ANÁLISIS DE SENTIMIENTO PARA DETERMINAR PATRONES PREDICTIVOS DE PROBLEMAS DE CRISIS REPUTACIONAL EN HOTELES DE BOGOTÁ

Juan Manuel Silva López
79544833

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
PROFESOR MARIO JULIÁN MORA

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 9 DE 2025

RESUMEN

Este proyecto se centró en el análisis de reseñas de hoteles en Bogotá con el objetivo de identificar patrones textuales y temporales asociados a posibles crisis de reputación empresarial. Mediante el uso de herramientas de ciencia de datos como *Python*, Jupyter Notebook y bibliotecas especializadas como Pandas, Scikit-learn y NLTK, se procesaron miles de opiniones de usuarios para construir un *modelo predictivo basado en sentimientos negativos*. Se realizó un análisis de series de tiempo sobre reseñas negativas, identificando patrones estacionales y periodos críticos que pueden servir como alertas tempranas. Este componente permitió incorporar una dimensión temporal valiosa para la toma de decisiones estratégicas. Se aplicaron técnicas de *procesamiento de lenguaje natural (NLP)*, incluyendo lematización y vectorización, para transformar el texto en variables cuantificables. A partir de una función de clasificación que distinguía entre estados de crisis y no crisis, se entrenaron cuatro algoritmos de *aprendizaje supervisado*: Regresión Logística, Random Forest, Support Vector Machine (SVM) y MLPClassifier. Cada modelo fue evaluado antes y después del ajuste de hiperparámetros mediante GridSearchCV, siendo el SVM y el MLP los que lograron mejores métricas de precisión y recall en la predicción de crisis. En conjunto, el proyecto demuestra la viabilidad de utilizar análisis de sentimientos y *aprendizaje automático* para fortalecer la gestión reputacional en el sector hotelero, con potencial de escalabilidad a otras industrias dependientes de plataformas de opinión digital.

ABSTRACT

This project focused on the analysis of hotel reviews in Bogotá with the aim of identifying textual and temporal patterns associated with potential corporate reputation crises. Using data science tools such as *Python*, Jupyter Notebook, and specialized libraries such as Pandas, Scikit-learn, and NLTK, thousands of user reviews were processed to build a *predictive model based on negative sentiment*. A time series analysis of negative reviews was performed, identifying seasonal patterns and critical periods that can serve as early warnings. This component allowed for the incorporation of a valuable temporal dimension for strategic decision-making. *Natural language processing (NLP)* techniques, including lemmatization and vectorization, were applied to transform the text into quantifiable variables. Based on a classification function that distinguished between crisis and non-crisis states, four *supervised learning* algorithms were trained: Logistic Regression, Random Forest, Support Vector Machine (SVM), and MLP Classifier. Each model was evaluated before and after hyperparameter tuning using GridSearchCV, with the SVM and MLP models achieving the best accuracy and recall metrics in crisis prediction. Overall, the project demonstrates the feasibility of using sentiment analysis and *machine learning* to strengthen reputation management in the hospitality sector, with potential for scalability to other industries that rely on digital review platforms

Tabla de contenido

i. LISTA DE FIGURAS

ii. LISTA DE TABLAS

iii. INTRODUCCIÓN

En la era actual altamente digitalizada, las empresas enfrentan un entorno dinámico y complejo en el que su nombre o prestigio online pueden verse afectados a causa de eventos inesperados o comentarios negativos en las redes sociales que, por la disponibilidad e inmediatez que permiten internet y las redes sociales, pueden tener un impacto devastador en la imagen de marca, las ventas y la rentabilidad de una empresa, llegando a una audiencia amplia o de interés en cuestión de minutos u horas.

La capacidad de anticipar y/o prevenir estas crisis se convierte en una necesidad crucial para las empresas que operan online u offline. El presente proyecto buscó, mediante metodologías y herramientas de Ciencia de Datos, hallar patrones relevantes de comportamiento de la opinión generalizada de los usuarios de los servicios hoteleros en la ciudad de Bogotá, que permitieran anticipar dichos fenómenos. Identificando los principales indicadores de crisis reputacional se implementó un método de recolección de reseñas que hicieran referencia a los hoteles con mayor número de menciones en la plataforma Trip Advisor, mediante técnicas de consulta a la API, las cuáles fueron almacenadas en una base de datos en formato csv.

Luego de una adecuada preparación y limpieza de los datos, se realizó una clasificación de sentimientos según su polaridad (positivo, negativo, neutral) por medio de un algoritmo de aprendizaje automático que sirvió de base para entrenar el modelo utilizando un conjunto de datos etiquetados.

El modelo se puso en marcha y se hizo una evaluación periódica del mismo para analizar los patrones de sentimiento a lo largo del tiempo durante el periodo planteado para la ejecución del proyecto.

Finalmente se proporcionaron recomendaciones para prevenir y gestionar crisis reputacionales, las cuales se incluyeron dentro de la redacción de este documento final con los hallazgos del proyecto experimental.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Las crisis reputacionales son fenómenos sociales que dañan la imagen pública de una empresa o persona y pueden tener un impacto negativo en la percepción que tiene la opinión pública hasta en sus oportunidades de negocios o su desempeño financiero, o como definió Combs [1]: "Una crisis de reputación es la percepción de un evento imprevisible que amenaza de manera importante las expectativas de los stakeholders y puede impactar seriamente el funcionamiento de una organización y generarle resultados negativos".

Diversos factores pueden ocasionar una crisis, por ejemplo, venta de productos de mala calidad, prestación de servicios de calidad diferente a la ofrecida, campañas de mercadeo equivocadas o mal enfocadas, abordajes éticos mal estimados, comentarios negativos de clientes o usuarios insatisfechos en las redes sociales o incluso competencia desleal.

Por la disponibilidad e inmediatez que se dan en internet y las redes sociales, estas crisis de reputación pueden tener un impacto devastador en la imagen de marca, las ventas y la rentabilidad de una empresa, llegando a audiencias masivas o de interés en cuestión de minutos u horas, lo que implica desarrollar metodologías o sistemas que permitan afrontar dichas situaciones con la misma velocidad y efectividad con que se generan.

Es en este punto en el que la Ciencia de Datos, con su capacidad de tratar cantidades ingentes de información en periodos cortos de tiempo que serían imposibles de procesar para un humano, otorga a las empresas o personas la capacidad de actuar en consecuencia y con rapidez ante estas contingencias, permitiendo tomar decisiones pertinentes en márgenes de tiempo menores, que redundarán en atención y ejecución más temprana de las soluciones.

La anticipación de crisis reputacionales mediante análisis del sentimiento en redes sociales tiene importantes aplicaciones en el ámbito del mercadeo y las relaciones públicas. Al identificar las señales de alerta temprana, las empresas pueden tomar medidas proactivas para prevenir o mitigar el impacto de una crisis. Esto puede incluir la implementación de estrategias de comunicación efectivas, la resolución rápida de problemas y el fortalecimiento de las relaciones con los clientes.

Las tecnologías de ciencia de datos, como el análisis del sentimiento, el aprendizaje automático y la minería de datos, son herramientas valiosas para la anticipación de crisis reputacionales. Mediante el análisis de grandes cantidades de datos de redes sociales, las empresas podrían identificar patrones y tendencias que podrían indicar una crisis en crecimiento o desarrollo.

Las crisis reputacionales corresponden al área de las relaciones públicas, siendo éstas contenidas en el ámbito de mercadeo de las empresas, a su vez inmersa dentro de la Administración de Empresas. Para este proyecto se enfocó el análisis en empresas del orden local de la ciudad de Bogotá, siendo los hoteles, un ámbito propicio para el estudio, y dado el tamaño de este clúster económico, su relevancia en la dinámica económica de la ciudad y la experiencia del autor en este sector, específicamente los cuarenta y nueve hoteles con mayor número de reseñas registradas en la plataforma Trip Advisor, por parte de usuarios en general de este tipo de servicios, en las que se refieren a la calidad o necesidad de los servicios que dichas empresas prestan.

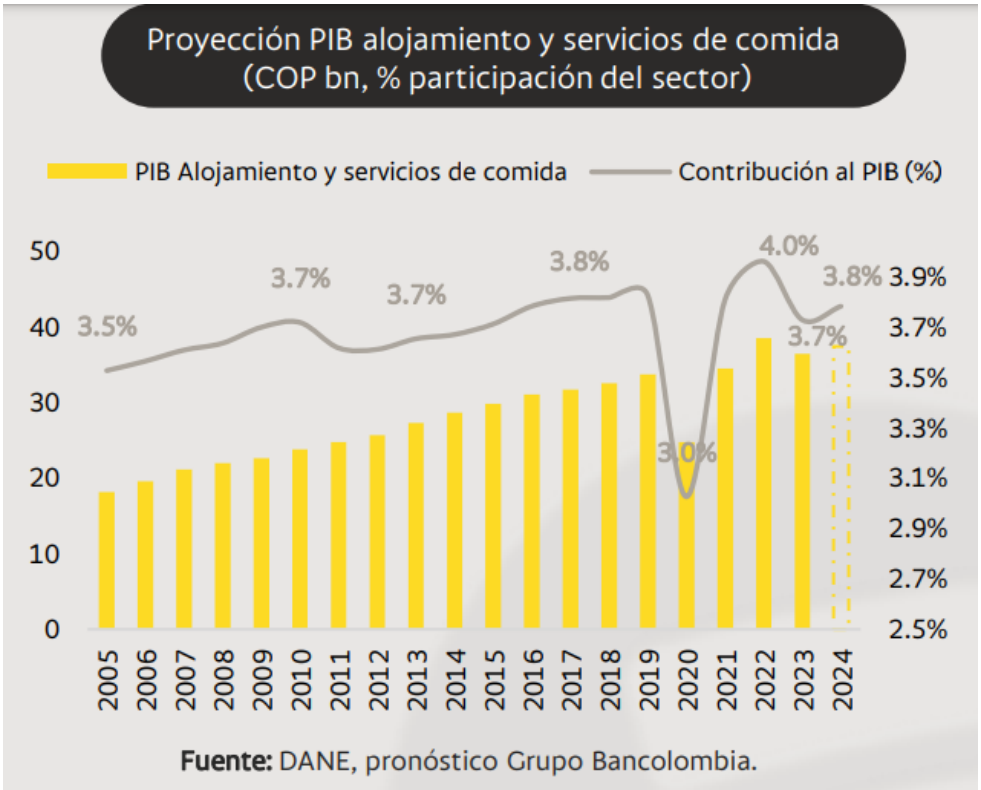


Fig. 1: Proyección PIB alojamiento y servicios de comida (% participación del sector)

De acuerdo con proyecciones suministradas en el reporte Panorama Hotelería y Turismo de Bancolombia, "el sector (de hotelería y turismo) alcanzaría en 2024 un crecimiento real de 2,7% y mantendría su participación en el PIB, al ubicarse en 3,8%" conforme a datos proporcionados por el DANE y pronósticos del mismo grupo Bancolombia. [2]

1.2. FORMULACIÓN DEL PROBLEMA

A pesar de las limitaciones, el análisis del sentimiento ofrece un gran potencial para mejorar la anticipación y gestión de crisis reputacionales. Existen oportunidades para investigar y desarrollar nuevas herramientas y metodologías de análisis del sentimiento que sean más precisas, confiables y efectivas para identificar y monitorear las señales de alerta temprana de crisis.

Esto nos lleva a resolver la pregunta del presente proyecto aplicado: ¿Existen patrones de sentimiento relevantes y/o similares en los procesos de formación y desarrollo de los fenómenos de crisis de reputación digital, en las reseñas registradas de los 50 hoteles con la mayor cantidad de éstas publicadas en la plataforma Trip Advisor?

Para buscar solución a la pregunta del proyecto aplicado, recurrimos a las siguientes preguntas de sistematización: ¿qué metodología de recopilación y almacenamiento de datos podremos aplicar para disponer de la data necesaria para el análisis? ¿cómo seleccionar el algoritmo apropiado de análisis de lenguaje natural para valorar adecuadamente la data obtenida? ¿cómo construir un modelo de ciencia de datos que permita hallar patrones de sentimiento representativos de los fenómenos de crisis de reputación digital?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Desarrollar un modelo de clasificación que permita identificar la ocurrencia de crisis reputacional en los 49 hoteles de Bogotá, con mayor número de reseñas publicadas.

2.2. OBJETIVOS ESPECÍFICOS

2.2.1 Objetivo específico 1

Implementar una metodología de obtención de las reseñas de los usuarios de la plataforma Trip Advisor en la que se mencionan los hoteles objeto del estudio, mediante una extensión del navegador Google Chrome.

2.2.2 Objetivo específico 2

Analizar, desde el NPL (procesamiento de Lenguaje Natural) el sentir del público acerca de los hoteles elegidos.

2.2.3 Objetivo específico 3

Desarrollar un modelo de clasificación que permita evidenciar patrones recurrentes de sentimiento en las publicaciones recogidas y valoradas.

2.2.4 Objetivo específico 4

Evaluar el desempeño del modelo de clasificación y proponer ajustes para mejorarlo.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

3.1.1 Elementos definitorios de las crisis reputacionales

La reputación de marca de una empresa, organización o persona, es un activo invaluable que sustenta el éxito y la permanencia en el mercado. Sin embargo, este activo no está exento de amenazas.

El modelo catalítico de gestión de asuntos (de crisis) propuesto por Crable y Vibbert, citado por Costa [3] nos indica que se puede dividir la gestión de crisis en cinco etapas:

- Primera etapa: Potencial: Durante esta etapa naciente del asunto de crisis, solo algunos individuos o grupos son conocedores de la situación.

- Segunda etapa: Inminente: En esta etapa otros individuos o grupos dan valor o credibilidad o legitiman la situación y se involucran de manera directa.

- Tercera etapa: Presente: Cuando la situación pasa a ser reconocida por grupos de interés y en la cual los medios de comunicación empiezan a hacer divulgación del tema.

- Cuarta etapa: Crítica: La presión para la toma de decisiones aumenta y hace inminente la toma de decisiones.

- Quinta etapa: Latente: Se llega a una decisión y el asunto se resuelve.

En el ámbito colombiano, la Superintendencia Financiera de Colombia indica que el riesgo reputacional es "la posibilidad de pérdida en que incurre una compañía por desprestigio, mala imagen, publicidad negativa, cierta o no, respecto de la institución y sus prácticas de negocios, que cause pérdida de clientes, disminución de ingresos o procesos judiciales"[4].

Una crisis reputacional es un evento o cadena de eventos negativos que impacta de forma significativa la percepción que los públicos clave tienen sobre una organización, empañando su imagen y poniendo en riesgo su credibilidad.

Las crisis reputacionales pueden adoptar diversas formas, pero algunas de las más comunes son:

Crisis por mala calidad de productos o servicios: Cuando los productos o servicios no cumplen con las expectativas de los consumidores, generando insatisfacción y posibles daños.

Crisis por prácticas antiéticas: Actos como la corrupción, el fraude o el engaño a los consumidores erosionan la confianza y dañan la imagen de la organización.

Crisis por incidentes de seguridad: Fugas de datos, ciberataques o accidentes ponen en riesgo la información sensible de los clientes, generando alarma y desconfianza.

Crisis por conflictos laborales: Malas condiciones de trabajo, despidos injustificados o huelgas pueden generar una imagen negativa de la empresa ante sus empleados y la sociedad en general.

Crisis por responsabilidad social: Daños al medio ambiente, negligencia en la responsabilidad social empresarial o apoyo a causas controvertidas pueden generar un rechazo por parte de diversos públicos.

Las consecuencias de una crisis reputacional para una marca, empresa, organización o persona pueden ser devastadoras y de largo alcance, afectando diversos aspectos de su funcionamiento y su imagen pública.

Algunos de los principales impactos negativos que puede generar una crisis de este tipo son el daño a la imagen y la reputación, pérdida de clientes y ventas, daño financiero, impacto en el valor de la marca, problemas

legales y regulatorios, daño a la moral y la productividad de los empleados, dificultades para atraer y retener el talento humano; y en el caso de las personas: impacto en la salud mental y emocional y dificultades en las relaciones personales y profesionales.

Al abordar el rol de Internet como amenaza a la reputación de las organizaciones, también agregó Costa [3] citando a Herrero y Smith (2008) quienes proponen que internet puede ser entendido de dos maneras:

"1) A veces simplemente Internet actúa como un agente que acelera el ciclo de noticias acerca de una crisis y rompe los límites geográficos, es decir Internet se convierte en un canal adicional ó 2) Internet también puede ser un factor desencadenante de un problema lo suficientemente importante como para ser considerado una crisis si no es manejado adecuadamente".

Ramos [5] indicó que desde la aparición de la web 2.0, se transformó el papel del internauta como un sujeto pasivo consumidor de información, a transformarse en un usuario activo que actúa desde tres roles: generador, transmisor y consumidor de contenidos, ocasionando que la opinión pública en general empezara a otorgar mayor credibilidad e importancia a la opinión y recomendaciones de sus contactos que a otros canales de comunicación más tradicionales.

3.1.2 Elementos definitorios de la Ciencia De Datos

El Análisis de Sentimiento es "una disciplina que comprende la tarea de identificar y clasificar fragmentos de texto que contengan opinión emotiva o subjetiva" [6]. En el contexto de las redes sociales se utiliza para comprender el sentimiento de los usuarios hacia una marca, producto, servicio o tema en particular. Esta información puede ser valiosa para las empresas para identificar posibles problemas de reputación, mejorar la atención al cliente y desarrollar estrategias de marketing más efectivas.

En el caso de los hoteles en Bogotá, el Análisis de Sentimiento puede ser una herramienta útil para identificar patrones en las opiniones de los usuarios que puedan predecir la probabilidad de ocurrencia de crisis reputacionales. Estas crisis pueden tener un impacto negativo en la imagen de un hotel, la satisfacción del cliente y, en última instancia, los resultados financieros.

