



Facultad de Ciencias Económicas y Empresariales

Análisis de percepciones de clientes con modelos de lenguaje de gran escala (LLMs) cuantizados

Autor: Irene Yongling Barrena Lorente

Director: Jenny Alexandra Cifuentes Quintero

MADRID | Mes 2025-2026

Resumen

El análisis de sentimientos se ha consolidado como una herramienta relevante para comprender las percepciones de los consumidores en entornos de comercio electrónico y apoyar la toma de decisiones empresariales. En este ámbito, los *Large Language Models* (LLMs) ofrecen una mayor capacidad contextual que los enfoques tradicionales, aunque su implementación práctica sigue condicionada por elevados requerimientos computacionales, la dependencia de APIs propietarias y, en muchos casos, la necesidad de datos etiquetados para el ajuste fino. Frente a estas limitaciones, la cuantización reduce el consumo de memoria y facilita la ejecución de LLMs en hardware convencional, mientras que el *prompting* permite adaptar su comportamiento a tareas específicas sin modificar sus parámetros internos.

En este contexto, el presente Trabajo de Fin de Grado evalúa el uso de LLMs cuantizados a 4 bits para la clasificación de sentimientos en reseñas de productos en español. La metodología se estructuró en tres etapas. Primero, se seleccionó un subconjunto de 200 reseñas del corpus *Amazon Reviews*, clasificadas en positivo, negativo y neutro. Segundo, se caracterizaron tres modelos cuantizados, *LLaMA 3.1 8B Instruct*, *Qwen 2.5 7B Instruct* y *Mistral 7B Instruct*, atendiendo a consumo de memoria, latencia y perplejidad. Posteriormente, se aplicaron seis estrategias de *prompting*, generando 18 combinaciones modelo-técnica evaluadas bajo las mismas condiciones experimentales. Finalmente, el desempeño se analizó mediante *F1 macro-promediada*, *accuracy*, matrices de confusión, métricas por clase e intervalos de confianza al 95 % mediante *bootstrap*. Los resultados muestran que los LLMs cuantizados a 4 bits pueden ofrecer un rendimiento competitivo sin recurrir a ajuste fino, APIs propietarias ni infraestructura especializada. Las mejores combinaciones alcanzaron valores de *F1 macro-promediada* superiores a 0,70 y una *accuracy* entre 0,80 y 0,84. *Mistral 7B* presentó el mejor perfil computacional, mientras que *Qwen 2.5* obtuvo los mejores valores clasificatorios. Considerando conjuntamente rendimiento global, simplicidad del *prompt*, menor número de inversiones de polaridad y recuperación de la clase neutra, *Qwen 2.5 zero-shot* se identificó como la alternativa más equilibrada. Asimismo, los resultados muestran que la clase neutra constituye el principal desafío clasificatorio, frente a una identificación más fiable de opiniones positivas y negativas.

Palabras Clave: análisis de sentimientos, LLM, cuantización, *prompting*, reseñas de productos

Abstract

Sentiment analysis has established itself as a relevant tool for understanding consumer perceptions in e-commerce environments and supporting business decision-making. In this domain, Large Language Models (LLMs) offer greater contextual capabilities than traditional approaches, although their practical implementation remains constrained by high computational requirements, dependence on proprietary APIs, and, in many cases, the need for labeled data for fine-tuning. To address these limitations, quantization reduces memory consumption and enables LLM execution on conventional hardware, while prompting allows adapting model behavior to specific tasks without modifying internal parameters.

In this context, this Bachelor’s Thesis evaluates the use of 4-bit quantized LLMs for sentiment classification of Spanish product reviews. The methodology was structured in three stages. First, a subset of 200 reviews from the *Amazon Reviews* corpus was selected, classified into positive, negative, and neutral. Second, three quantized models were characterized, *LLaMA 3.1 8B Instruct*, *Qwen 2.5 7B Instruct*, and *Mistral 7B Instruct*, in terms of memory consumption, latency, and perplexity. Subsequently, six prompting strategies were applied, generating 18 model-technique combinations evaluated under identical experimental conditions. Finally, performance was analyzed using macro-averaged F1-score, accuracy, confusion matrices, per-class metrics, and 95 % bootstrap confidence intervals. Results show that 4-bit quantized LLMs can deliver competitive performance without fine-tuning, proprietary APIs, or specialized infrastructure. The best combinations achieved macro-averaged F1-scores above 0.70 and accuracy between 0.80 and 0.84. *Mistral 7B* presented the best computational profile, while *Qwen 2.5* obtained the highest classification scores. Considering overall performance, prompt simplicity, fewer polarity inversions, and neutral class recall, *Qwen 2.5 zero-shot* was identified as the most balanced alternative. Results also show that the neutral class constitutes the main classification challenge, compared to more reliable identification of positive and negative opinions.

Keywords: sentiment analysis, LLM, quantization, prompting, product reviews.

