. Este paso es crucial para preparar los datos para la aplicación de las técnicas posteriores. La figura 2 muestra el flujo utilizado para convertir a documento los textos

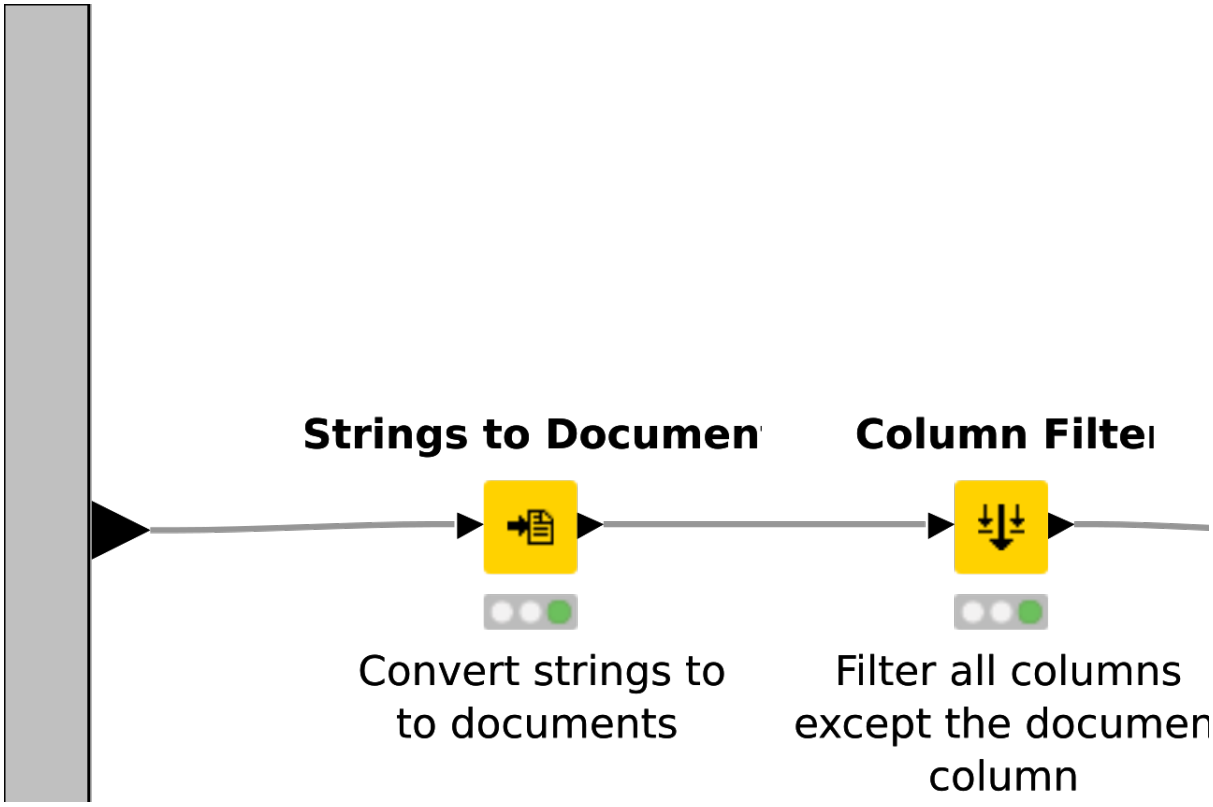

Figura 2:

El preprocesamiento es realizado por el metanodo *preprocessing* y constituye una etapa crucial en la minería de texto, ya que mejora la calidad y la consistencia de los datos antes de aplicar técnicas avanzadas de análisis. En este workflow, el preprocesamiento se divide en varias fases, cada una de las cuales realiza una tarea específica. A continuación, se explica cada nodo y su función en detalle.

La figura 3 muestra el flujo de preprocesamiento utilizado en este proyecto.