El Análisis de sentimiento en redes sociales se basa en la extracción de características lingüísticas de los textos publicados por los usuarios [7]. Estas características pueden incluir palabras clave, frases, emojis, hashtags y estructuras gramaticales. A continuación, se aplican algoritmos de aprendizaje automático para clasificar el sentimiento de los textos como positivo, negativo o neutro.

3.1.3 Enfoques para el Análisis de Sentimiento en redes sociales.

Existen diferentes enfoques para el Análisis de Sentimiento en redes sociales, que se pueden clasificar en dos categorías principales:

Análisis de Sentimiento Basado en Léxicos:

Este enfoque utiliza diccionarios de palabras o frases con una puntuación de sentimiento predefinida. La puntuación de sentimiento del texto se calcula como el promedio de las puntuaciones de las palabras o frases que contiene.

Análisis de Sentimiento Basado en Aprendizaje Automático:

Estos enfoques entrenan modelos de aprendizaje automático para clasificar el sentimiento de los textos. Los modelos se entrenan con un conjunto de datos de textos etiquetados con su sentimiento correspondiente.

Patrones predictivos de crisis reputacionales:

Son indicadores que pueden alertar a las empresas sobre la posibilidad de que se produzca una crisis. Estos patrones pueden identificarse mediante el análisis de las opiniones de los usuarios en las redes sociales. Algunos ejemplos de patrones predictivos de crisis reputacionales incluyen:

- Aumento repentino del volumen de publicaciones negativas: Si el número de publicaciones negativas sobre una empresa aumenta repentinamente, esto puede ser un indicador de que se está gestando una crisis reputacional.
- Cambio en el tono de las opiniones: Si el tono de las opiniones sobre una empresa cambia de positivo a negativo, esto puede ser otro indicador de una posible crisis.
- Aparición de palabras clave negativas: Si en las opiniones sobre una empresa comienzan a aparecer palabras clave negativas, como "estafa", "fraude" o "mala calidad", esto puede ser una señal de alerta.

Se destaca que el análisis del sentimiento no es una solución mágica para la anticipación y gestión de crisis reputacionales. Es una herramienta que debe combinarse con otras estrategias de gestión de crisis, como la comunicación efectiva y la resolución rápida de problemas. Además, el análisis del sentimiento puede ser susceptible a sesgos y errores, por lo que es importante utilizar herramientas y metodologías confiables.

3.1.4 Tecnologías a utilizar

NLP: Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) es un campo interdisciplinario de la inteligencia artificial y la lingüística computacional que se ocupa de la interacción entre computadoras y el lenguaje humano. En el contexto de ciencia de datos, NLP abarca un conjunto de técnicas y herramientas que permiten a las máquinas leer, interpretar, entender, generar y extraer información útil de datos textuales no estructurados [8].

Python: Lenguaje desarrollado por Guido van Rossum en 1991. Es un lenguaje de programación de alto nivel, interpretado y de propósito general, conocido por su sintaxis clara y legible. Es ampliamente utilizado en desarrollo web, análisis de datos, inteligencia artificial y automatización de tareas [9].

Anaconda: Es una plataforma de desarrollo de código abierto para Python y R, enfocada en la ciencia de datos y el aprendizaje automático. Desarrollado por Anaconda, Inc. y fue lanzado en 2012. Incluye una colección de más de 1,500 paquetes y herramientas como Jupyter Notebook y Conda para la gestión de entornos [10].

Jupyter Notebook: Es una aplicación web desarrollada por Fernando Pérez y Brian Granger, creada en 2014, que permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Es ampliamente utilizado en análisis de datos, modelado estadístico y aprendizaje automático.

Pandas: Es una biblioteca que proporciona estructuras de datos y herramientas de análisis de datos fáciles de usar y de alto rendimiento para el lenguaje de programación Python. Es especialmente útil para la manipulación y análisis de datos estructurados. Fue desarrollada por Wes McKinney y lanzada en 2008 [11].

NumPy: Es una biblioteca fundamental para la computación científica en Python. Proporciona soporte para arreglos y matrices multidimensionales, junto con una colección de funciones matemáticas de alto nivel para operar con estos arreglos. Fue desarrollada por Travis Oliphant y lanzada en 2006 [12]

Matplotlib: Es una biblioteca de gráficos 2D para Python que permite generar gráficos y visualizaciones de alta calidad. Es ampliamente utilizada para la visualización de datos en ciencia, ingeniería y análisis financiero. Fue desarrollada por John D. Hunter en 2003 [13].

Seaborn: Es una biblioteca de visualización de datos basada en Matplotlib que proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos. Facilita la creación de gráficos complejos con menos código. Desarrollada por Michael Waskom en 2014 [14].

NLTK: (Natural Language Toolkit) es una biblioteca de Python ampliamente utilizada para el procesamiento de lenguaje natural (NLP) en ciencia de datos y lingüística computacional. Proporciona un conjunto completo de herramientas, recursos y corpus que permiten realizar tareas como tokenización, etiquetado gramatical, análisis sintáctico, lematización, stemming, análisis de sentimiento, extracción de entidades, entre otros [15].

Scikit-learn: Es una biblioteca de aprendizaje automático en Python que proporciona herramientas simples y eficientes para el análisis predictivo de datos. Incluye algoritmos para clasificación, regresión, agrupamiento y reducción de dimensionalidad. Fue desarrollada por David Cournapeau en 2007 [16]

Algoritmos de Machine Learning

Regresión Logística: Tiene su origen en el Siglo XIX. Es un modelo estadístico utilizado para predecir la probabilidad de una variable dependiente binaria. Es ampliamente utilizada en clasificación binaria y análisis de riesgos [17].

Random Forest: Es un algoritmo de aprendizaje automático de conjunto desarrollado por Tin Kam Ho en 1995 y depurado por Leo Breiman en 2001. Construye múltiples árboles de decisión y combina sus resultados para mejorar la precisión y controlar el sobreajuste. Es eficaz para clasificación y regresión [18].

Máquinas de Vectores de Soporte (SVM): Es un algoritmo de aprendizaje supervisado que se utiliza para clasificación y regresión. Funciona encontrando el hiperplano que mejor separa las clases en el espacio de características. Fue desarrollado por Vladimir Vapnik en 1995 [19].

MLPClassifier: Es una implementación de perceptrón multicapa en Scikit-learn. Utiliza redes neuronales artificiales para modelar relaciones no lineales complejas en los datos, siendo útil para tareas de clasificación.

GridSearchCV: Es una herramienta de la herramienta Scikit-learn que permite realizar una búsqueda exhaustiva sobre un conjunto de parámetros especificado para un estimador. Es útil para encontrar la mejor combinación de hiperparámetros mediante validación cruzada [20].

3.2. ANTECEDENTES

3.2.1 Definiciones

La definición del concepto de *reputación* en la RAE se define como: 1. La opinión o consideración en que se tiene alguien o algo. 2. El prestigio o estima en que son tenidos alguien o algo.

Para Fombrun (1996), uno de los principales expertos en reputación corporativa "La reputación corporativa es una representación perceptual de las acciones pasadas de la empresa y expectativas que describen el atractivo general de la firma para sus grupos de interés clave, al compararla con sus principales rivales." [21].

Por su parte, Brown (1997) analiza cómo las asociaciones que los consumidores establecen entre una empresa y sus productos influyen en la percepción de la marca y, por ende, en la reputación corporativa, centrándose en la idea de que la reputación de una empresa está estrechamente ligada a su identidad corporativa. Es decir “la percepción que tienen los stakeholders o actores clave de una compañía, está influenciada por la manera en que esta se presenta a sí misma, sus valores, su cultura y su comportamiento” [22].

Rubio (2017) recogía que, “En la práctica, la reputación es una faceta que afecta al posicionamiento de las compañías, entendido este como la proposición de valor o ventaja comparativa que una empresa tiene frente a otras cuando el cliente las jerarquiza en su mente” [23].

Cervantes Flores y otros (2021) en su artículo “Análisis de sentimientos a comentarios de hoteles: Un caso práctico” realizaron una investigación documental, descriptiva y cuantitativa, mediante análisis de texto con un software comercial llamado SAS Enterprise Miner® para analizar sentimientos, respondiendo a la pregunta “¿Qué hacen los hoteles con las opiniones de los clientes al momento de entregar la habitación, y cómo se ha medido esa opinión?”. [24]

Dentro del marco de referencia de Ciencia de Datos, Lovera y otros (2023) presentan un trabajo comparativo de análisis de sentimientos realizado en Twitter, con una propuesta de una estrategia metodológica que abarca las fases de preprocesamiento de datos, construcción de modelos predictivos y su evaluación, cuyo objetivo fue “evaluar técnicas tanto de modelos inteligentes de Machine Learning, como Regresión Logística, Naive Bayes y Support Vector Machine (SVM), como de Deep Learning, como Convolutional Neural Network (CNN), Long Short Term Memory

(LSTM) y Bidirectional Long Short Term Memory (Bi-LSTM). El conjunto de datos (dataset) utilizado, para la evaluación de tales técnicas es el de Sentiment140 que contiene 1,600,000 tuits en Ingles etiquetados como positivo, negativo y neutral, así como metadatos que describen cada Tuit" [25].

3.2.2 Hechos relevantes de crisis reputacionales

En los últimos años, hemos presenciado diversas crisis reputacionales que han tenido un impacto significativo en empresas de diversos sectores. Algunos ejemplos notables incluyen:

Tabla I. Ejemplos de crisis reputacionales en los últimos años

AÑO	EMPRESA	SITUACIÓN	IMPACTO
2016	Uber	Esta empresa de transporte compartido fue criticada por prácticas discriminatorias y acoso sexual.	Caída en el valor de sus acciones y éxodo de usuarios [26]
2017	United Airlines	Un pasajero de origen asiático fue arrastrado por la fuerza de un avión de United Airlines	Ola de indignación en las redes sociales y boicots a la aerolínea y caída del valor de sus acciones [27]
2018	Facebook	La red social fue objeto de un escándalo por el uso indebido de datos de usuarios por Cambridge Analytica	Investigaciones por parte de reguladores y una pérdida de confianza entre los usuarios [28]
2019	Boeing	Esta compañía aeroespacial se vio envuelta en una crisis tras dos accidentes fatales del avión Boeing 737 Max	Suspensión de vuelos y un daño significativo a su reputación [29]
2020	Nestlé	Esta empresa de alimentos fue criticada por la comercialización de leche maternizada en polvo en países en desarrollo	Acusaciones de explotación y daños a la salud de los bebés [30]

En todos estos casos, el análisis del sentimiento en redes sociales podría haber sido utilizado como una herramienta valiosa para identificar y monitorear las señales de alerta temprana de una crisis en ciernes. Al analizar los comentarios y opiniones de los usuarios en las redes sociales, las empresas podrían haber detectado la creciente insatisfacción y el sentimiento negativo antes de que la crisis se intensificara y tuviera un impacto significativo en su reputación.

3.3. QUÉ ES TRIPADVISOR

TripAdvisor, mucho más que un sitio web de opiniones, es una comunidad global de viajeros que comparten sus experiencias, una fuente de información confiable para tomar decisiones de viaje y una plataforma que ha transformado la forma en que los usuarios entienden y disfrutan de sus viajes.

Nacido en febrero de 2000, *TripAdvisor* revolucionó la industria turística al ofrecer una plataforma donde los viajeros podían compartir sus experiencias de forma honesta y transparente. La idea de Stephen Kaufer y sus socios de crear un espacio donde la información sobre destinos y alojamientos fuera confiable y estuviera al alcance de todos, pronto se convirtió en una realidad. Con el paso de los años, TripAdvisor no solo creció en número de usuarios, sino también en funcionalidades. En 2004, fue adquirido por IAC (InterActiveCorp) y en 2005 se independizó como Expedia, Inc. Estas adquisiciones permitieron a la plataforma expandir sus horizontes y consolidarse como un referente mundial en el sector.

Hitos clave en su trayectoria incluyen: la compra de kuxun.cn en 2009, que le permitió adentrarse en el mercado chino, y la adquisición de holidaylettings.co.uk en 2010, fortaleciendo su oferta en el segmento de alquileres vacacionales. En 2012, la integración con Facebook marcó un antes y un después, permitiendo a los usuarios compartir sus opiniones con sus amigos y contactos.

Gracias a su constante innovación y a la confianza de millones de usuarios, TripAdvisor fue reconocido en 2014 como el portal de viajes más prestigioso, utilizado y fiable del mundo. En los últimos años, la plataforma ha ido más allá de las reseñas tradicionales, ofreciendo herramientas cada vez más sofisticadas para ayudar a los viajeros a planificar sus viajes de manera personalizada y eficiente.

En palabras de Rubio, “TripAdvisor hace de la reputación su actividad, celebrando cada año los “Travelers’ Choice Awards” (Premios de la elección del viajero). Para la concesión de los premios se diferencian varias categorías, como son el mejor hotel del mundo, Europa o España, premios a la mejor isla, resort o al mejor sitio de interés turístico. Los resultados se consiguen a partir de las reseñas, comentarios, valoraciones y opiniones que han proporcionado los usuarios sobre dichos hoteles (reputación)”.

El presente proyecto aplicado pretende abordar desde la ciencia de datos, la eventual anticipación a las crisis reputacionales corporativas, mediante un modelo de clasificación que permita evidenciar patrones recurrentes de sentimiento en las publicaciones recogidas y valoradas, pudiendo de esta manera permitir a empresas del sector hotelero, reconocer en series de tiempo, posibles escenarios de riesgo e implementar medidas que conjuren o mitiguen el impacto de dichas crisis.

4. OBTENCIÓN Y PREPARACIÓN DE DATOS

4.1. METODOLOGÍA DE RECOPILACIÓN DE RESEÑAS

Mediante el uso de la herramienta de software Scraper Plus, extensión de uso libre del navegador Chrome, sobre la página Trip Advisor, la cual utiliza la técnica consulta de web scraping, se hizo un barrido de cada una de las urls de los 50 hoteles con la mayor cantidad de reseñas publicadas por los usuarios de esta plataforma. El paquete de reseñas de cada hotel se descarga en un archivo separado por comas (csv).

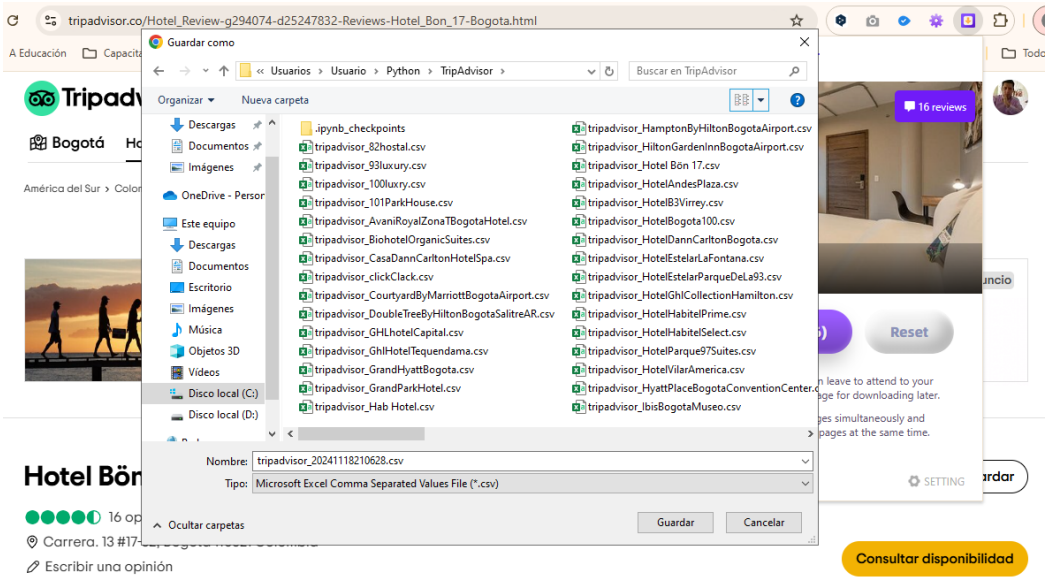
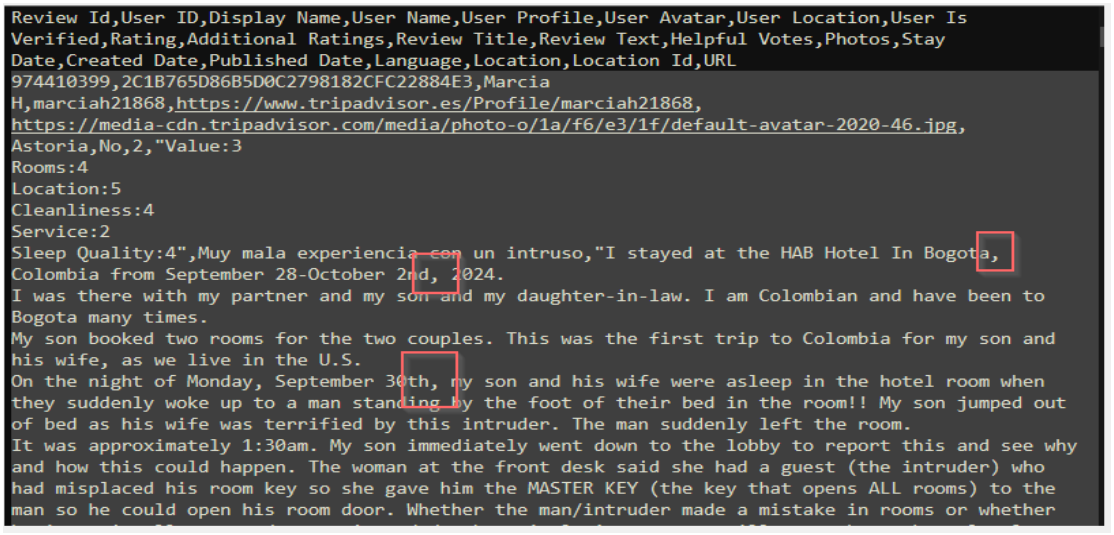


Fig. 2: Paquetes de reseñas por hotel en archivos csv

4.2. PREPROCESAMIENTO DE DATOS

4.2.1 Consolidación de la base de datos

En la consolidación de la base de datos se encontró un imprevisto consistente en que dentro de los archivos separados por comas (csv) que contenían las reseñas de cada hotel, la variable “Review Text” contiene caracteres de coma (,) que desconfigura la organización de los datos al momento de aplicar un análisis descriptivo.



```
Review Id,User ID,Display Name,User Name,User Profile,User Avatar,User Location,User Is
Verified,Rating,Additional Ratings,Review Title,Review Text,Helpful Votes,Photos,Stay
Date,Created Date,Published Date,Language,Location,Location Id,URL
974410399,2C1B765D86B5D0C2798182CFC22884E3,Marcia
H,marciah21868,https://www.tripadvisor.es/Profile/marciah21868,
https://media-cdn.tripadvisor.com/media/photo-o/1a/f6/e3/1f/default-avatar-2020-46.jpg,
Astoria,No,2,"Value:3
Rooms:4
Location:5
Cleanliness:4
Service:2
Sleep Quality:4",Muy mala experiencia con un intruso,"I stayed at the HAB Hotel In Bogota,
Colombia from September 28-October 2nd, 2024.
I was there with my partner and my son and my daughter-in-law. I am Colombian and have been to
Bogota many times.
My son booked two rooms for the two couples. This was the first trip to Colombia for my son and
his wife, as we live in the U.S.
On the night of Monday, September 30th, my son and his wife were asleep in the hotel room when
they suddenly woke up to a man standing by the foot of their bed in the room!! My son jumped out
of bed as his wife was terrified by this intruder. The man suddenly left the room.
It was approximately 1:30am. My son immediately went down to the lobby to report this and see why
and how this could happen. The woman at the front desk said she had a guest (the intruder) who
had misplaced his room key so she gave him the MASTER KEY (the key that opens ALL rooms) to the
man so he could open his room door. Whether the man/intruder made a mistake in rooms or whether
```

Fig. 3: Detalle de archivo separado por comas

Para resolver esto se recurrió a las librerias OS y Pandas, mediante un script se aplicó un proceso por lotes que reemplaza las comas (,) por punto y coma (;) en todos los archivos contenidos en la carpeta de data (data_reviews) y a su vez se agregó un nuevo campo denominado Hotel_Name usando el nombre de cada archivo que corresponde precísamente al nombre de cada hotel para permitir su filtrado posterior. Dichos archivos corregidos se guardaron en una subcarpeta “corregidos” para su posterior unificación. (Ver anexo 1: preprocessing_data_hotels.ipynb).