Índice general

Resumen	III
Abstract	IV
1. Introducción	1
1.1. Motivación	1
1.2. Objetivos	4
1.2.1. Objetivo General	4
1.2.2. Objetivos Específicos	4
1.3. Estructura del Documento	4
2. Revisión del Estado del Arte en Procesamiento de Lenguaje Natural aplicados al Análisis de Sentimientos	6
3. Diseño metodológico	14
3.1. Selección del conjunto de datos	15
3.2. Modelo de lenguaje cuantizado y estrategias de <i>prompting</i>	16
3.2.1. Cuantización de <i>LLMs</i>	17
3.2.2. Criterios de selección del <i>LLM</i> cuantizado	21
3.2.3. Estrategias de <i>prompting</i> para la clasificación de sentimientos	23
3.3. Evaluación del desempeño de los modelos	28
3.3.1. Matriz de confusión	29
3.3.2. Métricas de clasificación	30
3.3.3. Intervalos de confianza y significación estadística	32
4. Resultados	33
4.1. Caracterización de los <i>LLMs</i> cuantizados	33
4.2. Clasificación de sentimientos mediante técnicas de <i>prompting</i>	37
4.2.1. Resultados globales	43
5. Conclusiones	49

Índice de figuras

1.1. Encuesta sobre personalización de clientes de BCG de 2023 (n = 5000). . .	2
3.1. Etapas de la metodología seguida en este estudio	15
4.1. Comparativa de eficiencia de los tres modelos cuantizados. Mistral 7B destaca en las tres dimensiones, con el menor consumo de VRAM, la menor latencia por reseña y la perplejidad más baja.	37
4.2. Tiempo de ejecución total acumulado por modelo. Mistral 7B completa las 6 técnicas en 2.129 s (35 min), frente a los 5.202 s de Qwen 2.5 (87 min) y los 7.022 s de LLaMA 3.1 (117 min). Las técnicas <i>CoT</i> e <i>iterative</i> concentran la mayor proporción del tiempo.	38
4.3. <i>F1 macro-promediada</i> por modelo y técnica.	43

Índice de tablas

2.1. Resumen de los estudios sobre el análisis de sentimientos	12
3.1. Resumen de estudios empíricos sobre cuantización en los principales provee- dores de <i>large language models</i>	20
3.2. Estructura de la matriz de confusión para clasificación ternaria de sentimientos.	29
4.1. Modelos candidatos seleccionados para la evaluación comparativa.	35
4.2. Tamaño en disco de los modelos candidatos en su versión original (<i>float16</i>).	35
4.3. <i>Prompts</i> reales empleados (parte 1 de 2), ilustrados con la misma reseña del corpus.	40
4.4. <i>Prompts</i> reales empleados (parte 2 de 2). Fase 2 de <i>iterative</i> con clasificación previa “negativo”.	41
4.5. Respuestas reales de cada modelo para la reseña “mala calidad. Lo compra- mos hace pocos meses [...]”. En rojo los comportamientos problemáticos. Etiqueta real: <i>negativo</i>	42
4.6. <i>Accuracy</i> por modelo y técnica, sobre 200 reseñas.	44
4.7. <i>F1-score</i> por clase para cada combinación.	45
4.8. Matriz de confusión: <i>Qwen 2.5</i> con <i>few-shot</i> (F1-macro = 0,7216).	46
4.9. Matriz de confusión: <i>Qwen 2.5</i> con <i>zero-shot</i> (F1-macro = 0,7188).	47
4.10. Matriz de confusión: <i>Mistral 7B</i> con <i>role-based</i> (F1-macro = 0,7068).	47
4.11. <i>Precision</i> y <i>recall</i> por clase para las tres mejores combinaciones. En negrita se marca el mejor valor de cada métrica por clase.	47

Acrónimos

<i>API</i>	Application Programming Interface
<i>AWQ</i>	Activation-Aware Weight Quantization
<i>BCG</i>	Boston Consulting Group
<i>BERT</i>	Bidirectional Encoder Representations from Transformers
<i>CNN</i>	Convolutional Neural Network
<i>CoT</i>	Chain-of-Thought
<i>GPT</i>	Generative Pre-trained Transformer
<i>GPTQ</i>	Generative Pre-trained Transformer Quantization
<i>GPU</i>	Graphics Processing Unit
<i>GQA</i>	Grouped-Query Attention
<i>GRU</i>	Gated Recurrent Unit
<i>INT4</i>	Integer 4-bit
<i>INT8</i>	Integer 8-bit
<i>K-BERT</i>	Knowledge-enhanced BERT
<i>KL</i>	Kullback-Leibler
<i>KV</i>	Key-Value
<i>LCF-BERT</i>	Local Context Focus BERT
<i>LLM</i>	Large Language Model
<i>LoRA</i>	Low-Rank Adaptation
<i>LSTM</i>	Long Short-Term Memory
<i>NF4</i>	NormalFloat 4-bit
<i>NLP</i>	Natural Language Processing
<i>PTQ</i>	Post-Training Quantization
<i>QAT</i>	Quantization-Aware Training
<i>QLoRA</i>	Quantized Low-Rank Adaptation
<i>RAM</i>	Random Access Memory
<i>RNN</i>	Recurrent Neural Network
<i>RoBERTa</i>	Robustly Optimized BERT Pretraining Approach

<i>SO-CAL</i>	Semantic Orientation Calculator
<i>SVM</i>	Support Vector Machine
<i>SWA</i>	Sliding Window Attention
<i>T5</i>	Text-to-Text Transfer Transformer
<i>VADER</i>	Valence Aware Dictionary and sEntiment Reasoner
<i>VRAM</i>	Video Random Access Memory