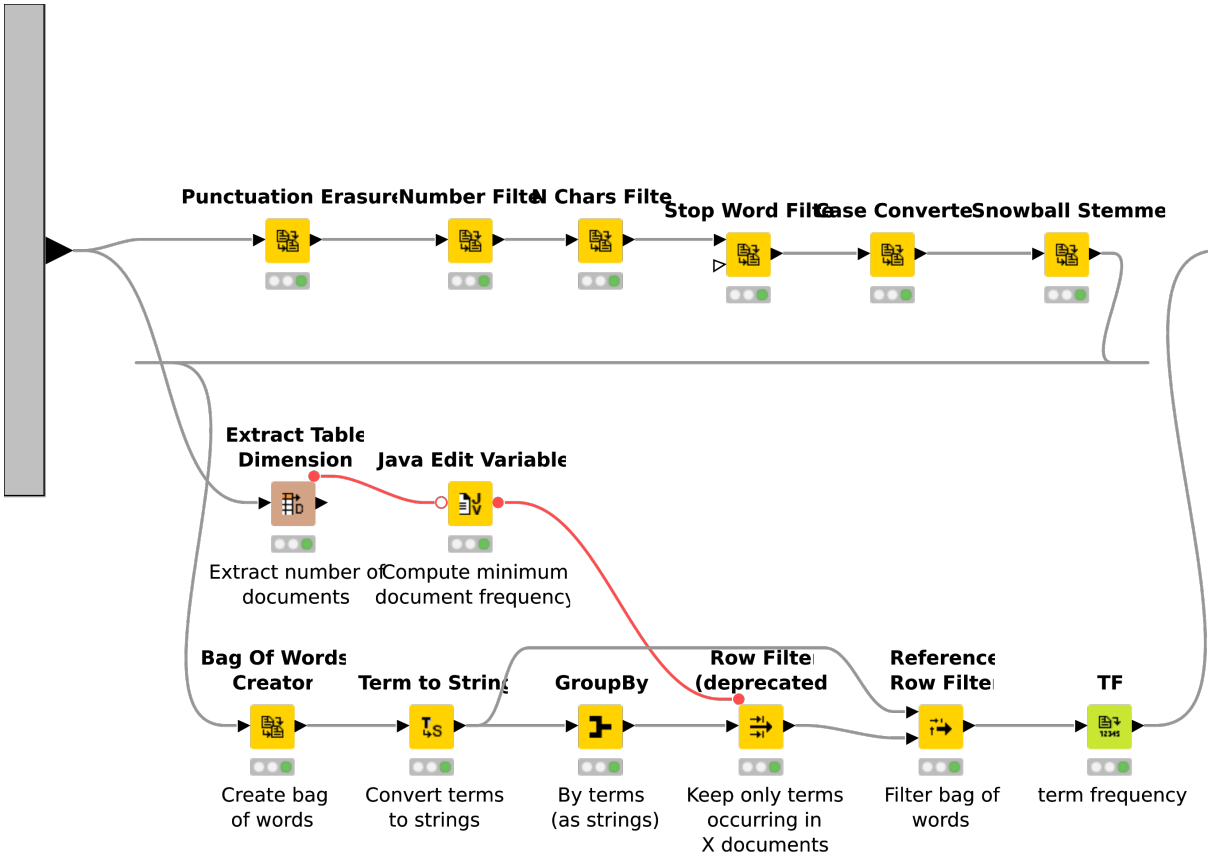

Figura 3:

El flujo de preprocesamiento descrito anteriormente fue diseñado teniendo en cuenta varios factores clave:

1. **Reducción de Ruido:** La eliminación de stopwords, caracteres no alfanuméricos y la normalización de espacios en blanco ayudan a eliminar ruido del texto, lo que mejora la calidad de los datos de entrada.

2. **Consistencia Lexical:** La conversión a minúsculas y la aplicación de stemming o lematización garantizan que las palabras similares se traten de manera consistente, reduciendo la dimensionalidad del vocabulario y mejorando la eficiencia del análisis.

3. **Preparación para Modelos Avanzados:** Las operaciones de tokenización y creación de documentos estructurados preparan los datos para técnicas avanzadas de modelado, como el modelado de temas (LDA) y el análisis de sentimientos.

4. **División de Datos:** La partición de los datos en conjuntos de entrenamiento y prueba es esencial para evaluar la efectividad del modelo de manera objetiva y evitar el sobreajuste.

1.2.2. Modelado de Temas

En esta sección, el workflow utiliza técnicas de modelado de temas para identificar patrones temáticos en los documentos procesados.

Los resultados obtenidos del metanodo preprocesamiento alimenta al nodo **Topic Extractor (Parallel LDA)**, que implementa el algoritmo Latent Dirichlet Allocation (LDA) para extraer temas latentes en los documentos. Este modelo estima la distribución de temas en cada documento y la distribución de palabras en cada tema.

El nodo **Table View** proporciona una visualización tabular de los resultados del modelado de temas, permitiendo inspeccionar los temas generados y sus términos más relevantes.

1.2.3. Minería de Redes

La minería de redes se utiliza para analizar las interacciones y relaciones entre entidades extraídas de los documentos. Primero se construyen las relaciones con el nodo **Term Co-Occurrence Counter** calcula la frecuencia de co-ocurrencia de términos en los documentos, lo que permite capturar relaciones semánticas entre palabras El nodo **Network Creator (local)** construye una red basada en las relaciones detectadas durante el proceso de extracción de temas y co-ocurrencias de términos.

El nodo **Network Viewer** visualiza la red generada, permitiendo explorar las conexiones entre nodos y identificar clusters o comunidades significativas. Esta visualización es valiosa para comprender las estructuras subyacentes en los datos.