Para unificar los 49 archivos con reseñas correspondientes a los hoteles analizados, se desarrolló un script que unió en un solo archivo todos los archivos csv corregidos en el aparte anterior. Dando como resultado un nuevo archivo csv denominado “bogota_hotel_reviews.csv” consolidando así una base de datos de 47.489 registros.

	Hotel_Name	Review Id	User ID	Display Name	User Name	User F
47473	York Luxury	789505096	D8757AABD6E87E5AF62BCD26DCD9A813	lfsandoval71	lfsandoval71	https:
47474	York Luxury	789201772	F627757FA28481375E7AB3DA70A8D7EE	danielamarin10	danielamarin10	https:
47475	York Luxury	787555685	093A8D011446804928C3212F678FA1FD	sagomez98	sagomez98	https:
47476	York Luxury	786387772	69BF313C392B32550C541B58AACC0892	Sergio V	620octaviol	https:
47477	York Luxury	786237852	1838C401AFA47B6BFFBD0CBB02DF1274	jesscastro07	jesscastro07	https:
47478	York Luxury	784518846	BC5CAF2000C8C43416BCFD6879040B8A	Sonia	Sons091990	https:
47479	York Luxury	780530129	BC1A5E6D80AC13774F5363FF8329B1EE	magutie2021	magutie2021	https:
47480	York Luxury	780454148	5773ECD5C21935A9616C8DF1122A0347	DayTrip751888	DayTrip751888	https:
47481	York Luxury	780224454	114261511E66F1D13EC2504957B77300	lauosgom	lauosgom	https:
47482	York Luxury	780150461	4EB42BDD262DC88DA2146BD5261DC5B4	Maria Camila	363mar_acamilam	https:
47483	York Luxury	779875557	17EC14E10501212D50E5EAE979C173FA	amontesmontoya	amontesmontoya	https:
47484	York Luxury	779839107	784A23FA305551B2A429A8BD5020501F	sirmansm	sirmansm	https:
47485	York Luxury	778275268	5F6912DA056CF7DF7A7A6AF1D883AA21	Francy M	FrancyMelo	https:
47486	York Luxury	778139539	50CD1E71979C3CD41436189DE610A98D	David Salas	davidsalas89	https:
47487	York Luxury	778082253	B909F2A9188BDEC8148E7AE6EF26EAAC	Yamile A	32yamilea	https:
47488	York Luxury	777478115	A4D701390BE0BF468485991F4E777887	Ana MarÃ-a Velasco	anamariavelasco	https:
47489	York Luxury	777477837	702AFD47924E88BFA00E14511A28B536	Felipe A	felipeaF1736DR	https:
47490	York Luxury	775243983	32366EC154DD3414D0CEFFC47273EF14	Gabriel Pareja	parejag25	https:
47491						

Fig. 4: Detalle base de datos resultante

4.2.2 Eliminar reseñas duplicadas

Se preparó un script para cargar el archivo csv con la data unificada y detectar y eliminar reseñas duplicadas, irrelevantes o incompletas.

Tabla II. Conteo de reseñas duplicadas por hotel

0	93 luxury	10
1	Casa Dann Carlton Hotel Spa	20
2	Click Clack	40
3	Courtyard By Marriott Bogota Airport	10
4	Double Tree By Hilton Bogota Salitre AR	100
5	GHL Hotel Capital	340
6	GHL Hotel Tequendama	30
7	Grand Hyatt Bogota	160
8	Hilton Garden Inn Bogota Airport	10
9	Hotel B3 Virrey	50
10	Hotel Dann Carlton Bogota	30
11	Hotel Estelar La Fontana	60
12	Hotel Estelar Parque De La 93	10
13	Hotel Habitel Select	160
14	Hotel Vilar America	10
15	Ibis Bogota Museo	80
16	Lancaster House	70
17	NH Collection Bogota WTC Royal	30
18	Novotel Bogota Parque 93	190
19	Sheraton Bogota Hotel	230
20	Sofitel Bogota Victoria Regia Hotel	62
21	Sonesta Hotel Bogota	20
22	Stelar Suites Jones	24
23	W Bogota	50

Se encontraron 1.846 reseñas duplicadas, las cuáles fueron eliminadas del dataframe.

```
Cantidad de registros en df_sin_duplicados: 45643
```

4.2.3 Identificación de idiomas de reseñas

Análisis de idiomas (Language): Se identificaron los diferentes idiomas en que están escritas las reseñas usando la variable Language, hallando 4.587 reseñas en idiomas diferentes al inglés y español; esto indica que el 90% de las reseñas están en estos dos idiomas:

Tabla III. Conteo de reseñas por idioma

	Language
es	26317
en	14739
pt	3235
fr	483
it	335
de	315
nl	85
ru	34
ja	25
zhcn	15
sv	15
zhtw	7
da	6
tr	6
iw	5
pl	5
ko	4
no	3
fi	2
cs	2
in	1
ar	1
sr	1
hu	1
el	1

Por practicidad del ejercicio, pues los modelos entrenados de análisis de sentimiento disponibles en las librerías de machine learning se encuentran en inglés y por razones de rendimiento de máquina se optó por trabajar la data en este idioma y traducir al inglés solamente los títulos y las reseñas que están en español usando la librería googletrans. Se mantienen los títulos y reseñas en inglés para finalmente sumarlas y tener un número final de 41.056 reseñas para el análisis.

Las reseñas escritas en idiomas diferente al español o inglés se omitieron pues no son un número representativo respecto del total de reseñas disponibles, toda vez que las librerías disponibles para su traducción ocasionan un consumo considerable de recursos de máquina y no garantizan una calidad óptima de traducción comparadas con las traducciones al idioma inglés, aparte que al ser idiomas ajenos al conocimiento del autor del proyecto no hay un elemento de juicio que permita validar o valorar la calidad de la traducción.

Fig. 5: Embudo de depuración de reseñas

4.2.4 Traducción de reseñas

Después de algunas pruebas de ensayo-error se halla que el proceso de traducción automatizada de reseñas de español a inglés mediante la librería googletrans demanda recursos importantes de máquina y rupturas de proceso, se opta por una técnica de traducciones parciales por lotes que van salvando las traducciones en archivos de menor tamaño. Para esto se implementa un script el proceso de traducción por lotes (Ver anexo 1: preprocesing_data_hotels.ipynb).

En las siguientes imágenes se aprecia la totalidad de tiempo del proceso (19 horas y 25 minutos) y la salida de lotes procesados:

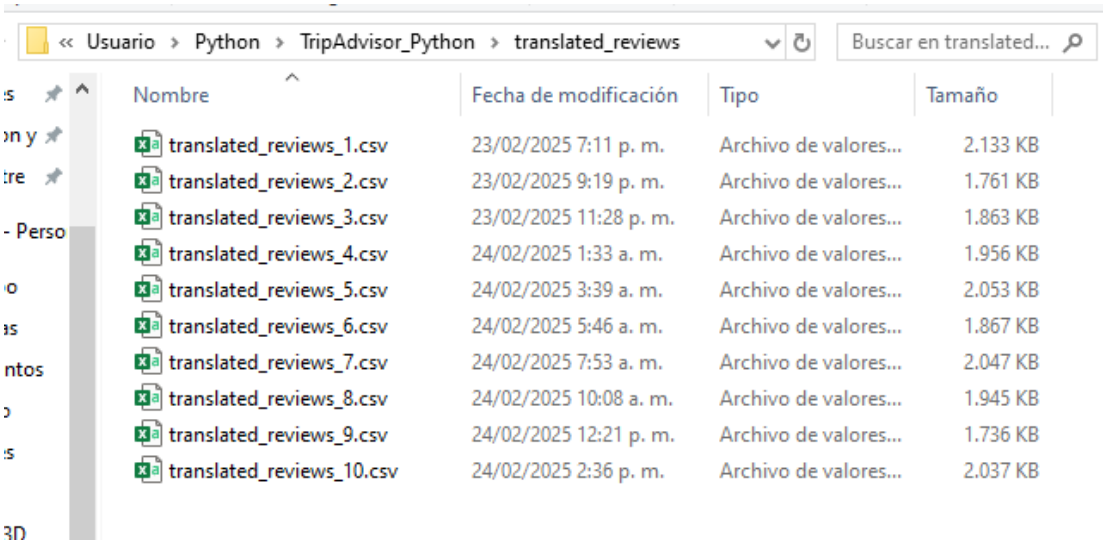


Fig. 6: Listado de archivos resultantes de reseñas traducidas

En la figura #7 se detalla el proceso de ejecución del script de traducción por lotes utilizado en la vista de Jupyter Notebook.

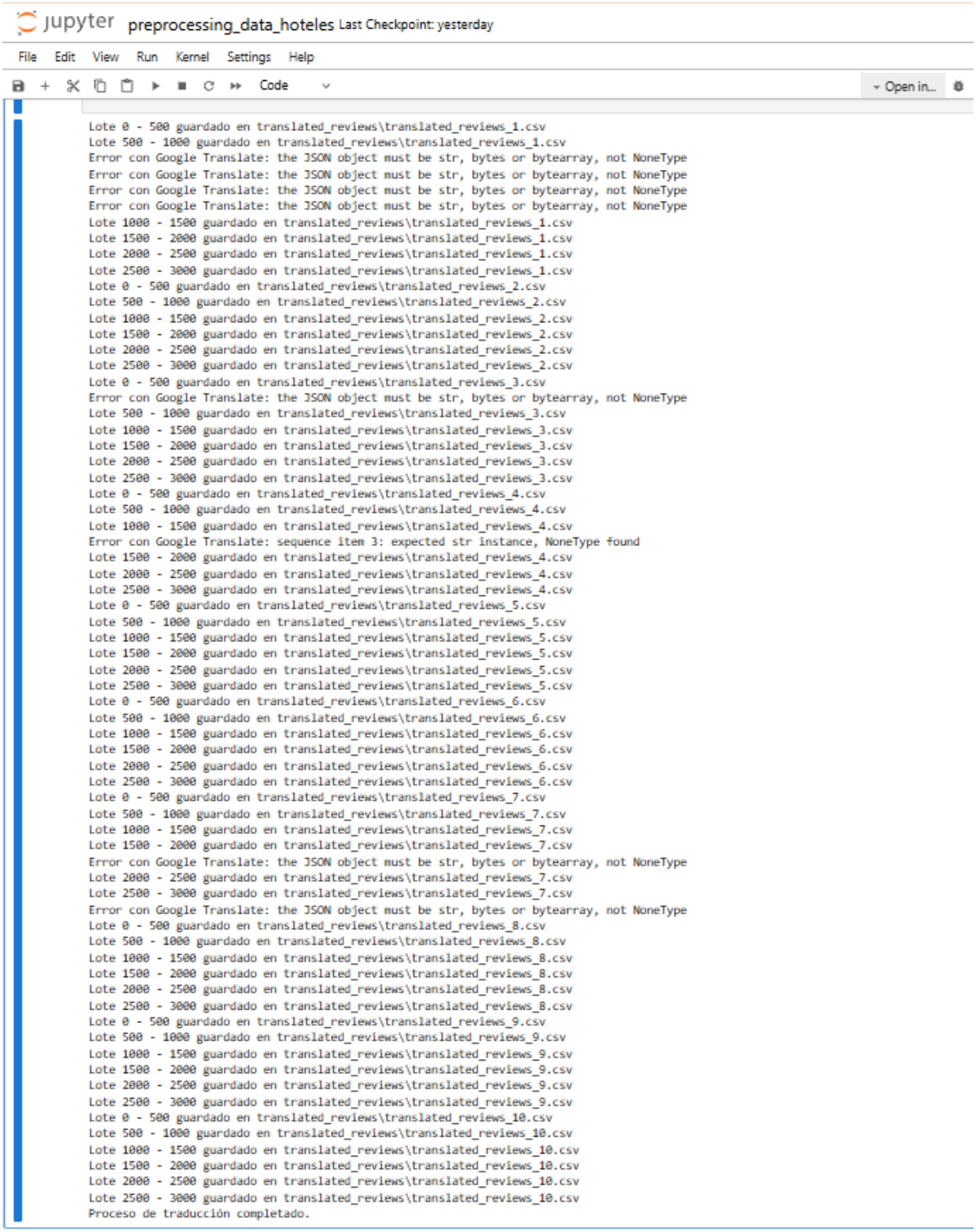


Fig. 7: Proceso de ejecución del script de traducción

Posteriormente se procedió a unir las reseñas traducidas de español a inglés con las reseñas en inglés en un nuevo dataframe y las reseñas unificadas al idioma inglés se guardan en una nueva base de datos csv (final_reviews.csv).

4.2.5 Conversión a minúsculas

A partir del archivo resultante final_reviews.csv se creó un nuevo dataframe df_final_reviews al cual se convierten a minúsculas todas las columnas de tipo texto: "Hotel_Name", "Review Title", “Review Text” y “Language” (anexo 1), para proceder con el análisis descriptivo.

4.3. ANÁLISIS DESCRIPTIVO

Se realizó una exploración preliminar de la información mediante análisis descriptivo al conjunto de datos final, el cual consta de 41056 filas y 9 columnas. Las cuales se describen a continuación:

Tabla IV. Descripción de variables

#	Nombre de Variable	Tipo	Cantidad	Descripción
1	Hotel_Name	String	41056	Nombre del hotel. Se cuenta con 49 hoteles diferentes
2	Review Id	Number	41056	Identificador único de cada reseña
3	User Location	String	23151	Ciudad de residencia del usuario
4	User Is Verified	String	41056	Variable categórica que indica si el usuario es verificado o no
5	Rating	Number	41056	Entero de 1 a 5 con el que el usuario califica el servicio
6	Published Date	String	41056	Fecha de publicación de la reseña
7	Language	String	41056	Idioma en que se publica la reseña
8	Review Title	String	41047	Título de la reseña
9	Review Text	String	41056	Contenido de la reseña

Se revisaron valores nulos. Se desestimaron los valores nulos de la variable User Location pues no es requerida por el sistema al momento que el usuario publica la reseña y no es de valor indispensable para el análisis.

Tabla V: Conteo de valores nulos

Nombre de Variable	Cantidad
Hotel_Name	0
Review Id	0
User Location	17905
User Is Verified	0
Rating	0
Published Date	0
Language	0
Review Title	0
Review Text	0

Se obtuvo un resumen estadístico de los datos numéricos con la descripción de la variable Rating del dataset.

Tabla VI: Descripción de variable Rating

Estadístico	Medida
Conteo	41056
Media	4.431118
Desviación estándar	0.948346
Mínimo	1.000000
25%	4.000000
50%	5.000000
75%	5.000000
Máximo	5.000000

También se obtuvo un resumen de los datos categóricos contenidos en las columnas de texto.

Tabla VII: Descripción de variables de texto

Nombre de Variable	Conteo	Valores únicos	Valor más frecuente	Cantidad del valor mas frecuente
Hotel_Name	41056	49	ghl hotel capital	2599
User Location	23151	2951	Bogota	3068
User Is Verified	41056	2	No	41055
Published Date	41056	5051	2019-09-23	42
Language	41056	2	es	26317
Review Title	41047	24287	excellent	1724
Review Text	41056	41150	i come to bogota maybe twice a year for busine...	2

4.3.1 Visualización

Por medio de las librerías Matplotlib y Seaborn se generaron algunas visualizaciones de la variable 'Rating' (Ver anexos 1 y 2).