Capítulo 1

Introducción

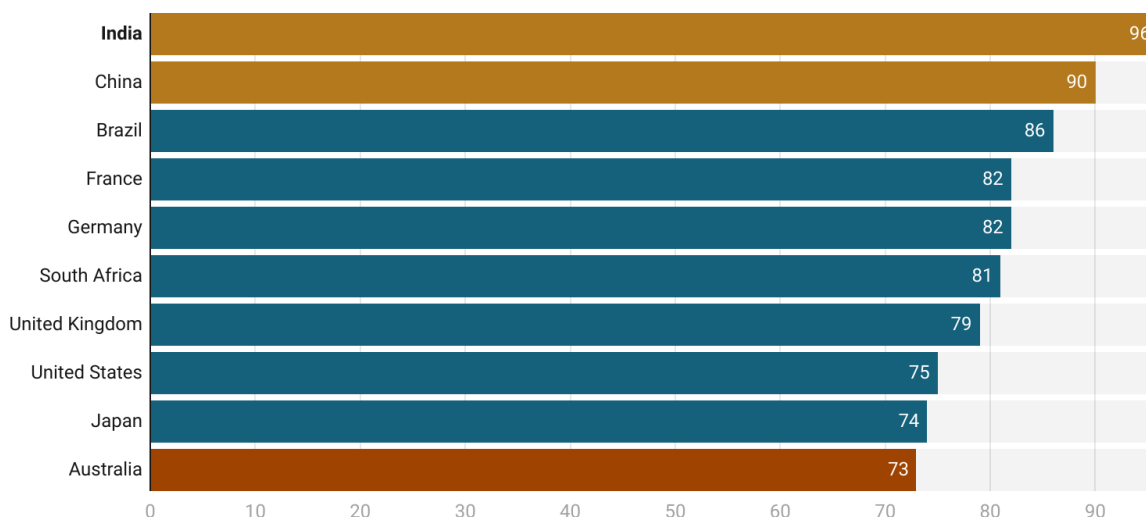
1.1. Motivación

En la era actual caracterizada por la disponibilidad de datos masivos, las empresas se enfrentan a un entorno altamente competitivo, dinámico y en constante transformación, donde alcanzar y mantener una posición de éxito constituye un desafío cada vez más complejo. A ello se suma que la mayoría de productos y servicios han alcanzado un grado significativo de madurez y los mercados presentan altos niveles de saturación, lo que dificulta la generación de innovaciones disruptivas (Christensen, Raynor, y McDonald, 2015). En este escenario, según (Hernández, 2020), uno de los factores decisivos para alcanzar el éxito empresarial reside en la construcción de una posición sólida en el mercado, sustentada principalmente en la fidelización de un número amplio y estable de clientes. En consecuencia, la verdadera diferenciación competitiva no depende exclusivamente de la calidad intrínseca de los productos o de la política de precios, sino de la capacidad de ofrecer experiencias personalizadas que fortalezcan el vínculo entre empresa y consumidor. La personalización, entendida como un mecanismo para fidelizar clientes y reducir el riesgo de fuga hacia la competencia, se sustenta en la creación de relaciones emocionales duraderas, en las que los clientes se sientan escuchados, valorados y comprendidos como individuos. De este modo, se incrementa no solo la probabilidad de recompra, sino también la propensión a recomendar la marca a otros consumidores, generando un círculo virtuoso de confianza y lealtad (Arora et al., 2021).

Bajo estas consideraciones, la personalización de la experiencia del cliente basada en el análisis sistemático de datos no puede concebirse como una simple tendencia coyuntural, sino como una práctica consolidada y con amplio respaldo en los mercados internacionales. Como se observa en la Figura 1.1, los resultados de la encuesta realizada por *Boston Consulting Group* (BCG) en 2023 resaltan que la personalización es altamente valorada a nivel global, con niveles de aceptación que no descienden del 73 % en ninguno de los nueve países analizados. Destacan especialmente los mercados asiáticos, donde India alcanza un 96 % y China un 90 %, reflejando la importancia atribuida a este enfoque por parte de los consumidores.

Estos resultados confirman que el uso de datos de clientes para personalizar su experiencia constituye una práctica relevante y generadora de valor tanto para las empresas como para los consumidores (Abraham, Geng, Kogler, y Taylor, 2024).

El nivel de comodidad de los clientes con las empresas que utilizan información pública sobre ellos para crear una experiencia personalizada, por país.(%)



Source: BCG • Created with Datawrapper

Figura 1.1: Encuesta sobre personalización de clientes de BCG de 2023 (n = 5000).

En este contexto, el análisis de sentimientos desempeña un papel de gran relevancia, al constituirse como la herramienta que permite comprender en profundidad las emociones, opiniones y comportamientos expresados por los consumidores en múltiples formatos, desde reseñas de productos hasta comentarios en redes sociales. Gracias a su capacidad para transformar datos textuales no estructurados en información procesable, las empresas pueden captar de manera sistemática las percepciones de sus clientes, identificar los atributos mejor valorados y detectar con antelación los factores que generan experiencias negativas o insatisfacción. De esta manera, se crea la base necesaria para diseñar experiencias más personalizadas, consolidar la fidelización y reforzar la relación empresa-cliente. El potencial de esta herramienta va más allá de la simple detección de opiniones. A través del análisis de sentimientos, las organizaciones pueden innovar con mayor precisión, creando productos y servicios ajustados a las necesidades emocionales y funcionales de los consumidores, en lo que se conoce como desarrollo de productos emocionales (Yan y Li, 2022). Asimismo, este tipo de análisis contribuye a identificar oportunidades de mejora, anticipar riesgos reputacionales y apoyar la formulación de decisiones estratégicas fundamentadas en evidencia (Pathak et al., 2023). En consecuencia, el análisis de sentimientos no solo aporta información descriptiva sobre la percepción del cliente, sino que se convierte en un mecanismo para generar ventajas competitivas sostenibles.