1.2.4. Análisis de Sentimientos

La última sección del workflow se centra en el análisis de sentimientos mediante aprendizaje automático. El nodo **Decision Tree Learner** entrena un modelo de árbol de decisión para clasificar los documentos según su polaridad (positiva, negativa o neutral). Este nodo utiliza características derivadas del preprocesamiento y el modelado de temas como entrada para el modelo.

El nodo **Decision Tree Predictor** aplica el modelo entrenado a nuevos datos para predecir la polaridad del sentimiento. Finalmente, el nodo **Scorer** evalúa la precisión del modelo comparando las predicciones con las etiquetas reales, proporcionando métricas de rendimiento como precisión, recall y F1-score.

1.2.5. Flujo General del Workflow

El workflow sigue un flujo lógico y secuencial:

1. Los datos se leen y preprocesan para prepararlos para el análisis.

2. Se extraen temas latentes utilizando LDA para identificar patrones temáticos.
3. Se construye y visualiza una red basada en las relaciones entre términos y documentos.
4. Se realiza un análisis de sentimientos utilizando un modelo de árbol de decisión para clasificar la polaridad de los textos.

Este enfoque integrado permite abordar múltiples aspectos de la minería de texto, desde la extracción de conocimiento hasta la interpretación de emociones y relaciones.

1.3. Parámetros Utilizados en el Proceso de Minería de Texto

Durante el desarrollo del flujo de trabajo en KNIME, se configuraron diversos parámetros con el objetivo de optimizar el desempeño y la interpretabilidad de las técnicas aplicadas. A continuación, se detallan los valores seleccionados para los componentes claves de la minería.

En primer lugar, en la fase de particionamiento, se utilizó un esquema de división estratificada donde el 70 % de los datos se asignó al conjunto de entrenamiento y el 30 % restante al conjunto de prueba (ver figura 1). Esta proporción es ampliamente adoptada en estudios de minería de texto y aprendizaje automático, ya que permite disponer de suficientes datos para entrenar modelos robustos, manteniendo al mismo tiempo un tamaño muestral representativo para la evaluación final del desempeño. El uso de esta partición ayuda a reducir el riesgo de sobreajuste y proporciona una estimación más realista del comportamiento del modelo en datos no vistos.

Por otro lado, en la etapa de modelado de tópicos, se empleó el algoritmo Latent Dirichlet Allocation (LDA) implementado en el nodo *Topic Extractor (Parallel LDA)*. Para este análisis, se definió un número fijo de tres (3) tópicos o clusters temáticos. La elección de este valor se basó en consideraciones prácticas y exploratorias: dado el volumen relativamente reducido de documentos y la naturaleza general del corpus analizado, un número bajo de tópicos facilita la interpretación semántica de los resultados sin caer en una granularidad excesiva que pudiera dificultar su comprensión. Además, trabajar con un número moderado de tópicos reduce la complejidad computacional y mejora la estabilidad del modelo durante el proceso de inferencia.

Estos parámetros fueron definidos tras un análisis preliminar de los datos y pruebas piloto que permitieron identificar configuraciones que equilibran eficiencia computacional, interpretabilidad y calidad del resultado final.

2. Resultados

La matriz de confusión obtenida tras aplicar el modelo de clasificación a las reseñas de los clientes de la cadena hotelera es la siguiente:

Real \Pred	POS	NEG
POS	117	19
NEG	16	48

A partir de esta matriz, se han calculado las siguientes métricas clave:

1. **Accuracy**: 82.5 %. El modelo acierta correctamente el 82.5 % de las reseñas totales.
2. **Precisión**: 87.97 %. De todas las reseñas predichas como positivas, el 87.97 % realmente lo son.
3. **Recall**: 86.03 %. El modelo detecta correctamente el 86.03 % de las reseñas positivas reales.
4. **F1-Score**: 86.98 %. Esta medida equilibrada entre precisión y recall refleja un buen rendimiento global del modelo.
5. **Especificidad**: 75 %. Detecta correctamente el 75 % de las reseñas negativas reales.

Estos resultados indican que el modelo tiene un desempeño sólido en la identificación de reseñas positivas, lo cual es crucial para una cadena hotelera que busca medir la satisfacción del cliente. La alta sensibilidad y el F1-score sugieren que el modelo no está pasando por alto muchas experiencias positivas, lo cual puede ayudar a tomar decisiones estratégicas sobre marketing, mejora de servicios y fidelización de clientes.

Sin embargo, la especificidad del 75 % indica que hay margen de mejora en la detección de reseñas negativas, ya que aún se están etiquetando como positivas algunas reseñas que deberían ser negativas. Esto podría llevar a una visión algo optimista de la percepción real de los huéspedes si no se corrige.

Discusión e Interpretación

Conclusiones

Recomendaciones