4.3.2.1 Distribución de Calificaciones

El gráfico de columnas evidenció una concentración importante de las referencias con calificación positiva, sin embargo, se precisa (por la experticia en el área de mercadeo) que es en las referencias o reseñas negativas donde recae el riesgo de las crisis reputacionales, aun cuando estas sean muy pocas o casi nulas, tal como se citó en el aparte de antecedentes del presente proyecto. Las 5292 reseñas negativas (para este caso de análisis las puntuadas con rating 3, 2 o 1) corresponden al 13% del total disponible.

Tabla VIII: Conteo de calificaciones por variable Rating

Rating	Conteo
1	1153
2	1185
3	2954
4	9281
5	26483

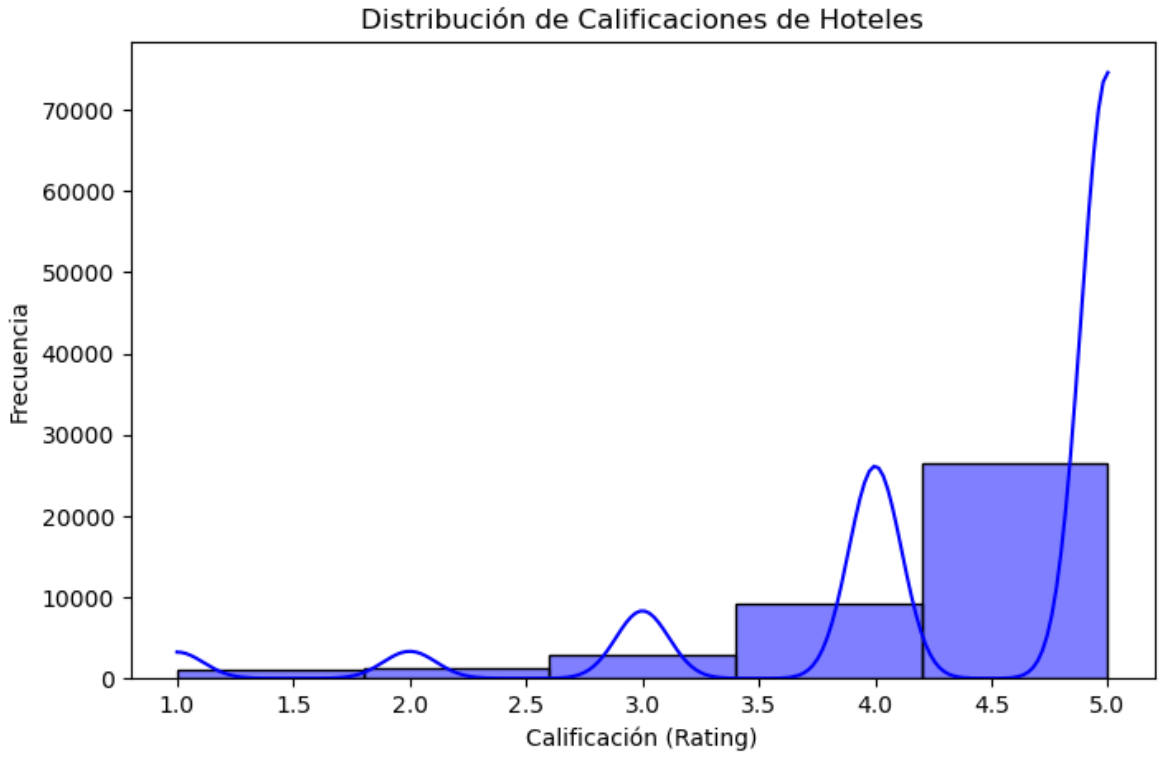


Fig. 8: Histograma de distribución de calificaciones (rating) por hoteles

4.3.2.2 Análisis de la Longitud de las Reseñas

Para un análisis preliminar de la longitud de las reseñas según la cantidad de palabras utilizadas por los usuarios en la columna “Text Review”, se recurrió a la herramienta Counter. Se evidenció que la mayoría de las reseñas se componen de 1 a 50 palabras y de 51 a 100 palabras, en un total de 25.633 y 9.887 reseñas respectivamente, que representan el 86.5% del total de reseñas.

Tabla IX. Tabla de frecuencias por longitud de reseñas

	Rango	Frecuencia	Frecuencia Acumulada	Frecuencia Relativa %	Frecuencia Relativa Acumulada %
1	1-50	25633	25633	62%	62%
2	51-100	9887	35520	24%	86%
3	101-150	2795	38315	7%	93%
4	151-200	1225	39540	3%	96%
5	201+	1516	41056	4%	100%

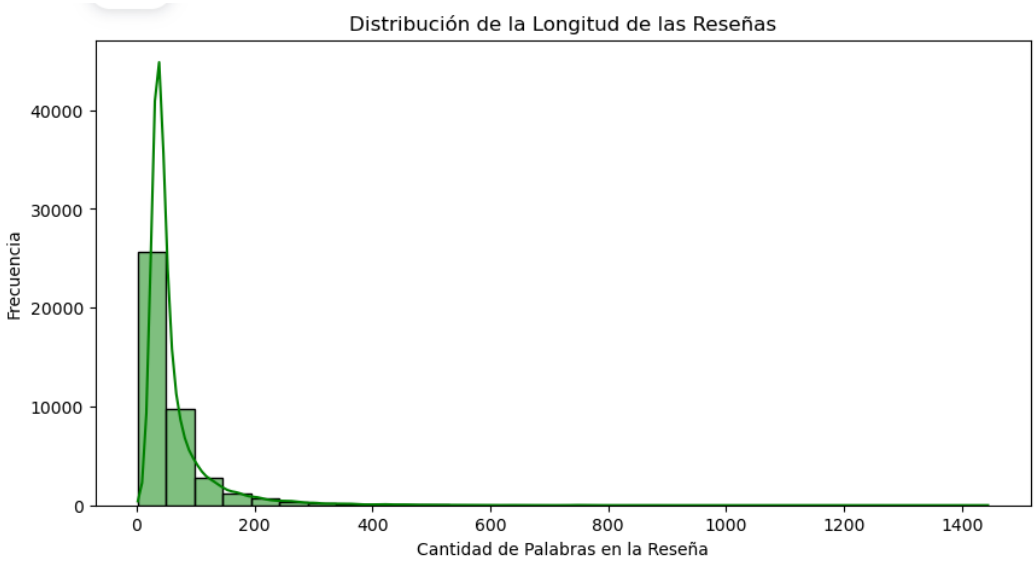


Fig. 9: Histograma de distribución de cantidad de palabras por reseña

4.3.2.3 Frecuencia de Palabras Más Usadas:

Para revisar la frecuencia de las palabras más usadas se utilizó el corpus Stopwords de la librería Nltk.

```
Palabras más frecuentes en Review Title:
[('excellent', 9238), ('hotel', 7341), ('good', 6452), ('service', 3841), ('great',
2783), ('stay', 2397), ('location', 1779), ('experience', 1751), ('bogota', 1624),
('best', 1385), ('attention', 1341), ('nice', 1153), ('place', 1021), ('comfortable',
886), ('business', 875), ('option', 866), ('bogotá', 790), ('well', 769), ('hotel,',
745), ('staff', 741)]
```

```
Palabras más frecuentes en Review Text:
[('hotel', 32087), ('good', 22664), ('service', 13897), ('room', 13598),
('excellent', 12933), ('staff', 11936), ('rooms', 9556), ('breakfast', 9510),
('stay', 8535), ('great', 7631), ('well', 7049), ('would', 6908), ('attention',
6839), ('one', 6590), ('comfortable', 6504), ('location', 6397), ('food', 6354),
('restaurant', 6053), ('friendly', 5500), ('recommend', 5212)]
```

4.3.3 Análisis de series temporales

Para el análisis de series temporales se empezó por una visión general del total de reseñas disponibles y posteriormente el enfoque se hizo solamente en aquellas que representan un riesgo de reputación, esto es las que representan calificaciones bajas en la columna de 'Rating'.

4.3.3.1 Análisis General de Fecha de Publicación

Después de un periodo escéptico entre 2004 y 2011, se reveló un aumento pequeño de la frecuencia de publicaciones entre 2012 y 2014, para luego presentar un aumento exponencial entre 2014 y 2017, manteniéndose relativamente estable entre 2014 y 2020, coincidiendo con la presencia de la pandemia por Covid 19 que afectó al planeta entero y que golpeó con fuerza particular al sector hotelero. La cantidad de reseñas tomó luego un comportamiento creciente sostenido desde 2021 hasta 2024, sin aún haber llegado a los mismos indicadores pre-pandemia (ver anexo 2).

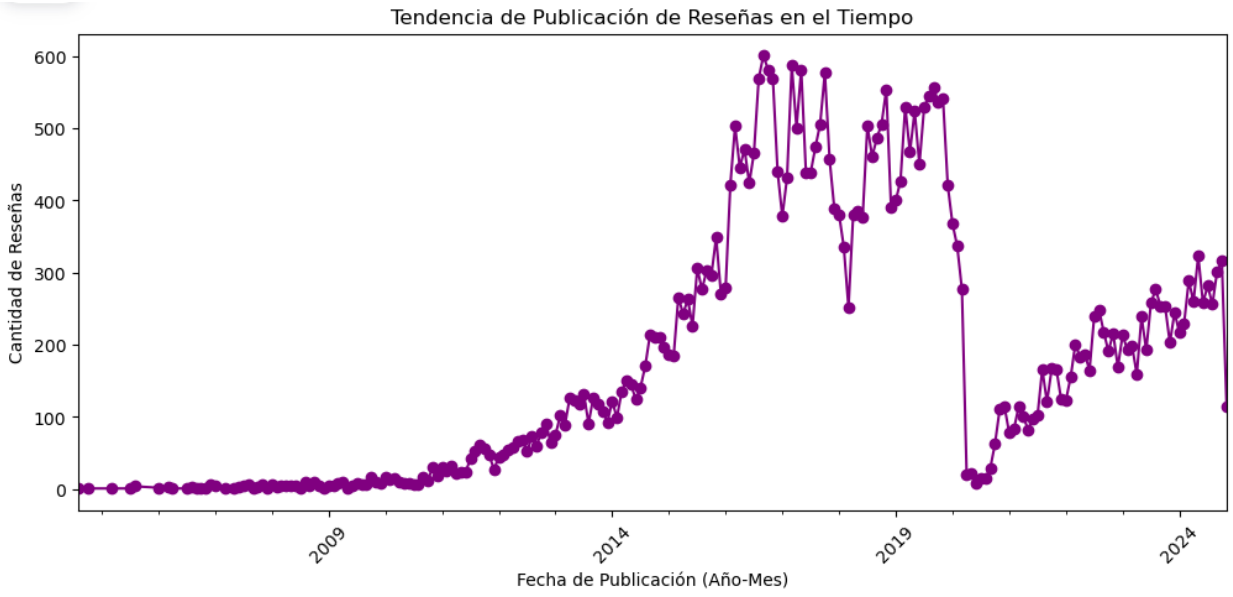


Fig. 10: Línea de tendencia de publicación de reseñas en el tiempo

4.3.3.2 Filtrado de reseñas negativas: Se filtraron las reseñas negativas para analizar su comportamiento mes a mes. A continuación se presenta la tendencia de publicación de reseñas negativas durante todo el periodo disponible, desde el año 2005 al 2024 (ver anexo 2).

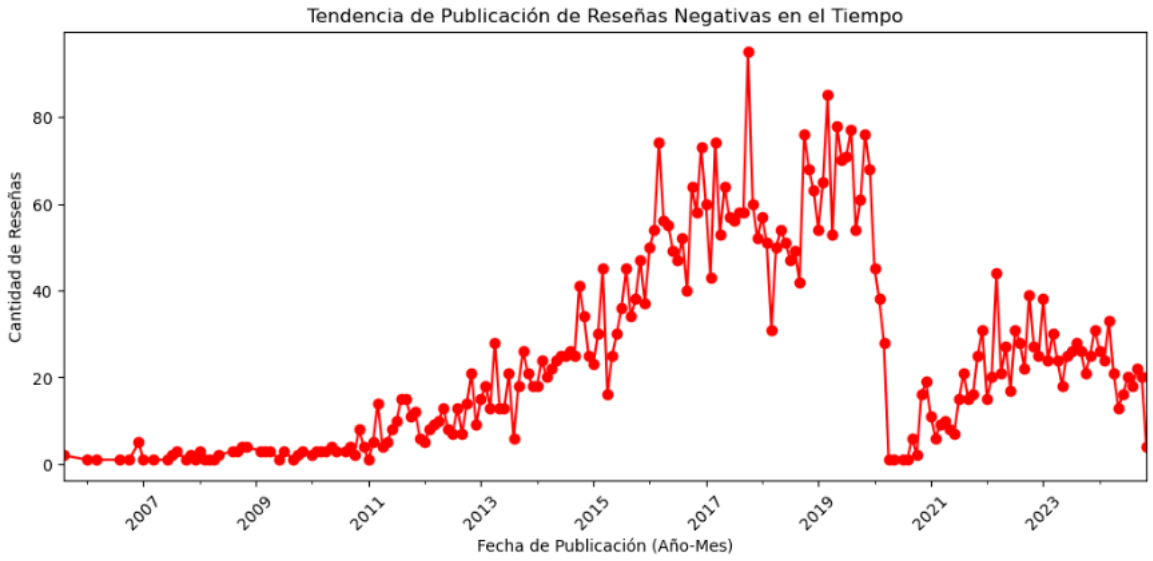


Fig. 11: Línea de tendencia de publicación de reseñas negativas en el tiempo

Como se evidenció un comportamiento similar (esperado) en cuanto al comportamiento de las reseñas totales y las negativas, y visto el evidente efecto atípico de la pandemia en el comportamiento de estas, se hizo necesario incorporar al script parámetros temporales de fecha_inicio y fecha_fin para facilitar el filtrado en diferentes periodos de tiempo.

En este sentido se procedió a filtrar el periodo posterior a la pandemia para analizar el comportamiento de las reseñas en el quinquenio comprendido entre los años 2020-01-01 y 2024-10-30.

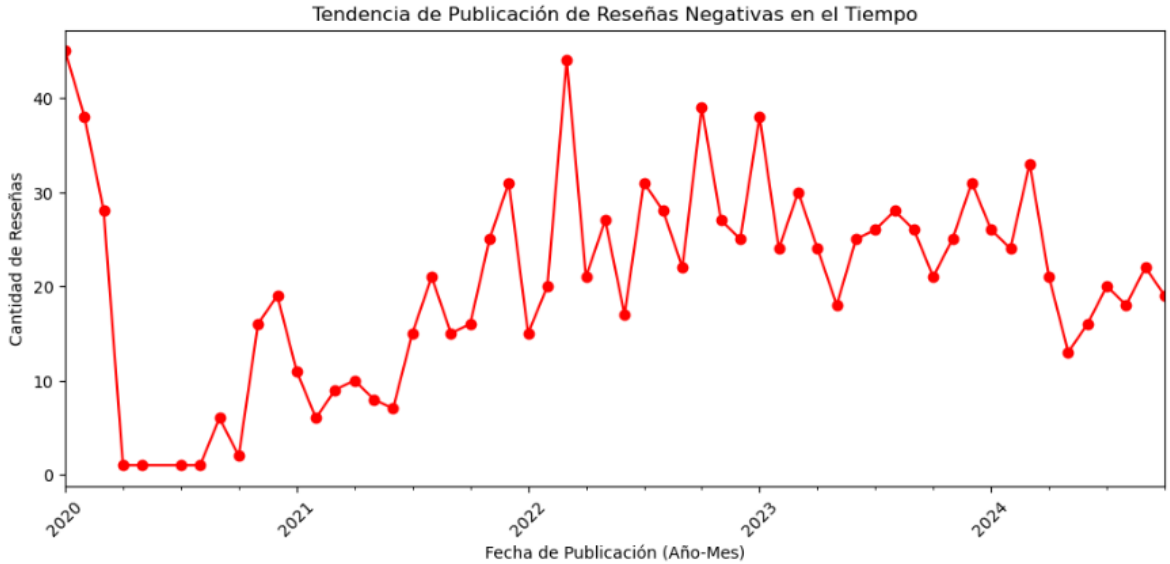


Fig 12: Tendencia de publicación de reseñas negativas en periodo acotado

4.3.3.3 Descomposición estacional (ver anexo 2)**:**

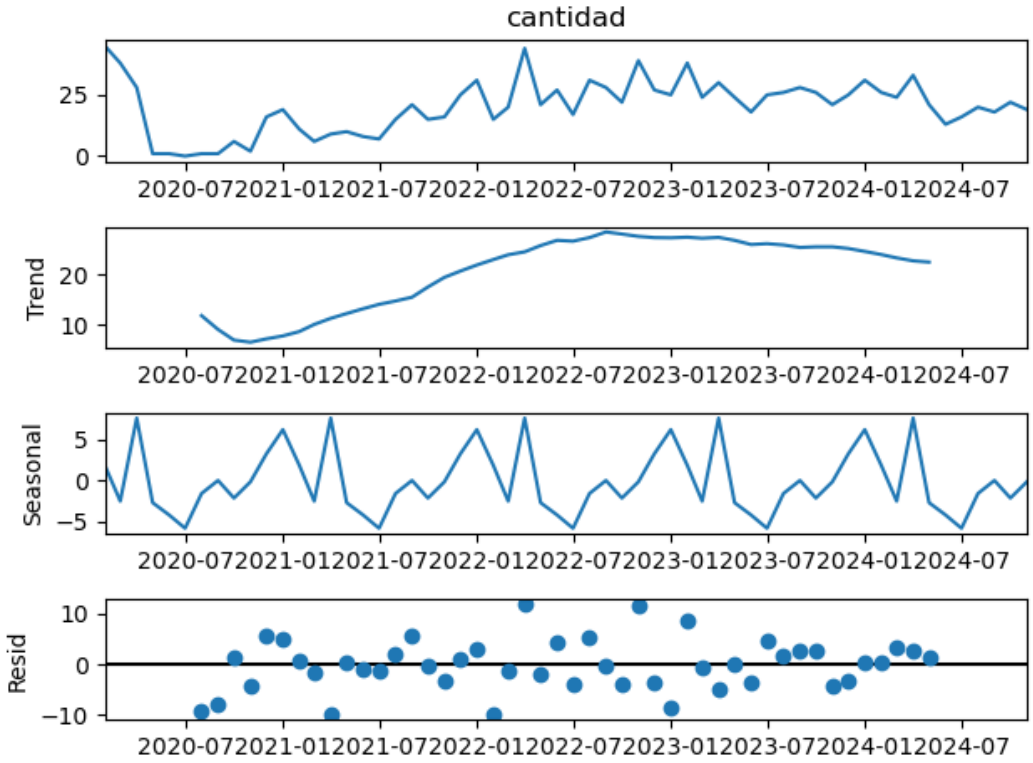


Fig. 13. Descomposición estacional

4.3.3.4 Explicación de los gráficos de descomposición estacional

Serie original (gráfico “cantidad”): Muestra el número total de reseñas negativas por período mensual desde 2020 hasta mediados de 2024. Se nota un repunte inicial en 2020 (rezago asociado a disrupciones por la pandemia COVID-19). Luego, hay un comportamiento fluctuante, con ciertos picos regulares. La serie es ruidosa, lo cual hace muy útil separar sus componentes para ver patrones más claros.