Su utilidad resulta particularmente relevante en el ámbito del *marketing*, donde permite

gestionar de manera más eficiente los recursos y esfuerzos de captación y fidelización. El conocimiento profundo del comportamiento del cliente posibilita diseñar campañas de publicidad alineadas con los intereses y expectativas del público objetivo, lo que incrementa la eficacia de las acciones implementadas. Además, la focalización de los recursos en las áreas de mayor impacto evita el desperdicio en iniciativas de escaso rendimiento. Otro aspecto relevante es que el análisis de sentimientos constituye una alternativa ventajosa frente a los métodos tradicionales, al ofrecer mayor efectividad, menores costes y la posibilidad de obtener resultados en tiempo real. Estos beneficios reflejan la necesidad de integrar este tipo de herramientas en las organizaciones modernas, tanto para optimizar los flujos de trabajo como para respaldar la toma de decisiones con datos objetivos, reduciendo así el riesgo de obsolescencia en el entorno competitivo actual (Aguilar-Moreno, Palos-Sanchez, y Pozo-Barajas, 2024).

En consecuencia, el análisis de sentimientos puede considerarse una herramienta indispensable dentro de la gestión empresarial contemporánea, al facilitar tanto la personalización de la experiencia del cliente como la generación de ventajas competitivas sostenibles. No obstante, su implementación práctica presenta una notable diversidad metodológica que abarca desde enfoques basados en diccionarios hasta modelos de aprendizaje profundo y, más recientemente, la aplicación de modelos de lenguaje de gran escala (*Large Language Models*, LLMs). Esta evolución evidencia el progreso de la disciplina y, al mismo tiempo, destaca los desafíos asociados a la elección de la estrategia más adecuada en función del contexto, la naturaleza de los datos y los objetivos específicos de la organización. En este escenario, los LLMs representan un hito en el tratamiento de datos textuales, ya que ofrecen una capacidad sin precedentes para comprender, generar y clasificar lenguaje natural de manera flexible y adaptativa (Ghatora, Hosseini, Pervez, Iqbal, y Shaukat, 2024). Sin embargo, su uso plantea también retos vinculados al elevado coste computacional y a las dificultades de implementación en entornos empresariales con recursos limitados. La cuantización de estos modelos surge como una alternativa prometedora, al permitir reducir de manera significativa el consumo de memoria y el tiempo de inferencia, manteniendo un desempeño competitivo en tareas de procesamiento del lenguaje natural.

Partiendo de esta necesidad de encontrar un equilibrio entre precisión analítica y eficiencia operativa, el presente Trabajo de Fin de Grado se centra en el diseño y la evaluación de una metodología de análisis de sentimientos basada en LLMs cuantizados y técnicas de *prompting*. Esta propuesta pretende no solo contrastar el rendimiento de este enfoque frente a metodologías tradicionales, sino también explorar su potencial como herramienta práctica para que las empresas puedan captar percepciones de sus clientes de manera estructurada, generar información de valor y respaldar la toma de decisiones estratégicas en un entorno cada vez más dinámico y competitivo.

1.2. Objetivos

1.2.1. Objetivo General

El presente Trabajo de Fin de Grado en Business Analytics tiene como objetivo principal proponer una metodología de análisis de percepciones de clientes en reseñas de productos mediante el uso de LLMs cuantizados, evaluando su capacidad para clasificar sentimientos en reseñas de comercio electrónico en español mediante técnicas de *prompting*, y analizando tanto el rendimiento clasificatorio como la eficiencia computacional en un entorno con recursos limitados.

1.2.2. Objetivos Específicos

- Contextualizar la importancia del análisis de sentimientos como herramienta estratégica para las empresas en la era digital, destacando su papel en la mejora de la experiencia del cliente, la identificación de ventajas competitivas y la toma de decisiones basada en datos.
- Identificar, clasificar y analizar, a partir de la literatura académica y profesional, las principales técnicas de análisis de sentimientos empleadas en la actualidad, con el fin de establecer un marco comparativo sólido que sirva de base para la propuesta metodológica.
- Diseñar una metodología de análisis de sentimientos apoyada en modelos LLMs cuantizados y técnicas de *prompting*, y evaluar su desempeño mediante métricas estándar de clasificación multiclase, considerando tanto el rendimiento clasificatorio como la eficiencia computacional, y contextualizando los resultados obtenidos con los reportados en la literatura para enfoques basados en diccionarios, modelos supervisados y LLMs propietarios.

1.3. Estructura del Documento

El presente Trabajo de Fin de Grado se estructura en cinco capítulos. El Capítulo 2 está dedicado a la revisión de la literatura, cuyo propósito es examinar de manera sistemática las distintas técnicas de análisis de sentimientos desarrolladas en los últimos años. En este apartado se abordan tanto los enfoques tradicionales, como los métodos basados en diccionarios y los modelos supervisados clásicos, como las aproximaciones más recientes que incorporan LLMs para la detección y clasificación de emociones y opiniones. El Capítulo 3 describe la metodología seguida en la investigación, detallando las etapas de selección del conjunto de datos, selección y cuantización de los modelos de lenguaje y diseño de las técnicas de

prompting, así como los criterios establecidos para la evaluación del desempeño del sistema.