Tendencia (gráfico “Trend”): Este componente muestra la evolución general de largo plazo en la cantidad de reseñas negativas, suavizando las fluctuaciones de corto plazo. Desde mediados de 2020 hasta mediados de 2022 hay una tendencia ascendente, lo que sugiere un aumento sostenido de reseñas negativas (posiblemente asociado a reapertura de servicios o aumento en el volumen de viajeros). Desde mediados de 2022 hasta 2024 se ve una leve estabilización y luego descenso, lo que podría interpretarse como mejora en la experiencia del cliente o menor volumen de reseñas negativas.

Estacionalidad (gráfico “Seasonal”): Este componente representa patrones que se repiten con una frecuencia constante (en este caso, mensual). La forma cíclica muestra picos y valles que se repiten regularmente cada año, indicando que hay meses con mayor carga de reseñas negativas. Podría tratarse de temporadas altas con mayor presión operativa o meses donde el servicio baja su calidad percibida. Este patrón es clave para detectar meses críticos o recurrentes para activar alertas o reforzar el servicio.

Residuo (gráfico “Resid”): Este componente representa lo que queda fuera de la tendencia y estacionalidad: el “ruido” o eventos aleatorios e impredecibles. Aunque hay dispersión, no se observan outliers extremos. Sin embargo, algunos puntos negativos (fuertes caídas) podrían corresponder a eventos puntuales (crisis, quejas masivas, incidentes específicos).

4.3.3.5 Visualización del Promedio Acumulado Mensual (ver anexo 2)

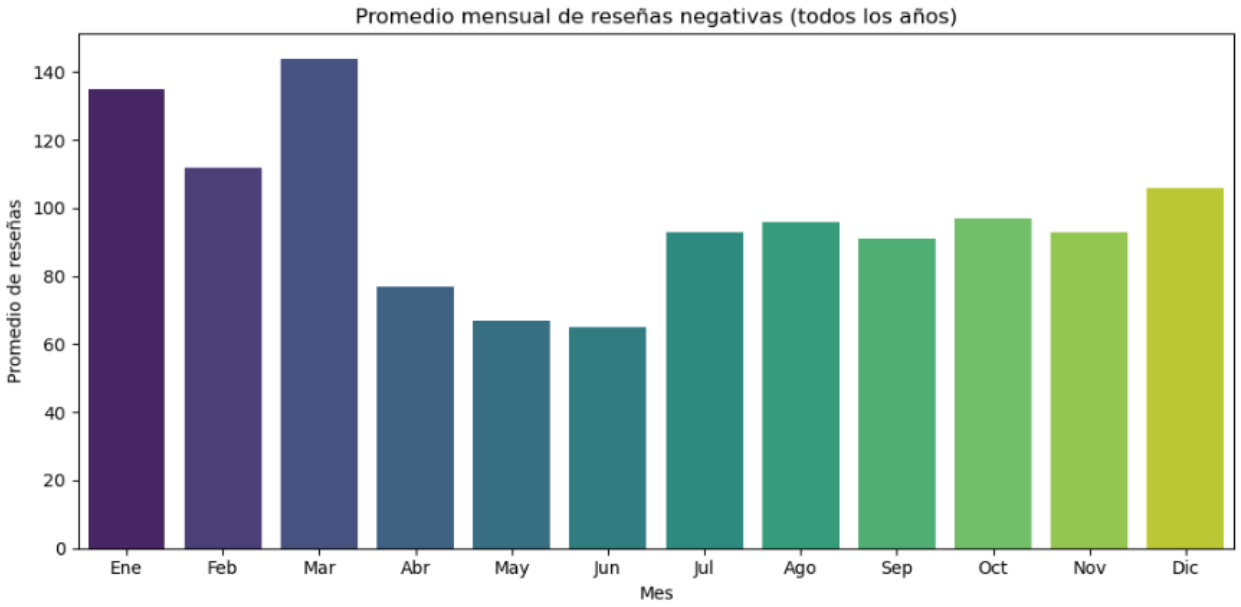


Fig. 14. Promedio mensual de reseñas negativas (años acumulados)

Esta visualización permitió evidenciar los meses del año con mayor incidencia de reseñas negativas en el periodo analizado (2020-2024). El primer trimestre del año correspondiente a los meses de enero a marzo son los de mayores incidencias negativas, mientras que el segundo trimestre (abril-mayo-junio) las incidencias negativas disminuyen y volviendo a un repunte intermedio en los trimestres tercero y cuarto del año.

4.3.3.6 Detección de meses críticos de reseñas negativas:

Usando un umbral basado en el percentil 75 se indicó una posible línea de alerta que permita tomar acciones correctivas a nivel de mercado a partir de la ocurrencia de una cantidad de reseñas negativas mayor al 75% de las ocurrencias.

Esto puede ser utilizado como herramienta de alerta temprana a nivel de gremio para anticiparse a posibles situaciones de crisis reputacionales.

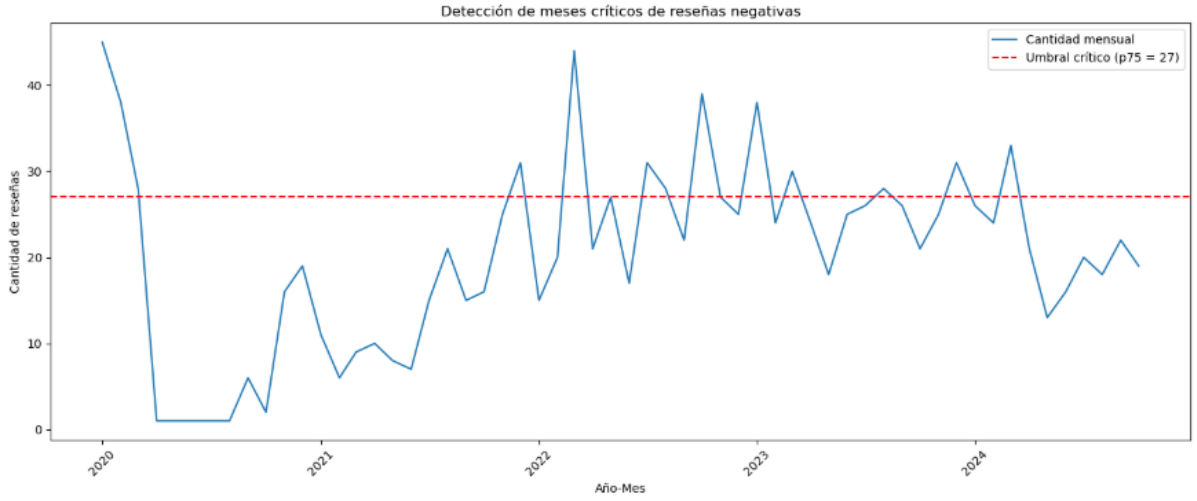


Fig. 15: Detección de meses críticos con umbral crítico de percentil 75

4.3.3.7 Análisis Particular de Fecha de Publicación

Las visualizaciones posteriores se enfocaron en gráficos mensuales de reseñas negativas por hotel, permitiendo identificar cuáles presentan patrones anómalos o señales de alerta (ver anexo 2).

En la gráfica de la figura 17 se detallan los 5 hoteles con mayor cantidad de reseñas negativas para el periodo analizado (2020-2024). Se evidencian picos de reseñas negativas para el hotel Habitel Select en el primer y tercer trimestres de 2022 y el tercer trimestre de 2024. El hotel Grand Hyatt Bogtá también presenta picos intermedios en la mitad del año 2021 y el tercer trimestre de 2022.

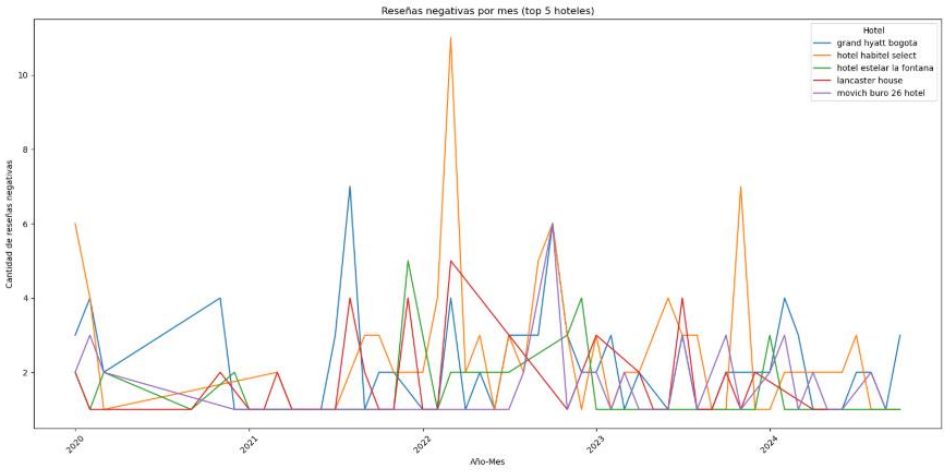


Fig. 16: Top 5 de reseñas negativas por hoteles en periodo acotado

También se generó gráficos individuales por hotel para los hoteles del top 5 de reseñas negativas del periodo analizado.

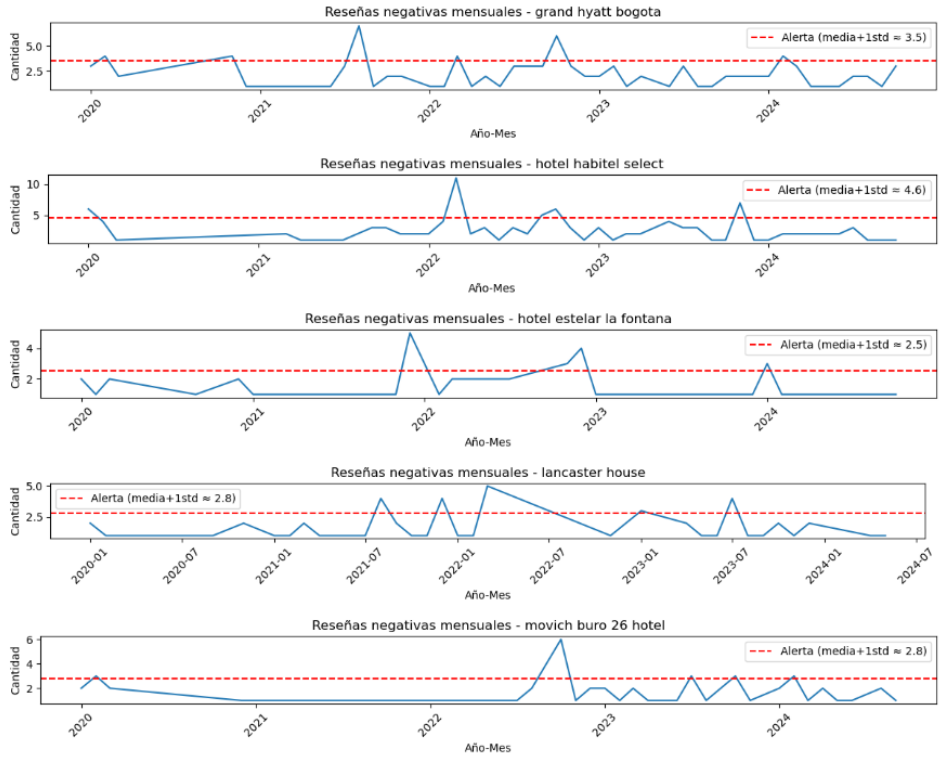


Fig. 17: Top 5 de reseñas negativas por hotel en periodo acotado

4.4. LEMMATIZATION

¿Stemming o Lematización?: La lematización es un proceso en el análisis de texto que busca reducir las palabras a su forma base o "lema". A diferencia de la stemming, que simplemente recorta las palabras eliminando sufijos, la lematización utiliza reglas lingüísticas y diccionarios para transformar una palabra a su forma canónica [31].

Dado que se están analizando reseñas para detectar crisis reputacionales mediante análisis de sentimiento, la lematización ayuda a:

- Reducir la variabilidad del lenguaje: "habitaciones", "habitación" y "habitacional" se pueden reducir a "habitación".
- Mejorar la precisión del análisis de sentimiento: Al estandarizar palabras, los modelos pueden identificar mejor patrones y asociaciones.
- Optimizar recursos computacionales: Se reduce el número de términos únicos, lo que facilita el procesamiento de grandes volúmenes de datos.

Tiene como ventajas que produce resultados más precisos y correctos desde el punto de vista lingüístico y considera el contexto de la palabra, por lo que es más preciso en términos de significado. En contrapartida es más lento y requiere un análisis morfológico y sintáctico más profundo y depende de un diccionario o modelo de lenguaje.

El proceso de lematización reduce las palabras a su forma base o lemática, considerando el contexto y la parte del discurso (por ejemplo, "running" se tomará como "run", "better" como "good").

4.4.1 Implementación de Lematización

Para el presente proyecto aplicado se realizó el proceso de lemmatización con la biblioteca spaCy, una de las más populares para NLP. Se aplicó a las variables que contienen texto utilizando el modelo de lenguaje en inglés en_core_web_sm (ver anexo 3).

Al código de lematización implementado se le incluyeron métricas para evaluar el tiempo de procesamiento, usando la biblioteca time para medir el tiempo total y el tiempo promedio por reseña, se añadió una nueva columna de texto lematizado llamada “Review Text Lemmatized” y se guardó el resultado en una base de datos csv (lemmatized_reviews.csv) para ser utilizada en el posterior proceso de vectorización (ver anexo 3).

```
Total de reseñas procesadas: 41056
Tiempo total de procesamiento: 471.14 segundos
Tiempo promedio por reseña: 0.0115 segundos
```

4.5. VECTORIZACIÓN

En el procesamiento de lenguaje natural (NLP), la vectorización es el proceso de convertir texto en representaciones numéricas que las máquinas pueden entender y procesar. Existen diversas técnicas para realizar esta conversión, entre las cuales destacan:

- Bag of Words (BoW): Representa el texto mediante la frecuencia de aparición de palabras en un documento, ignorando su contexto.
- TF-IDF (Term Frequency - Inverse Document Frequency): Asigna pesos a las palabras según su importancia en un conjunto de documentos.
- Word Embeddings (Word2Vec, GloVe, FastText, etc.): Representaciones densas y de alta dimensionalidad que capturan relaciones semánticas entre palabras.

Los Word Embeddings como Word2Vec son especialmente útiles porque representan palabras en un espacio vectorial donde palabras con significados similares tienen representaciones cercanas. Word2Vec tiene dos enfoques principales:

- CBOW (Continuous Bag of Words): Predice una palabra a partir de su contexto.
- Skip-gram: Predice el contexto a partir de una palabra dada.

Este método permite capturar relaciones semánticas y sintácticas entre palabras, lo que lo hace ideal para análisis de sentimientos como el que se realizó.

4.5.1 Implementación de vectorización:

Para la implementación del modelo de vectorización se utilizaron las librerías Pandas y Numpy, además de utilidad Word2Vec de la librería gensim y la utilidad word_tokenize de la librería NLTK.

La tokennización se aplicó a la nueva columna “Review Text Lemmatized” y se entrenó el modelo word2Vec con las reseñas lematizadas y se guardó para futuras referencias.

4.5.2 Entrenamiento del modelo utilizando un conjunto de datos etiquetado.

Se aplicó el embedding Word2Vec utilizando la biblioteca gensim. Se entrenó el modelo sobre las reseñas de hoteles ya lematizadas y luego se generaron representaciones vectoriales.

Paso seguido se ejecutó una prueba del funcionamiento del modelo ya entrenado, buscando en la lematización palabras similares al término ‘bad’ basado en los vectores obtenidos, con los siguientes resultados:

```
Palabras similares a 'bad':
[('terrible', 0.803429901599884), ('poor', 0.718224823474884), ('horrible',
0.6607425212860107), ('disappointing', 0.61928391456604), ('lousy',
0.6144813895225525), ('strange', 0.6103461384773254), ('regular',
0.5931856632232666), ('slow', 0.5896086692810059), ('unfortunately',
0.5733848810195923), ('unpleasant', 0.5404871106147766)]
```

Con el modelo Word2Vec ya entrenado, se representó cada reseña como un único vector en lugar de una lista de palabras. Para ello, se usó la técnica de "word embeddings promedio", que consiste en calcular el promedio de los vectores de todas las palabras de una reseña. Se convirtió cada reseña en un vector promedio, ignorando palabras no presentes en el modelo.

Se generó un nuevo DataFrame con las representaciones vectoriales y se guardaron los vectores en un archivo CSV (review_vectors.csv - ver anexo 4) para utilizarlo posteriormente en modelos de machine learning.

4.5.3 Exploración de los Vectores Generados de Reseñas

Una vez generados los vectores representativos de cada reseña en review_vectors.csv, se exploraron para entender su estructura y verificar su calidad antes de utilizarlos en modelos de Machine Learning.

Para ello, se cargaron los vectores en un DataFrame para revisar si los valores son numéricos y no hay datos faltantes y que la cantidad de columnas coincidiera con la dimensión del modelo Word2Vec (en este caso 100).

Para obtener la información general del DataFrame se usó df_vectors.info() para saber cuántas filas y columnas se obtuvieron y si hay valores nulos.

Con df_vectors.describe() se mostró el rango de valores, la media y la desviación estándar de cada dimensión del vector.

Tabla X. Estadísticos del dataframe de vectorización

Estadística	0	1	2	3	4
count	41.056.000	41.056.000	41.056.000	41.056.000	41.056.000
mean	0.461391	-0.282568	-0.170359	0.042132	-0.060475
std	0.164059	0.286283	0.177368	0.179409	0.189805
min	-1.013.138	-1.287.396	-1.428.294	-0.722966	-0.822665
25%	0.356177	-0.486013	-0.283947	-0.073819	-0.190197
50%	0.458871	-0.314087	-0.169685	0.049480	-0.066249
75%	0.566384	-0.103926	-0.057683	0.165127	0.064503
max	1.497.245	1.308.827	0.890431	0.862245	0.783738

Posteriormente se revisó si habían filas con vectores de ceros. Si alguna reseña no contenía palabras en el vocabulario de Word2Vec, se le asignó un vector de ceros. Al verificar cuántos casos resultaron de este tipo se encontró que no habían casos en ceros, lo que permitió validar que el modelo Word2Vec cubrió bien el vocabulario.

4.5.4 Visualización de los Vectores en 2D Usando Reducción de Dimensiones

Como los vectores generados tienen muchas dimensiones (en este caso 100), se redujeron a 2 dimensiones usando PCA (Análisis de Componentes Principales) o t-SNE para visualizar cómo se agrupan las reseñas.

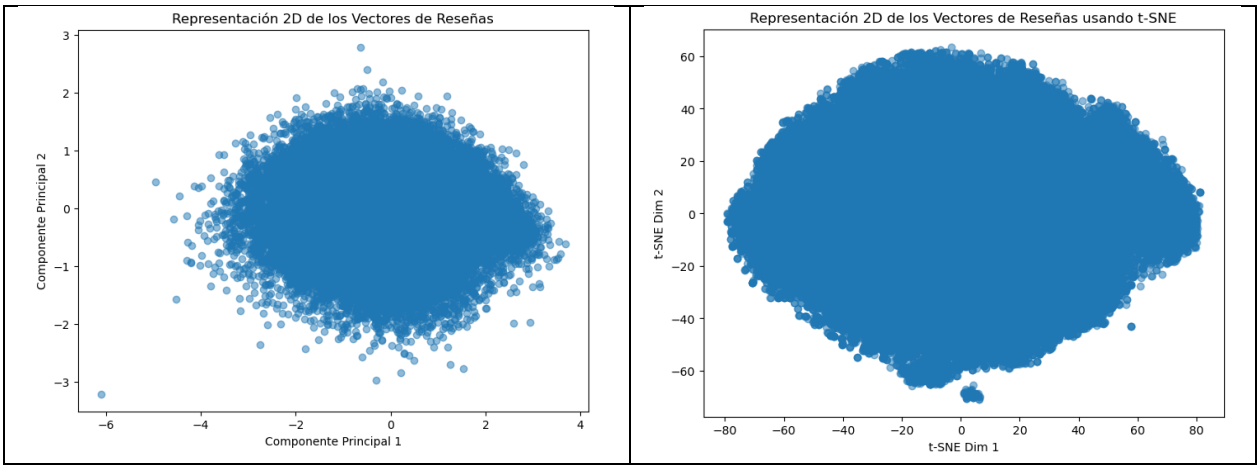


Fig. 18: Representación 2D de los vectores de reseñas usando PCA y t-SNE

4.5.5 Medir la Distancia Entre Reseñas

Finalmente se calculó qué tan similares son dos reseñas midiendo la distancia entre sus vectores con la métrica de coseno:

```
Similitud entre la reseña 1 y la reseña 2: 0.7916
```

Tomando en cuenta que valores cercanos a 1 indican reseñas muy similares y valores cercanos a 0 indican reseñas diferentes, se encontró una similitud importante.

Realizadas estas pruebas, se determinó que el corpus de datos estaba listo para ser probado en modelos de machine learning.

5. CONSTRUCCIÓN Y REVISIÓN DEL MODELO DE MACHINE LEARNING

5.1. SELECCIÓN DEL MODELO DE CLASIFICACIÓN

En este punto del proceso, se hizo necesario decidir cual modelo de clasificación sería en apropiado para abordar el proceso: bivariado o multivariado, tomando en cuenta que se tiene una columna referente llamada 'Rating' que contiene calificaciones asignadas por los usuarios en dígitos de 1 a 5, donde los valores menores significan insatisfacción y los valores mayores satisfacción frente al servicio recibido.

Dado que la situación de análisis se centra en determinar posibles patrones de ocurrencia de crisis, se opta por la opción bivariada, definiendo solo dos posibles escenarios a partir de las calificaciones dadas, esto es, escenario 1: crisis, escenario 2: no crisis.

5.1.1 Definir las Etiquetas Binarias (Crisis vs No Crisis)

Dado que Rating es un valor de 1 a 5, se asignan las etiquetas de la siguiente forma:

Crisis (1) → Reseñas con puntuaciones bajas (por ejemplo, 1, 2 y 3).

No Crisis (0) → Reseñas con puntuaciones altas (por ejemplo, 4 y 5).

Utilizando la base de datos vectorizada y la base de datos lematizada, se implementó una función que asignara la etiqueta de crisis (1) o no crisis (0) a cada reseña mediante la columna de 'rating' y el identificador común id, que permitiendo discriminar de esta manera el total de reseñas de caso crisis o no crisis con el siguiente resultado:

Tabla XI. Conteo de reseñas por etiqueta

Etiqueta	Conteo
0 (No crisis)	35764
1 (crisis)	5292

Para los correspondientes análisis posteriores en las matrices de confusión resultantes, se tomó en cuenta que el escenario de crisis detectado habría de considerarse como un verdadero positivo, mientras que un escenario de no crisis detectado sería tomado como un verdadero negativo.

5.2. SELECCIÓN DE ALGORITMOS DE CLASIFICACIÓN

Para alcanzar el objetivo principal del presente estudio, detectar patrones predictivos de crisis reputacionales en hoteles a partir del análisis de reseñas textuales, se optó por entrenar y comparar el desempeño de cuatro algoritmos de aprendizaje supervisado: Regresión Logística, Random Forest, Support Vector Machine (SVM) y MLPClassifier (Multilayer Perceptron). La elección de estos modelos respondió a criterios de diversidad metodológica, rendimiento contrastado en tareas de clasificación textual y potencial de escalabilidad para aplicaciones futuras.

La Regresión Logística se incluyó por tratarse de un modelo lineal clásico ampliamente utilizado como punto de referencia en problemas de clasificación binaria. Su capacidad para proporcionar resultados interpretables a través de coeficientes y su eficiencia en contextos con relaciones lineales entre variables justifican su presencia en este conjunto comparativo.

Por su parte, Random Forest constituye un modelo de tipo ensemble que combina múltiples árboles de decisión. Es especialmente eficaz en la captura de relaciones no lineales y en la gestión de datos con ruido o alta dimensionalidad, como los vectores derivados del procesamiento de texto. Su robustez frente al sobreajuste lo convierte en una opción confiable para tareas de clasificación complejas.

El algoritmo Support Vector Machine (SVM) fue seleccionado por su buen rendimiento en espacios vectoriales de alta dimensión, condición común en el análisis de texto vectorizado. Su enfoque basado en la maximización del margen entre clases y su flexibilidad a través del uso de diferentes núcleos (kernels) permiten abordar problemas donde las fronteras entre categorías no son linealmente separables.

Finalmente, se incluyó el MLPClassifier, una red neuronal de tipo feedforward, que destaca por su capacidad para modelar relaciones no lineales complejas. Aunque su interpretación resulta menos intuitiva, su adaptabilidad y capacidad de aprendizaje lo convierten en una alternativa idónea para escenarios donde se prevé la incorporación de características adicionales, como metadatos, representaciones semánticas del texto (embeddings), o información temporal.

En conjunto, estos cuatro algoritmos representan una selección balanceada de enfoques que van desde modelos lineales interpretables hasta arquitecturas más complejas y escalables. Esta diversidad metodológica permitió evaluar el problema desde distintas perspectivas, facilitando no solo la identificación del modelo con mejor desempeño actual, sino también la exploración de opciones robustas para aplicaciones futuras de mayor complejidad.

También se tomó como metodología entrenar el modelo inicialmente con los parámetros por defecto de cada algoritmo de clasificación y luego ajustando hiperparámetros con la ayuda de la librería GridSearchCV.

5.3. ENTRENAMIENTOS PRELIMINARES DE LOS ALGORITMOS CON PARÁMETROS POR DEFECTO

5.3.1 Distribución de los conjuntos de datos

Para la ejecución de los modelos sin optimización de hiperparámetros, se utilizó un esquema clásico de división del conjunto total de datos: 80% para entrenamiento y 20% para prueba (test). En este caso, no se hizo validación separada, sino solo se evaluó el rendimiento del modelo final con el conjunto de test 20% (Ver anexo 5).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

5.3.2 Acotación importante para el posterior análisis de la matríz de confusión

Por cuanto se ha definido la clase 1 como crisis reputacional (es decir, sentimiento negativo), y la clase 0 como no crisis reputacional (es decir, sentimiento positivo o neutral); de esta manera:

- Una reseña con sentimiento negativo correctamente clasificada como clase 1 es un Verdadero Positivo (VP).
- Una reseña con sentimiento negativo clasificada erróneamente como clase 0 es un Falso Negativo (FN).
- Una reseña con sentimiento positivo o neutral correctamente clasificada como clase 0 es un Verdadero Negativo (VN).
- Una reseña con sentimiento positivo o neutral clasificada erróneamente como clase 1 es un Falso Positivo (FP).

Por lo tanto, cuando el modelo encuentra una reseña de sentimiento positivo o neutral y correctamente la clasifica como no crisis (clase 0), eso se cuenta como un Verdadero Negativo (VN), todo esto dentro del contexto del modelo y la forma en que se definió la etiqueta de salida, porque en el proyecto la clase 1 representa "crisis" (sentimiento negativo), y la clase 0 representa "no crisis" (sentimiento positivo o neutro).

5.3.2 Modelo Regresión Logística (parámetros por defecto):

Tomamos como base los parámetros por defecto que aplica el modelo MLPClassifier, según la documentación misma de la librería Scikit Learn [32]:

Tabla XII. Parámetros por defecto Regresión Logística:

Parámetro	Descripción técnica	Valores por defecto
C	Inversa de la regularización L2	Valor por defecto: 1.0 Mayor C = menor regularización Menor C = evita sobreajuste
solver	Algoritmo de optimización	lbfgs' (por defecto): eficiente para datasets medianos
penalty	Tipo de regularización aplicada	l2' (por defecto): penalización de tipo Ridge
max_iter	Número máximo de iteraciones	Por defecto: 100. Se puede aumentar si el modelo no converge.

5.3.3 Resultados del modelo de Regresión Logística (con parámetros por defecto)

Tabla XIII. Resultados del modelo de Regresión Logística (parámetros por defecto)

Métrica / Clase	Precision	Recall	F1-Score	Support
Clase 0 (No crisis)	0.94	0.97	0.95	7143
Clase 1 (Crisis)	0.71	0.58	0.64	1069
Exactitud global (accuracy)	—	—	**0.9149**	8212
Promedio macro	0.83	0.77	0.80	8212
Promedio ponderado	0.91	0.91	0.91	8212

Tabla XIV. Matriz de Confusión Regresión logística (parámetros por defecto)

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6894	Falsos Positivos (FP) 249
Clase 1 (Crisis)	Falsos Negativos (FN) 450	**Verdaderos Positivos (VP)** 619

5.3.4 Análisis de Resultados Regresión Logística (parámetros por defecto)

Tabla XV: Análisis de Resultados Regresión Logística (parámetros por defecto)

Métrica	Valor	Interpretación
Precisión (precision)	0.71	De todas las predicciones etiquetadas como crisis, el 71% realmente lo eran. Es una buena señal para evitar falsos positivos.
Exhaustividad (recall)	0.58	El modelo solo detecta el 58% de los casos reales de crisis. Esto implica una cantidad significativa de falsos negativos.
F1-score	0.64	Promedio armónico entre precisión y recall. Puede considerarse aceptable, pero hay margen de mejora.

La exactitud total (Accuracy) del 91.49% es una cifra alta, pero puede ser engañosa dado el desbalance de clases (muchas más no crisis que crisis). El recall bajo (0.58) es una debilidad importante en contextos donde detectar las crisis reales es crucial. Podría estar dejando pasar más del 40% de los eventos críticos.

En cuanto a la matriz de confusión:

- Falsos positivos (FP): 249 casos fueron etiquetados como crisis sin serlo.
- Falsos negativos (FN): 450 crisis reales no fueron detectadas como tal.

Este balance indica que el modelo es más conservador (tiende a indicar que no hay crisis cuando si las hay), lo que en un sistema de alerta temprana podría significar una falla grave si el objetivo es actuar preventivamente.

En conclusión, el modelo de Regresión Logística tiene un buen desempeño global, pero un recall bajo para la clase de crisis, lo que es un problema, pues el objetivo es detectar todos los eventos de crisis posibles, aunque hayan equivocaciones con algunos falsos positivos. Este modelo podría ser una buena línea base, pero no es suficiente para producción sin mejoras (como balanceo de clases, tuning de umbrales, o mejor optimización).

5.3.5 Modelo Random Forest (parámetros por defecto):

Tomamos como base los parámetros por defecto que aplica el modelo Random Forest, según la documentación misma de la librería Scikit Learn [33]:

Tabla XVI. Parámetros por defecto de Random Forest

Parámetro	Descripción técnica	Valor por defecto
n_estimators	Número de árboles en el bosque	100
max_depth	Profundidad máxima de cada árbol	None (los nodos se expanden hasta que todas las hojas son puras o contienen menos de min_samples_split)
min_samples_split	Número mínimo de muestras requeridas para dividir un nodo interno	2
min_samples_leaf	Número mínimo de muestras en una hoja	1
max_features	Número de características consideradas al dividir un nodo	'sqrt' (raíz cuadrada del número total de características)
bootstrap	Si se deben usar muestras con reemplazo	True

5.3.6 Resultados del modelo Random Forest (con parámetros por defecto)

Tabla XVII. Resultados Random Forest (parámetros por defecto)

Métrica / Clase	Precisión	Recall	F1-Score	Support
Clase 0 (No crisis)	0.92	0.97	0.95	7143
Clase 1 (Crisis)	0.72	0.45	0.56	1069
Exactitud global	—	—	**0.9062**	8212
Promedio macro	0.82	0.71	0.75	8212
Promedio ponderado	0.90	0.91	0.90	8212

Tabla XVIII. Matriz de Confusión Random Forest (parámetros por defecto)

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6956	Falsos Positivos (FP) 187
Clase 1 (Crisis)	Falsos Negativos (FN) 583	**Verdaderos Positivos (VP)** 486

5.3.7 Análisis de resultados de Random Forest (con parámetros por defecto)

Tabla XIX: Análisis de resultados de Random Forest (parámetros por defecto)

Métrica	Valor	Interpretación
Precisión (precision)	0.72	Muy similar a la de la regresión logística (0.71). Muestra buena capacidad para evitar falsos positivos.
Exhaustividad (recall)	0.45	Sensiblemente **peor** que la regresión logística (0.58). Detecta solo el 45% de las crisis reales.
F1-score	0.56	Más bajo que en RL (0.64). Refleja el pobre equilibrio entre precisión y recall.

El recall bajo (aún menor que en la regresión logística) es preocupante si el foco del proyecto es la detección de crisis. Este modelo se muestra más conservador aún que la regresión logística. La exactitud total (Accuracy) de 90.62% es muy cercana a la de la Regresión Logística (91.49%), aunque ligeramente inferior.

Analizando la matriz de confusión:

- Falsos positivos (FP): 187 (menos que RL → 249) casos fueron etiquetados como crisis sin serlo.
- Falsos negativos (FN): 583 (más que RL → 450) crisis reales que no fueron detectadas como tal.

El modelo comete menos errores de alarma falsa, pero deja escapar más crisis reales que la regresión logística, lo que es más grave para un sistema de detección proactiva.

En conclusión, el modelo muestra una precisión aceptable, pero un recall claramente deficiente para la clase de crisis. Detecta menos crisis reales que la Regresión Logística pero comete menos falsas alarmas. No sería adecuado como modelo final sin mejoras significativas (reescalado de clases, optimización de hiperparámetros, técnicas como SMOTE o modificación del umbral de decisión).