Por su parte, el Capítulo 4 presenta los resultados obtenidos tras la aplicación de la metodología propuesta. Se incluyen los indicadores de rendimiento alcanzados por los modelos de lenguaje cuantizados evaluados con las distintas técnicas de *prompting*, analizando tanto las métricas globales de clasificación como el comportamiento diferenciado por clase de sentimiento y los patrones de error más frecuentes. Finalmente, el Capítulo 5 recoge las conclusiones del estudio, destacando las principales aportaciones, las implicaciones prácticas para el ámbito empresarial y académico, así como las limitaciones detectadas en el desarrollo de la investigación. Del mismo modo, se plantean posibles líneas de trabajo futuro que permitan profundizar y ampliar los resultados alcanzados en este trabajo.

Capítulo 2

Revisión del Estado del Arte en Procesamiento de Lenguaje Natural aplicados al Análisis de Sentimientos

La creciente necesidad de diferenciación y posicionamiento de marca en los mercados actuales ha impulsado a las empresas a desarrollar estrategias orientadas a fortalecer la fidelización de sus clientes (Hernández, 2020). En un entorno caracterizado por una competencia intensa y por una oferta de productos y servicios cada vez más homogénea, comprender las percepciones de los consumidores se ha convertido en un elemento fundamental para la toma de decisiones empresariales. En respuesta a esta necesidad, han surgido herramientas analíticas capaces de identificar y examinar las emociones, opiniones y actitudes que los clientes expresan en distintos formatos digitales, como reseñas, valoraciones y comentarios en plataformas en línea. Entre estas herramientas, el análisis de sentimientos destaca por su capacidad para detectar atributos bien valorados de una marca, anticipar señales de insatisfacción y contribuir al diseño de experiencias más personalizadas. De este modo, esta técnica favorece la diferenciación competitiva y la creación de ventajas sostenibles. Desde el punto de vista metodológico, el análisis de sentimientos se fundamenta en técnicas de *Natural Language Processing* (NLP), orientadas a identificar, clasificar y cuantificar la polaridad emocional expresada en un texto, generalmente mediante categorías positivas, negativas o neutras. Así, permite transformar grandes volúmenes de datos textuales no estructurados en información organizada, interpretable y útil para el diagnóstico empresarial y la toma de decisiones.

La evolución tecnológica ha dado lugar a una amplia diversidad de enfoques para implementar el análisis de sentimientos. Entre los métodos más tradicionales se encuentran los enfoques basados en diccionarios, en los que la polaridad de un texto se determina comparando sus palabras con léxicos previamente clasificados según su orientación emocional. En esta línea, (Cho, Kim, Lee, y Lee, 2014) proponen un método basado en la integración y adaptación de diccionarios léxicos para el análisis de reseñas de Amazon en tres categorías

de producto: smartphones, películas y libros. Los autores combinan distintos recursos de sentimiento, entre ellos *AFINN*, *General Inquirer*, *Opinion Lexicon*, *SenticNet*, *SentiWordNet*, *SO-CAL*, *Subjectivity Lexicon* y *WordNet-Affect*, con el fin de ampliar la cobertura léxica y homogeneizar formatos originalmente heterogéneos. Posteriormente, ajustan el diccionario resultante con reseñas etiquetadas, eliminando términos con baja capacidad discriminativa y modificando la polaridad de aquellas palabras cuyo uso en el dominio analizado difiere de la polaridad asignada inicialmente. Los resultados muestran una precisión balanceada aproximada del 82,6 % en smartphones, 80 % en películas y 81,8 % en libros. Este tipo de aproximación presenta ventajas relevantes, como su elevada interpretabilidad, su facilidad de implementación, la ausencia de entrenamiento previo y su bajo coste computacional. Sin embargo, su dependencia de léxicos predefinidos también limita su capacidad para capturar matices contextuales, variaciones propias de cada dominio y fenómenos lingüísticos complejos.

Las restricciones asociadas a los enfoques léxicos han sido analizadas de forma comparativa en estudios que los contrastan con modelos de aprendizaje automático. En esta línea, (Nguyen, Veluchamy, Diop, y Iqbal, 2018) evalúan distintas técnicas de análisis de sentimientos aplicadas a reseñas de productos de Amazon, comparando modelos basados en diccionarios, como *VADER*, *Pattern* y *SentiWordNet*, con algoritmos supervisados como *Logistic Regression*, *Support Vector Machine* y *Gradient Boosting*. Sus resultados muestran que los modelos supervisados alcanzan un rendimiento superior, con precisiones cercanas al 87–90 %, frente a un desempeño más limitado y variable en los enfoques léxicos. Esta comparación evidencia que, aunque los diccionarios ofrecen ventajas en términos de interpretabilidad y bajo coste computacional, presentan dificultades para capturar connotaciones dependientes del contexto, así como fenómenos lingüísticos complejos como la negación, la ambigüedad léxica o el sarcasmo. En consecuencia, los métodos basados en aprendizaje automático permiten superar parcialmente estas dificultades al aprender patrones directamente a partir de los datos, aunque su desempeño depende de la disponibilidad y calidad de conjuntos etiquetados.