5.3.8 Modelo Support Vector Machines SVM (parámetros por defecto):

Tomamos como base los parámetros por defecto que aplica el modelo SVM, según la documentación misma de la librería Scikit Learn [34]:

Tabla XX. Parámetros por defecto de SVM

Parámetro	Descripción técnica	Valor por defecto
C	Parámetro de penalización	1.0
kernel	Tipo de función núcleo	'rbf' (Radial Basis Function)
gamma	Coeficiente del kernel RBF o polinomial	'scale' (1 / (n_features * X.var()))
degree	Grado del kernel polinómico	3
probability	Habilita la estimación de probabilidad (predict_proba)	False
shrinking	Usa heurística de reducción	True
max_iter	Número máximo de iteraciones	-1

5.3.9 Resultados del modelo de SVM (con parámetros por defecto)

Tabla XXI. Resultados del modelo SVM con parámetros por defecto

Métrica / Clase	Precision	Recall	F1-Score	Support
Clase 0 (No crisis)	0.94	0.97	0.95	7143
Clase 1 (Crisis)	0.72	0.59	0.65	1069
Exactitud global	—	—	**0.9163**	8212
Promedio macro	0.83	0.78	0.80	8212
Promedio ponderado	0.91	0.92	0.91	8212

Tabla XXII. Matriz de Confusión SVM con parámetros por defecto

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6894	Falsos Positivos (FP) 249
Clase 1 (Crisis)	Falsos Negativos (FN) 438	**Verdaderos Positivos (VP)** 631

5.3.10 Análisis de resultados de SVM (parámetros por defecto)

Tabla XXIII: Análisis de Resultados SVM (parámetros por defecto)

Métrica	Valor	Interpretación
Precisión	0.72	Igual que RF y RL. El modelo se mantiene constante en evitar falsos positivos.
Recall	0.59	Mejor que RF (0.45) y ligeramente mejor que RL (0.58). Esto implica que detecta más casos reales de crisis.
F1-score	0.65	El más alto hasta ahora para clase 1, señal de mejor balance entre precisión y recall.

La exactitud total (Accuracy) de 91.63% es el valor más alto hasta ahora entre los tres modelos evaluados, aunque la diferencia con la Regresión Logística (91.49%) es marginal.

En cuanto a la matriz de confusión:

Falsos positivos (FP): 249 (igual que RL)

Falsos negativos (FN): 438 (mejor que RF y RL)

SVM mantiene un nivel de falsos positivos moderado, pero reduce los falsos negativos respecto a los otros modelos, lo que lo hace más útil si la prioridad es detectar correctamente las crisis sin incrementar demasiado las falsas alarmas.

En conclusión, SVM logra el mejor F1-score para la clase minoritaria (crisis) hasta ahora. Tiene la mejor combinación entre precisión y recall de los tres modelos evaluados. La exactitud global es la más alta, aunque solo por unas décimas. Requiere optimización posterior (por ejemplo, con el parámetro C o el tipo de kernel), pero ya muestra un excelente punto de partida.

5.3.11 Modelo de Redes Neuronales MLPClassifier (parámetros por defecto)

Tomamos como base los parámetros por defecto que aplica el modelo MLPClassifier, según la documentación misma de la librería Scikit Learn [35]:

Tabla XXIV. Parámetros por defecto del modelo de redes neuronales MLPClassifier

Parámetro	Descripción técnica	Valor por defecto
hidden_layer_sizes	Tamaño de las capas ocultas	(100,) (una capa oculta con 100 neuronas)
activation	Función de activación	'relu' (Rectified Linear Unit)
solver	Algoritmo de optimización	'adam'
alpha	Parámetro de regularización L2	0.0001
learning_rate	Política de tasa de aprendizaje	'constant'
learning_rate_init	Tasa de aprendizaje inicial	0.001
max_iter	Máximo de iteraciones	200
early_stopping	Detiene entrenamiento si no mejora	False
random_state	Semilla aleatoria	None

5.3.12 Resultados del modelo de MLPClassifier (con parámetros por defecto)

Tabla XXV. Resultados del modelo de MLPClassifier (con parámetros por defecto)

Métrica / Clase	Precision	Recall	F1-Score	Support
Clase 0 (No crisis)	0.94	0.94	0.94	7143
Clase 1 (Crisis)	0.62	0.63	0.62	1069
Exactitud global	—	—	**0.9009**	8212
Promedio macro	0.78	0.79	0.78	8212
Promedio ponderado	0.90	0.90	0.90	8212

Tabla XXVI. Matriz de Confusión MLP Classifier (con parámetros por defecto)

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6725	Falsos Positivos (FP) 418
Clase 1 (Crisis)	Falsos Negativos (FN) 396	**Verdaderos Positivos (VP)** 673

5.3.14 Análisis de resultados de Redes Neuronales MLPClassifier (parámetros por defecto)

Tabla XXVII: Análisis de Resultados Redes Neuronales MLPClassifier (parámetros por defecto)

Métrica	Valor	Comparación con modelos anteriores
Precisión	0.62	Inferior a SVM, RL y RF (todos ~0.71–0.72)
Recall	0.63	**La más alta** entre todos los modelos, lo que implica que detecta más crisis reales.
F1-score	0.62	Menor que SVM (0.65) y RL (0.64), pero mejor que RF (0.56).

La exactitud total (Accuracy) de 90.09% es el valor más bajo de los 4 modelos, aunque sigue siendo una precisión aceptable en términos generales.

Analizando la matriz de confusión:

Falsos positivos (FP): 418 (el más alto de todos)

Falsos negativos (FN): 396 (el más bajo de todos)

Este modelo sacrifica precisión (más falsos positivos) a cambio de mejorar el recall (menos falsos negativos). Si el objetivo del proyecto fuera minimizar el riesgo de no detectar una crisis real, este comportamiento podría ser útil. Sin embargo, esto viene a costa de generar más alarmas falsas.

En conclusión, el modelo MLPClassifier es el modelo más "sensible" (recall más alto) a eventos de crisis. Tiene un balance menos equilibrado que SVM, debido a su menor precisión. Podría ser una opción válida si se optara por sobre-alertar a sub-alertar, especialmente en contextos donde las consecuencias de omitir una crisis son graves. El desempeño podría mejorar mucho con ajuste de hiperparámetros (tamaño de capa, activación, optimizador, tasa de aprendizaje).

5.4. RETROALIMENTACIÓN

En una comparativa preliminar de desempeño de los 4 modelos teniendo como foco la detección de clase 1 – crisis, se presenta la siguiente tabla:

Tabla XXVIII: Comparativo de resultados por modelo (parámetros por defecto)

Modelo	Precisión	Recall	F1-score	Exactitud
Logística	0.71	0.58	0.64	91.49%
RandomForest	0.72	0.45	0.56	90.62%
SVM	0.72	0.59	**0.65**	**91.63%**
MLP	0.62	**0.63**	0.62	90.09%

En este punto del proceso podría sugerirse considerar como mejor opción al modelo SVM por su aparente mejor equilibrio y precisión, aunque el modelo que detecta la mayor cantidad de crisis posible es el de redes neuronales MLPClassiffier.

Sin embargo, teniendo en cuenta la opción de aplicar la optimización de hiperparámetros, se procede a realizar una nueva ronda de entrenamiento del modelo con esta función.

En este punto tenemos tanto una opción manual de ajuste de hiperparámetros, como una función automatizada de ajuste con una librería ya existente para tal fín como lo es GridsearchCV. Dado que el ajuste manual de hiperparámetros de prueba significaría un ingente consumo de recursos de computación y tiempo, se optó por implementar la optimización mediante la mencionada herramienta GridsearchCV.

5.5. MEJORAMIENTO DEL MODELO Y LA SOLUCIÓN EN GENERAL

Tomando en cuenta los hiperparámetros por defecto anotados en sección anterior (5.3), se procedió a aplicar el ajuste de parámetros utilizando la herramienta GridSearchCV de la librería Scikit Learn, obteniendo su correspondiente análisis de hiperparámetros optimizados y resultados.

Para la optimización de hiperparámetros con GridSearchCV, se utilizó validación cruzada con 5 pliegues (5-fold CV) aplicada sobre el conjunto de entrenamiento. En cada iteración, se empleó ~64% del total de datos para

entrenamiento interno y ~16% para validación interna, manteniéndose aparte el 20% del conjunto de prueba para la evaluación final. Ver anexo 5.

```
grid_search = GridSearchCV(lr, param_grid, cv=5, scoring="accuracy", n_jobs=-1)
```

5.4.1 Ajuste de parámetros con GridSearchCV en Regresión Logística:

Tabla XXIX. Comparativo de hiperparámetros por defecto y ajustados en Regresión Logística

Modelo	Hiperparámetro	Valor por defecto Scikit-learn	Mejor valor con GridSearchCV
Regresión Logística (LogisticRegression)	C	1.0	1
	solver	'lbfgs'	'lbfgs'
	penalty	'l2'	'l2' (no cambió)
	max_iter	100	100 (no optimizado explícitamente)

5.4.2 Resultados del modelo de Regresión Logística con parámetros ajustados por GridsearchCV

Tabla XXX. Resultados del modelo de Regresión Logística con parámetros ajustados por GridsearchCV

Métrica / Clase	Precision	Recall	F1-Score	Support
Clase 0 (No crisis)	0.94	0.97	0.95	7143
Clase 1 (Crisis)	0.71	0.58	0.64	1069
Exactitud global	—	—	**0.9186**	8212
Promedio macro	0.83	0.77	0.80	8212
Promedio ponderado	0.91	0.91	0.91	8212

Tabla XXXI. Matriz de Confusión Regresión Logística con ajuste de parámetros con GridsearchCV

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6894	Falsos Positivos (FP) 249
Clase 1 (Crisis)	Falsos Negativos (FN) 450	**Verdaderos Positivos (VP)** 619

5.4.3 Análisis de aplicar GridSearchCV al algoritmo de Regresión Logística

Parámetros Óptimos (GridSearchCV)

C = 1: Regula la fuerza de regularización (mayor C = menos regularización).

solver = 'lbfgs': Algoritmo de optimización eficiente para conjuntos de datos de tamaño mediano y clases múltiples.

Comparando los resultados antes y después de usar GridSearchCV, los cambios son mínimos, pero hay una mejora importante en la clase 1 (crisis).

Tabla XXXII: Análisis de resultados de Regresión Logística con ajuste de hiperparámetros

¿Qué mejoró con GridSearchCV?	¿Qué quedó igual?	¿Qué desmejoró?
Exactitud (Accuracy): 0.9186 ligeramente mejor que la original 0.9149	Precisión clase 1 (crisis): 0.71 igual al modelo original Recall clase 1 (crisis): 0.58 igual al modelo original F1-score clase 1: 0.64 igual al modelo original Matriz de Confusión: igual al modelo original	Nada desmejoró

Hay mejoras mínimas tras el ajuste. GridSearchCV confirmó que el modelo con C=1 y solver='lbfgs' ya era cercano al óptimo con los parámetros por defecto.

Clase minoritaria (1 - crisis): Sigue mostrando bajo recall (58%), lo que indica que muchos casos de crisis siguen sin ser detectados (falsos negativos).

Clase mayoritaria (0 - no crisis): Excelente comportamiento, con más del 96% de recall, pero no es el foco del proyecto.

En conclusión, el ajuste de hiperparámetros no mejora sustancialmente los resultados, pues el modelo ya era bastante ajustado con los parámetros por defecto. Es posible que el desempeño esté limitado por la distribución de clases (desbalanceada). Estos resultados más bien discretos eran de esperarse ya que la regresión logística es un modelo lineal y puede no capturar bien patrones complejos. Con los modelos no lineales que se probaron a continuación se esperaba mejorar la detección de crisis.

5.4.4 Ajuste de Parámetros en Random Forest con GridSearchCV:

Tabla XXXIII. Comparativo de hiperparámetros por defecto y ajustados en Random Forest

Modelo	Hiperparámetro	Valor por defecto Scikit-learn	Mejor valor con GridSearchCV
Random Forest (RandomForestClassifier)	n_estimators	100	50
	max_depth	None	None
	min_samples_split	2	5
	min_samples_leaf	1	1
	max_features	‘sqrt’	‘sqrt’ (no cambió)

5.4.5 Resultados del modelo de Random Forest con parámetros ajustados por GridsearchCV

Fitting 5 folds for each of 81 candidates, totalling 405 fits: Se evaluaron 81 combinaciones distintas de hiperparámetros con validación cruzada con 5 particiones (folds) y se ejecutaron 405 entrenamientos de modelos en total.

Arrojando los siguientes resultados:

Tabla XXXIV. Resultados del modelo de Random Forest con parámetros ajustados por GridsearchCV

Métrica / Clase	Precision	Recall	F1-Score	Support
Clase 0 (No crisis)	0.92	0.97	0.95	7143
Clase 1 (Crisis)	0.70	0.46	0.55	1069
Exactitud global	—	—	**0.90**	8212
Promedio macro	0.81	0.71	0.75	8212
Promedio ponderado	0.89	0.90	0.90	8212

Tabla XXXV. Matriz de Confusión de Random Forest con ajuste de parámetros con GridsearchCV

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6934	Falsos Positivos (FP) 209
Clase 1 (Crisis)	Falsos Negativos (FN) 579	**Verdaderos Positivos (VP)** 490

5.4.6 Análisis de aplicar GridSearchCV con el algoritmo Random Forest

Como el enfoque está en la clase 1 (crisis), así que se analizó recall, precisión y la matriz de confusión para evaluar el impacto de GridSearchCV.

El set buscó árboles más profundos, pero ligeramente más robustos al overfitting (por min_samples_split > 2)

Tabla XXXVI: Análisis de resultados de Random Forest con ajuste de hiperparámetros

¿Qué mejoró con GridSearchCV?	¿Qué no cambió?	¿Qué desmejoró?
Recall clase 1 (crisis): 0.46 (antes 0.45) Matriz de confusión: 490 VP (antes 486)	-	Exactitud (Accuracy): 0.90 (bajó ligeramente antes 0.9062) Precisión clase 1 (crisis): 0.70 (bajó, antes 0.72) F1-score clase 1: 0.56 (bajó, antes 0.55)

El ajuste no mejora sustancialmente el rendimiento, especialmente para la clase minoritaria (1 - crisis). Aunque el recall de la clase 1 mejora ligeramente (de 0.45 a 0.46), el modelo sigue teniendo dificultades para identificar correctamente muchos casos de crisis. El número de árboles (50) puede haber limitado la capacidad de generalización en comparación con usar, por ejemplo, 100–200 árboles (valor común por defecto), pero el efecto en consumo computacional sería considerable.

5.4.7 Ajuste de Parámetros en Support Vectors Machines SVM con GridsearchCV

Tabla XXXVII. Comparativo de hiperparámetros por defecto y ajustados en SVM

Modelo	Hiperparámetro	Valor por defecto Scikit-learn	Mejor valor con GridSearchCV
SVM (SVC)	C	1.0	100
	kernel	'rbf'	'rbf'
	gamma	'scale'	'scale'

5.4.8 Resultados del modelo de SVM con parámetros ajustados por GridsearchCV

Fitting 5 folds for each of 32 candidates, totalling 160 fits: Esto es, se aplicó validación cruzada con 5 particiones (folds), generando 32 combinaciones distintas de hiperparámetros posibles, para un total de 160 entrenamientos del modelo, cada uno con diferentes datos de entrenamiento y validación. Presentando los siguientes resultados:

Tabla XXXVIII. Resultados del modelo de SVM con parámetros ajustados por GridsearchCV

Clase / Métrica	Precision	Recall	F1-score	Support
Clase 0	0.95	0.96	0.95	7143
Clase 1	0.71	0.65	0.68	1069
Promedio Macro	0.83	0.80	0.82	8212
Promedio Ponderado	0.92	0.92	0.92	8212
Exactitud general			**0.92**	

Tabla XXXIX. Matriz de Confusión de SVM con ajuste de parámetros con GridsearchCV

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6856	Falsos Positivos (FP) 287
Clase 1 (Crisis)	Falsos Negativos (FN) 376	**Verdaderos Positivos (VP)** 693

5.4.9 Análisis de aplicar GridSearchCV con el algoritmo SVM

Mejores Hiperparámetros: Un valor de C alto y kernel RBF permiten al modelo captar patrones más complejos, aunque con riesgo de sobreajuste.

Tabla XL: Análisis de resultados de SVM con ajuste de hiperparámetros

¿Qué mejoró con GridSearchCV?	¿Qué no cambió?	¿Qué desmejoró?
Exactitud (Accuracy): 0.92 (mejoró ligeramente, antes 0.9163) Recall: 0.65 (mejoró, antes 0.59) F1-score: 0.68 (mejoró, antes 0.65) Matriz de confusión: VP 693 (mejoró, antes 631)		Precisión (crisis): 0.71 (bajó, antes 0.72)

El modelo SVM mejora su recall en la clase minoritaria (1) de 0.59 a 0.65, lo que es valioso para reducir falsos negativos (detectar más crisis). Hay una ligera pérdida en precisión para clase 1 (de 0.72 a 0.71), lo cual es esperable cuando se gana en sensibilidad. Accuracy general también mejora levemente, acercándose al 92%. El balance entre precisión y recall (f1-score = 0.68) lo posiciona mejor que los modelos anteriores hasta ahora en cuanto a clasificación de la clase crítica (1).

En conclusión, el modelo SVM muestra la mejor capacidad para identificar reseñas de crisis hasta ahora, con una mejora sensible en el recall, que es clave si el objetivo es detección temprana de alertas. El SVM con kernel RBF puede ser una excelente opción de base

5.4.10 Ajuste de Parámetros en Redes Neuronales MLPClassifier con GridsearchCV

Tabla XLI. Comparativo de hiperparámetros por defecto y ajustados en Redes Neuronales MLPClassifier

Modelo	Hiperparámetro	Valor por defecto Scikit-learn	Mejor valor con GridSearchCV
MLPClassifier	activation	'relu'	'relu'
	alpha	0.0001	0.001
	hidden_layer_sizes	-100	(50, 50)
	learning_rate	'constant'	'adaptive'
	solver	'adam'	'sgd'

5.4.11 Resultados del modelo de Redes Neuronales MLPClassifier con parámetros ajustados por GridsearchCV

Fitting 5 folds for each of 144 candidates, totalling 720 fits: Se evaluaron 144 combinaciones diferentes de hiperparámetros, usando una validación cruzada con 5 particiones (folds), para un número total de 720 entrenamientos (fits), dando los siguientes resultados:

Tabla XLII. Resultados del modelo de MLPClassifier con parámetros ajustados por GridsearchCV

Clase	Precisión	Recall	F1-Score	Soporte
0 (No Crisis)	0.95	0.97	0.96	7143
1 (Crisis)	0.74	0.63	0.68	1069
Promedio Macro	0.84	0.80	0.82	8212
Promedio Ponderado	0.92	0.92	0.92	8212
Exactitud Total	—	—	**0.92**	8212

Tabla XLIII. Matriz de Confusión de MLPClassifier con ajuste de parámetros con GridsearchCV

Clase 0 (No crisis)	Verdaderos Negativos (VN) 6908	Falsos Positivos (FP) 235
Clase 1 (Crisis)	Falsos Negativos (FN) 393	**Verdaderos Positivos (VP)** 676

5.4.12 Análisis de MLPClassifier con ajuste de hiperparámetros

Mejores hiperparámetros seleccionados: Dos capas ocultas de 50 neuronas cada una. ReLU como función de activación (eficiente en redes profundas). Regularización L2 (alpha=0.001) para evitar sobreajuste. SGD como optimizador, junto con un learning rate adaptable, lo que permite mejorar la convergencia especialmente en datos no

lineales y desbalanceados.

Tabla XLIV: Análisis de resultados de MLPClassifier con ajuste de hiperparámetros

¿Qué mejoró con GridSearchCV?	¿Qué no cambió?	¿Qué desmejoró?
Accuracy (Exactitud): 0.92 (mejoró, antes 0.9009) Precisión: 0.74 (mejora significativa, antes 0.62) F1-score: 0.68 (mejoró, antes 0.62) Matriz de confusión: VP 676 (mejora leve, antes 673)	Recall (crisis) 0.63 (igual)	

Mejora significativa en precisión para la clase 1 (crisis): pasa de 0.62 a 0.74, lo cual indica menos falsos positivos (menos "falsas alarmas"). El recall se mantiene constante en 0.63, lo cual sigue siendo competitivo. El f1-score de 0.68 lo ubica como uno de los mejores hasta ahora, muy similar al SVM ajustado. El aumento en la accuracy general a 0.92 confirma que el ajuste mejora también el rendimiento general del modelo.

En conclusión, el modelo MLPClassifier optimizado compite fuertemente con el SVM ajustado, pues ambos logran f1-scores de 0.68 para la clase 1, aunque MLP ofrece mejor precisión y un modelo no lineal más flexible.

5.6. SELECCIÓN DE ALGORITMO DE APRENDIZAJE AUTOMÁTICO ADECUADO

5.5.1 Análisis Comparativo de Todos los Modelos de Machine Learning

Finalmente se realizó un análisis comparativo de todos los modelos para un escenario real. Como el objetivo principal era detectar crisis reputacionales en hoteles a partir del análisis de sentimiento en reseñas, dando especial énfasis en la clase 1 (crisis). Con base en los resultados obtenidos de todos los modelos, se compararon todos los resultados en función de los criterios más relevantes para el caso.

Tabla XLV: Análisis Comparativo de resultados de todos los modelos ML

Modelo	Accuracy	Precision (Clase 1)	Recall (Clase 1)	F1-Score (Clase 1)	Mejores Hiperparámetros
Logistic Regression	0.9186	0.71	0.58	0.64	C=1, solver='lbfgs'
Random Forest	0.9138	0.73	0.55	0.63	n_estimators=200, max_depth=10
SVM (RBF Kernel)	0.9225	0.71	0.65	0.68	C=100, kernel='rbf', gamma='scale'
MLPClassifier	0.9200	**0.74**	0.63	**0.68**	activation='relu', hidden_layer_sizes=(50,50), solver='sgd', alpha=0.001, learning_rate='adaptive'

Se mencionan los criterios clave a tener en cuenta en un entorno de producción:

- Balance entre precisión y recall en la clase minoritaria (crisis).
- Estabilidad del rendimiento con nuevos datos.
- Costo computacional y facilidad de mantenimiento.
- Interpretabilidad vs. capacidad predictiva.

Interpretación del análisis comparativo:

Tabla XLVI: Comparativo de resultados de todos los modelos ML

Logistic Regression	**Random Forest**
• Rápida, interpretable y fácil de implementar. • Pierde capacidad predictiva en casos más complejos (recall = 0.58). • Útil si se requieren explicaciones claras y decisiones auditables.	• Robusto y buena generalización. • Bajo recall en clase crítica (0.55). • Más estable que SVM y MLP, pero menos preciso en detección de crisis.
SVM (RBF)	**MLPClassifier**
• Excelente balance entre precisión y recall. • Computacionalmente costosa en grandes volúmenes. • Gran opción si se tienen recursos y necesidad de detección precisa.	• Mejor precisión en clase minoritaria (0.74), buen f1-score. • Más difícil de interpretar y requiere ajuste fino. • Muy buena opción si el foco es rendimiento predictivo puro.

5.5.2 Recomendación para un Escenario Real

El mejor modelo, de los evaluados, para la detección de crisis es MLPClassifier porque:

- Presenta un mejor balance general en clase minoritaria.
- Tiene buen f1-score, alta precisión y recall decente.
- Posee buena capacidad de generalizar relaciones no lineales y variables complejas.

Alternativa: SVM. Tiene métricas similares a MLP, pero con resultados ligeramente menores de precisión y recall. Puede ser útil si se optara por un modelo más interpretable y con menos demanda computacional.

Modelos descartados:

Random Forest: Mal desempeño en recall, demasiadas crisis no detectadas.

Logistic Regression: Mal desempeño en recall, aunque es más simple de interpretar.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

El presente proyecto logró consolidar un enfoque metodológico basado en ciencia de datos para analizar reseñas de usuarios en hoteles de Bogotá, con el objetivo de identificar señales tempranas de crisis de reputación. A partir de una base de datos extensa y diversa, se desarrollaron herramientas de procesamiento, análisis y visualización de la información con el propósito de brindar apoyo estratégico a la toma de decisiones en el sector hotelero.

Se implementaron técnicas de procesamiento de lenguaje natural (PLN) que permitieron transformar texto no estructurado en información cuantificable, empleando métodos como la lematización, eliminación de ruido textual, vectorización con Word2Vec y análisis de sentimientos. Estos procesos fueron fundamentales para clasificar automáticamente las reseñas según su polaridad y convertirlas en insumos útiles para modelos predictivos.

En la fase de modelado, se exploraron distintos algoritmos de clasificación como Regresión Logística, Random Forest, Support Vector Machines (SVM) y MLPClassifier, evaluando su desempeño mediante métricas como precisión, recall, F1-score y exactitud. Los resultados indicaron que todos los modelos alcanzaron una buena capacidad de predicción general, aunque con diferencias importantes en la identificación de clases minoritarias, lo cual dió ventajas de desempeño a los modelos de red neuronal MLPClassifier y SVM, por tratarse de clases con bastante desigualdad.

Tras aplicar ajuste de hiperparámetros con GridSearchCV, se evidenció una mejora significativa en modelos como SVM y MLPClassifier, los cuales lograron mayor balance entre precisión y recall en la detección de reseñas clasificadas como potenciales señales de crisis. Esto subraya la importancia de una etapa de optimización en proyectos de machine learning aplicados al análisis reputacional.

A lo largo del proyecto se utilizaron herramientas y tecnologías consolidadas como Python, Jupyter Notebook, y bibliotecas como Pandas, Numpy, Scikit-learn, Matplotlib y Seaborn. Estas permitieron desarrollar un flujo de trabajo reproducible, escalable y documentado, facilitando tanto la exploración inicial de los datos como el desarrollo de modelos avanzados.

Adicionalmente, Se incorporó una dimensión temporal al análisis descriptivo mediante visualizaciones por mes y año, así como mediante la descomposición estacional de series de tiempo. Esta exploración permitió detectar ciertos patrones cíclicos en el volumen de reseñas negativas, especialmente entre los años 2016 y 2019, lo cual podría estar vinculado a dinámicas externas como eventos económicos, políticos o coyunturas sectoriales y así mismo permitió evidenciar temporadas del año en que las reseñas negativas se acentúan para permitir al gremio tomar medidas de anticipación y prevención.

La inclusión de umbrales críticos basados en percentiles y varianzas aportaron una perspectiva más profunda sobre la confiabilidad y sensibilidad de los modelos, ayudando a definir alertas tempranas y a priorizar acciones según la severidad y frecuencia de las señales detectadas. El análisis desagregado por nombre de hotel (mediante la variable "Hotel_Name") brindó una vista más granular del comportamiento individual de cada establecimiento, permitiendo detectar casos particulares de deterioro reputacional o recuperación progresiva. Esta información puede ser altamente útil para planes de mejora y estrategias de fidelización de clientes.

Es importante destacar que, si bien el enfoque predictivo mostró buenos resultados, las crisis de reputación son fenómenos complejos influenciados por múltiples factores cualitativos. Por tanto, el modelo propuesto debe considerarse como un sistema de apoyo a la toma de decisiones y no como una herramienta determinista o definitiva.

Finalmente, este proyecto sienta las bases para futuras líneas de investigación y mejora, tales como el análisis semántico profundo, el entrenamiento de modelos con datos multilingües, la incorporación de variables externas (como precios, ocupación o eventos locales) y el desarrollo de sistemas en tiempo real para la vigilancia reputacional continua en el sector hotelero.

6.2. TRABAJOS FUTUROS

A partir de los hallazgos y metodologías desarrolladas en este proyecto, se abren múltiples posibilidades de profundización, expansión y aplicación práctica que pueden fortalecer la capacidad predictiva y estratégica del análisis de reputación en el sector hotelero. Así, en el futuro se puede pensar en expandir el modelo con características textuales, embeddings o datos temporales, MLP (Multilayer Perceptron) o SVM los cuáles serían más escalables.

Una primera línea de trabajo futuro contempla la detección de insatisfacciones ocultas en reseñas con calificaciones positivas. Este fenómeno, donde el texto contiene críticas encubiertas a pesar de una puntuación aparentemente favorable, puede abordarse mediante modelos de análisis de sentimiento por aspectos (aspect-based sentiment analysis), permitiendo identificar dimensiones específicas del servicio que generan descontento (por ejemplo, limpieza o atención al cliente), incluso cuando la opinión general parece positiva.

Dado que el MLP puede capturar relaciones complejas, también es una buena opción al expandir el modelo a una entrada con múltiples variables (por ejemplo, incluyendo características textuales, fechas o metadata). Incorporando la dimensión temporal al análisis, permitiría observar la evolución de la reputación en el tiempo e identificar patrones de deterioro progresivo que anteceden a una crisis. Este enfoque facilitaría la creación de sistemas de alerta temprana, capaces de notificar a los gestores cuando las señales negativas comienzan a consolidarse, brindándoles tiempo para reaccionar oportunamente.

Otra línea valiosa consiste en el análisis comparativo entre hoteles de una misma categoría o zona. Esto no solo permitiría evaluar el desempeño relativo de cada establecimiento, sino también detectar ventajas competitivas o debilidades estructurales frente a la competencia.

Asimismo, el modelo podría enriquecerse integrando características del entorno geográfico y social, como temporadas turísticas, eventos masivos o problemáticas locales, lo que daría lugar a un modelo reputacional contextualizado, más robusto y sensible al entorno dinámico del sector.

Finalmente, una aplicación práctica relevante sería el desarrollo de tableros de visualización interactivos, orientados a la gestión hotelera, que presenten los resultados del análisis de sentimiento, alertas tempranas, nubes de palabras clave negativas, y comparaciones con la competencia. Esta herramienta permitiría transformar los resultados del modelo en inteligencia accionable para la toma de decisiones estratégicas.

En conjunto, estas líneas apuntan a consolidar un sistema integral de gestión de la reputación hotelera, basado en ciencia de datos y centrado en la voz del cliente como fuente primaria de información.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] W. T. Coombs, *Ongoing crisis communication: Planning, Managing, and Responding*. SAGE Publications, Incorporated, 2014.

[2] Bancolombia. (28 de junio de 2024). Reportes del sector hotelería y turismo en Colombia en 2024. [En línea] Disponible: https://www.bancolombia.com/empresas/capital-inteligente/especiales/informes-sectoriales/sector-hoteleria-turismo

[3] C. V. Costa, *Comunicación de crisis, redes sociales y reputación corporativa.* ESIC, 2019.

[4] Superintendencia Financiera de Colombia, Circular interna sobre "Marco De Gestión De Riesgos De Los Conglomerados Financieros (Mgr)", Bogotá, Colombia. Julio 7 de 2020. Circular Externa 025.

[5] J. Ramos, *Gestión de la reputación online. Claves y estrategias. Cómo monitorizar y gestionar su reputación online en la Web*. XinXii, 2012.

[6] R. Hernández y X. Li. *Sentiment analysis of texts in spanish based on semantic approaches with linguistic rules*. Mexico DF: CINVESTAV-IPN-Department of Computer Science, 2014.

[7] D. Das, A. Kumar, S. Sarkar, A. Basu. *Sentiment analysis and computational intelligence*, Academic Press, 2023.

[8][pag 18] Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed. draft). Stanford University. [Online]. Disponible en: https://web.stanford.edu/~jurafsky/slp3/

[9][pag 18] D. Kuhlman. *A Python Book: Beginning Python, Advanced Python, and Python Exercises*. 2011

[10][pag 18] Anaconda Inc.(2025). About us. [En línea]. Disponible: https://www.anaconda.com/about-us

[11] [pag 18] Numfocus Org. (2025). About Pandas. [En línea]. Disponible: https://pandas.pydata.org/about/index.html

[12][pag 18] C. R. Harris et al., *Array programming with NumPy*, Nature, vol. 585, n.º 7825, pp. 357–362, septiembre de 2020. Accedido el 18 de mayo de 2025. [En línea]. Disponible: https://doi.org/10.1038/s41586-020-2649-2

[13][pag 18] J. D. Hunter. (2012). Matplotlib History. [En línea]. Disponible: https://matplotlib.org/stable/project/history.html

[14][pag 18] M. Waskom. (2024). seaborn: statistical data visualization. [En línea] Disponible: https://seaborn.pydata.org/

[15][pag 18] E. Klein, S. Bird y E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly , Inc., 2009.

[16][pag 18] D. Cournapeau. (2007). *Google Summer of Code project*. [En línea]. Disponible: https://scikit-learn.org/stable/about.html

[17][pag 19] D. S. Jurafsky, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall, 2000.

[18] [pag 19] Tin Kam Ho, *The random subspace method for constructing decision forests*, IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, n.º 8, pp. 832–844, 1998. [En línea]. Disponible: https://doi.org/10.1109/34.709601

[19][pag 19] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. New York, NY: Springer N. Y., 2006.

[20][pag 19] Geeks For Geeks Educational Portal. (2024). *Performing Feature Selection with gridsearchcv in Sklearn.*[En línea]. Disponible: https://www.geeksforgeeks.org/performing-feature-selection-with-gridsearchcv-in-sklearn/

[21] Fombrun, Charles. *Reputation: Realizing Value from the Corporate Image*. Harvard Business School Press, 1996.

[22] S. Brown, Et. Al. *The Company And The Product: Corporate Associations and Consumer Product Responses*. Journal of Marketing, 1997.

[23] Á. R. Gil, I. C. J. Barandalla, and C. M. Idoeta. *Reputación corporativa online en la hotelería: el caso TripAdvisor*. ESIC MARKET Economic and Business Journal, 2017.

[24] O. Cervantes Flores et al. *Análisis de sentimientos a comentarios de hoteles: Un caso práctico*. Escuela de Negocios Universidad Popular Autónoma del Estado de Puebla. Puebla, México, 2021.

[25] F. Lovera et. Al. *Análisis de sentimientos en Twitter: Un estudio comparativo*. Revista Científica de Sistemas e Informática Universidad Nacional de San Martín, Perú, 2023.

[26] Change Management Insight. (2016, marzo 7). Uber Crisis Management Approach and Lessons Learned [En línea], Disponible: https://changemanagementinsight.com/uber-crisis-management/

[27]CNN en español. (2017, abril 11). Acciones de United Airlines, en caída libre tras escándalo de expulsión de pasajero [En línea], Disponible: https://cnnespanol.cnn.com/2017/04/11/acciones-de-united-airlines-en-caida-libre-tras-escandalo-de-expulsion-de-pasajero/

[28]Cadena SER. (2018, julio 27). El problema reputacional de Facebook por la crisis de Cambridge Analytica [En línea]. Disponible: https://cadenaser.com/ser/2018/07/27/economia/1532671825_061954.html

[29] CNN en español. (2024, Enero 10). Crisis del 737 Max 9 de Boeing [En línea]. Disponible: https://cnnespanol.cnn.com/2024/01/10/crisis-boeing-737-max-9-aviones-problemas-error-trax

[30] IO Investigación social y comunicación. (2021, Junio 2). Crisis de reputación online 'Caso Nestlé' [En línea]. Disponible: https://www.io-siscom.com/crisis-de-reputacion-online/

[31][pag 34] J. H. Martin y D. Jurafsky, *Speech and Language Processing (2nd Edition)*, 2a ed. Prentice Hall, 2006.

[32] Scikit-learn Docs LogisticReseression [En línea].
Disponible: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[33] Scikit-learn Docs Random Forest Classifier [En línea].
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[34] Scikit-learn Docs Support Vector Machines SVM [En línea].
https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[35] Scikit-learn Docs MLPClassifier [En línea].
https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

ANEXOS

ANEXO 1. Script Python. 1-preprocessing_data_hoteles.ipynb

ANEXO 2. Script Python. 2-visualization_time_series.ipynb

ANEXO 3. Script Python. 3-lemmatization.ipynb

ANEXO 4. Script Python. 4-vectorizacion.ipynb

ANEXO 5. Script Python. 5-Modelo-Machine-Learning.ipynb

ANEXO 6. Link del proyecto en Github:
https://github.com/juanmanuelsilva/SentimentAnalysis-NaturalLanguageProcessing-NLP.git

ANEXO 6

Los scripts correspondientes al presente proyecto aplicado se encuentran disponibles en el siguiente repositorio del servicio Github.com:

https://github.com/juanmanuelsilva/SentimentAnalysis-NaturalLanguageProcessing-NLP