La necesidad de representar con mayor precisión la complejidad semántica del lenguaje impulsó el uso de modelos de aprendizaje profundo en el análisis de sentimientos. A diferencia de los enfoques tradicionales, estas técnicas permiten aprender representaciones más ricas de los textos y detectar relaciones no lineales entre palabras, estructuras sintácticas y polaridades emocionales. En este sentido, (Ain et al., 2017) revisan la aplicación de modelos de *deep learning* al análisis de opiniones en textos digitales, como redes sociales, reseñas y blogs, destacando su capacidad para mejorar el rendimiento frente a métodos más simples. Estos modelos resultan especialmente relevantes porque reducen la dependencia de reglas manuales y léxicos predefinidos, y pueden adaptarse a distintos dominios mediante el aprendizaje de patrones contextuales.

A pesar de las mejoras introducidas por el aprendizaje profundo, su aplicación prácti-

ca también plantea desafíos relevantes. (Panthati, Bhaskar, Ranga, y Challa, 2018) analizan estas restricciones en el contexto de reseñas de productos y proponen un enfoque que combina *Word2Vec*, para la representación semántica del texto, con una red recurrente *LSTM*, orientada a la clasificación de la polaridad. Este tipo de arquitectura permite capturar relaciones secuenciales y dependencias contextuales que los modelos más simples no representan adecuadamente. Sin embargo, su desempeño depende en gran medida del tamaño y la calidad del conjunto de datos, así como de una etapa de preprocesamiento adecuada. Además, estos modelos pueden presentar riesgo de sobreajuste, elevada complejidad computacional, mayores costes de implementación y menor interpretabilidad frente a los enfoques léxicos tradicionales. Por tanto, aunque el aprendizaje profundo amplía la capacidad de capturar patrones lingüísticos complejos, su adopción continúa condicionada por restricciones técnicas y operativas.

Como respuesta a estos desafíos, los modelos basados en arquitecturas *transformer* han adquirido una relevancia creciente en el análisis de sentimientos. A diferencia de las redes recurrentes, estos modelos incorporan mecanismos de atención que permiten ponderar la importancia relativa de las palabras dentro de una secuencia y capturar relaciones contextuales de forma más eficiente. En esta línea, (Tan, Lee, Anbananthan, y Lim, 2022) proponen un modelo híbrido que combina *RoBERTa* y *LSTM* con el objetivo de mejorar la detección de sentimientos en reseñas y textos breves. Los resultados reportados muestran que esta combinación permite aprovechar las fortalezas de ambas arquitecturas: mientras la *LSTM* contribuye a modelar dependencias secuenciales, *RoBERTa* aporta representaciones bidireccionales más ricas y sensibles al contexto. En consecuencia, este tipo de enfoques alcanza niveles elevados de rendimiento, con *F1-scores* situados aproximadamente entre el 90 % y el 93 %, lo que evidencia el potencial de los *transformers* para mejorar la clasificación de sentimientos en textos con mayor complejidad semántica.

Además de los modelos híbridos, también se han desarrollado investigaciones centradas exclusivamente en arquitecturas *transformer*. (Perikos y Diamantopoulos, 2024) evalúan distintos modelos de este tipo en tareas de análisis de sentimientos basado en aspectos, incorporando además técnicas de interpretabilidad para analizar qué palabras y contextos influyen en las predicciones. El estudio muestra que *RoBERTa* obtiene un rendimiento destacado, especialmente por su capacidad para capturar matices contextuales y asociar polaridades específicas a determinados atributos mencionados en las reseñas. Este resultado refuerza la relevancia de los *transformers* en escenarios donde no solo interesa clasificar la polaridad general del texto, sino también comprender qué elementos concretos del producto o servicio explican dicha valoración.

No obstante, estos avances no eliminan completamente los retos observados en los modelos de aprendizaje profundo. Las arquitecturas *transformer* suelen requerir mayores recursos computacionales, memoria *GPU* y tiempos de procesamiento más elevados, especialmente cuando se trabaja con modelos de gran tamaño o conjuntos de datos extensos. Asimismo, su

rendimiento puede depender de la calidad de los datos etiquetados, del idioma, del dominio analizado y de la presencia de fenómenos lingüísticos complejos como la ironía, el sarcasmo o la ambigüedad. Estas limitaciones han impulsado una nueva línea de investigación orientada al uso de *LLMs*, los cuales amplían las capacidades de comprensión y generación del lenguaje natural, y permiten abordar tareas de análisis de sentimientos de forma más flexible y adaptativa (Ghatora et al., 2024).

Dentro de esta línea de avance, algunos estudios han explorado el uso de *LLMs* adaptados específicamente a tareas de análisis de sentimientos. (Zhan, Shi, Shi, Li, y Lin, 2024) investigan la optimización de *GPT-3* mediante técnicas de ajuste fino (*fine-tuning*) para mejorar la clasificación de opiniones en reseñas etiquetadas. Este enfoque se apoya en el aprendizaje por transferencia, mediante el cual el conocimiento adquirido por el modelo durante su entrenamiento general puede reutilizarse y ajustarse a una tarea específica. Los autores muestran que la adaptación del modelo al dominio del análisis de sentimientos permite mejorar su capacidad de clasificación y alcanzar una exactitud del 85 %. Este tipo de aproximación resulta relevante porque evidencia que los *LLMs* pueden capturar representaciones contextuales complejas y adaptarse a dominios concretos, aunque dicha adaptación puede seguir implicando costes computacionales y disponibilidad de datos etiquetados.

Junto con el ajuste fino, el *prompting* se ha consolidado como una estrategia alternativa para orientar el comportamiento de los *LLMs* sin modificar sus parámetros internos. Esta técnica consiste en diseñar instrucciones específicas que guían al modelo en la realización de una tarea, permitiendo estructurar sus respuestas y adaptar el análisis al objetivo del usuario. En el ámbito del análisis de sentimientos, (Miriyala, Bukkapatnam, y Prahallad, 2025) comparan un *prompt* simple con una estrategia de *Chain-of-Thought prompting* aplicada a reseñas de productos, observando que esta última mejora la precisión del modelo al inducir un proceso de razonamiento más estructurado antes de emitir la clasificación final. De este modo, el *prompting* representa una vía flexible y de menor coste que el reentrenamiento completo, ya que permite mejorar la coherencia y utilidad de las respuestas del modelo. No obstante, su efectividad depende de la calidad del diseño del *prompt*, del tipo de tarea y de la complejidad del lenguaje presente en las reseñas analizadas.

En esta misma línea, (C. Li et al., 2021) proponen *SentiPrompt*, un método de *prompt-tuning* orientado al análisis de sentimientos basado en aspectos. A diferencia de los *prompts* genéricos, esta aproximación incorpora conocimiento semántico y emocional para modelar de forma explícita la relación entre el aspecto evaluado, la opinión expresada y la polaridad asociada. Los autores comparan esta estrategia con distintas arquitecturas, entre ellas *BERT*, *K-BERT*, *T5* y *LCF-BERT*, observando mejoras de entre 2,7 y 3,9 puntos porcentuales en *F1-score*, así como una precisión promedio del 84,5 %. Estos resultados sugieren que el diseño de instrucciones específicas y orientadas al dominio puede mejorar la capacidad de los modelos para interpretar relaciones sentimentales complejas en reseñas, especialmente cuando el objetivo no se limita a clasificar la polaridad global, sino a identificar sentimientos asociados

a atributos concretos del producto o servicio.

A pesar de estos avances, la aplicación práctica de los LLMs continúa condicionada por sus elevados requerimientos computacionales, tanto en memoria como en tiempo de inferencia. Por ello, una línea reciente de investigación se ha centrado en el uso de *LLMs* cuantizados. La cuantización consiste en reducir la precisión numérica con la que se representan los parámetros del modelo, por ejemplo, pasando de representaciones de 16 o 32 bits a formatos de menor precisión, como 8 o 4 bits. Esta reducción permite disminuir el consumo de memoria y acelerar la inferencia, aunque puede implicar pérdidas de rendimiento si no se aplica de forma adecuada. En este contexto, (T. Li, Li, y Zhou, 2025) proponen un enfoque basado en *LLaMA3* ajustado mediante *QLoRA*, combinando cuantización a 4 bits con el entrenamiento de adaptadores *LoRA*. Esta estrategia permite adaptar el modelo al dominio de reseñas de productos sin requerir infraestructuras computacionales de gran escala. En su estudio, el modelo ajustado alcanza una precisión del 86 %, frente al bajo rendimiento del modelo sin adaptación, lo que pone de manifiesto que la cuantización, combinada con técnicas de adaptación eficiente, puede constituir una alternativa viable para aplicar LLMs en tareas de análisis de sentimientos en contextos empresariales con recursos limitados.

La Tabla 2.1 sintetiza los principales estudios revisados, destacando el tamaño de los conjuntos de datos utilizados, los algoritmos empleados y los resultados más relevantes reportados en cada caso. En términos generales, se observa una evolución desde estudios basados en enfoques léxicos, aplicados sobre conjuntos de reseñas de tamaño moderado, hacia investigaciones que emplean modelos supervisados, arquitecturas de *deep learning*, *transformers* y LLMs sobre volúmenes de datos más amplios y diversos. Esta progresión refleja no solo el aumento en la complejidad de los algoritmos utilizados, sino también una mayor exigencia en cuanto a disponibilidad de datos, capacidad computacional y calidad del preprocesamiento. Asimismo, la tabla permite identificar que los modelos basados en diccionarios destacan por su facilidad de implementación y bajo coste computacional, aunque tienden a mostrar un rendimiento más limitado frente a los modelos supervisados y de aprendizaje profundo. Los enfoques tradicionales de aprendizaje automático, como la regresión logística, Máquinas de Vectores Soporte y *Gradient Boosting*, presentan mejoras relevantes en precisión cuando se dispone de datasets etiquetados de mayor tamaño. Por su parte, las arquitecturas basadas en *deep learning* y *transformers* alcanzan métricas más elevadas al capturar mejor el contexto y las relaciones semánticas del texto, aunque requieren más recursos y una preparación más rigurosa de los datos.

Finalmente, los estudios recientes con LLMs, *prompting* y modelos cuantizados muestran una tendencia hacia soluciones más flexibles y eficientes, orientadas a equilibrar rendimiento, escalabilidad y viabilidad de implementación en entornos empresariales. No obstante, la literatura también señala que los LLMs convencionales mantienen desafíos relevantes. En este sentido, (Ponnusamy, Sonvani, Traylor, y Nanda, 2025) comparan modelos como *ChatGPT*, *Claude AI*, *BARD* y *Azure AI* en tareas de análisis de sentimientos sobre más de 18.000

reseñas de productos de Amazon y BestBuy, contrastando sus predicciones con evaluaciones humanas. Sus análisis indican que, aunque estos modelos presentan una alta capacidad para analizar opiniones de consumidores, todavía pueden mostrar dificultades para identificar sentimientos neutros, sesgos en la interpretación emocional y elevados requerimientos computacionales. Estos aspectos refuerzan la pertinencia de explorar alternativas basadas en LLMs cuantizados y estrategias de *prompting*, como propone el presente trabajo.

A partir de esta revisión, el presente Trabajo de Fin de Grado se sitúa en la intersección entre el análisis de sentimientos, el uso de LLMs cuantizados y el diseño de estrategias de *prompting*. En concreto, se propone evaluar si un *LLM* optimizado mediante cuantización puede ofrecer un desempeño competitivo en la clasificación de sentimientos de reseñas de clientes, reduciendo al mismo tiempo las barreras computacionales asociadas a los LLMs convencionales. De este modo, el trabajo busca contribuir tanto al ámbito académico, mediante la comparación metodológica con enfoques tradicionales, como al ámbito empresarial, al explorar una alternativa eficiente, replicable y aplicable al análisis de percepciones de clientes en contextos reales.

Tabla 2.1: Resumen de los estudios sobre el análisis de sentimientos

Referencia	Tamaño del Dataset	Objetivo	Algoritmo	Resultados
(Cho et al., 2014)	17.500 reseñas de smartphones, 35.000 de películas y 90.000 de libros.	Desarrollar un método de análisis de sentimientos basado exclusivamente en diccionarios léxicos, capaz de mejorar la clasificación de reseñas de productos mediante la integración y adaptación automática de múltiples diccionarios de sentimiento a distintos dominios.	Diccionarios: AFINN, General Inquirer, Opinion Lexicon, SenticNet, SentiWordNet, SO-CAL, Subjectivity Lexicon, WordNet-Affect.	Se concluye una precisión balanceada aproximada del 82.6 % en reseñas de smartphones, 80.1 % en películas y 81.8 % en libros.
(Nguyen et al., 2018)	43.620 reseñas de productos de Amazon.	Comparar el rendimiento de los enfoques basados en diccionarios léxicos frente a técnicas de aprendizaje automático.	Seis modelos de clasificación: tres enfoques de aprendizaje automático (<i>Logistic Regression</i> , <i>Support Vector Machine</i> y <i>Gradient Boosting</i>) y tres enfoques basados en diccionarios léxicos (<i>VADER</i> , <i>Pattern</i> y <i>SentiWordNet</i>).	Los resultados muestran que los modelos de aprendizaje automático alcanzan mayores niveles de precisión (90 % para Logistic Regression, 89 % para SVM y 87 % para Gradient Boosting) que los modelos basados en diccionarios (69 % para Pattern, 83 % para VADER y 80 % para SentiWordNet), evidenciando diferencias claras en el rendimiento de ambos enfoques.
(Ain et al., 2017)	Revisión Comparativa de modelos.	Analizar los avances del aprendizaje profundo frente a métodos tradicionales.	Modelos de aprendizaje profundo: CNN, RNN, LSTM, GRU.	Los investigadores señalan que las redes neuronales muestran mayor precisión y capacidad contextual, al mismo tiempo que reducen la intervención humana y admiten datos multimodales (texto, imagen, vídeo).
(Panthati et al., 2018)	3.000 reseñas de productos de Amazon.	Desarrollar y evaluar un modelo de análisis de sentimientos basado en Deep Learning que permita clasificar automáticamente las reseñas de productos como positivas o negativas.	Combina <i>Word2Vec</i> para la generación de embeddings de palabras y una red neuronal recurrente <i>LSTM</i> (<i>Long Short Term Memory</i>).	Los resultados experimentales muestran que el modelo basado en LSTM alcanza una precisión aproximada del 93,33 % en la tarea de clasificación de sentimientos sobre las reseñas de productos.
(Tan et al., 2022)	IMDb: 50.000 reseñas, Senti-ment140: 1.6 millones de <i>tweets</i> , US Airlines: 14.640 <i>tweets</i> .	Evaluar el desempeño de un modelo que combina un transformer con una red recurrente, para mejorar coherencia y precisión.	Modelo híbrido Transformer + RNN.	Esta propuesta de modelo consigue un F1-score entre 90–93 %. También, reduce el tiempo de entrenamiento y mantiene coherencia contextual.
(Perikos y Diamantopoulos, 2024)	7.000 reseñas sobre el restaurante SemEval-2014 y ordenadores portátiles.	Analizar el rendimiento de modelos transformer explicables.	Modelos de aprendizaje profundo <i>transformers</i> : BERT y RoBERTa.	Se determina que RoBERTa obtiene la mayor precisión y coherencia contextual, aunque con alto costo computacional.
(Zhan et al., 2024)	Entre 5.000-10.000 textos (IMDb y conjuntos de reseñas empresariales).	Aplicar fine-tuning para adaptar GPT-3 a dominios específicos.	LLM con ajuste fino (fine-tuning).	Obtención de una precisión del 85 %. Este modelo presenta una elevada generalización y adaptación contextual mediante aprendizaje por transferencia.
(Miriyyala et al., 2025)	2.000 reseñas de la aplicación de Amazon (Amazon App Store reviews).	Evaluar si el uso de técnicas de prompting basadas en razonamiento explícito (<i>Chain-of-Thought</i>) mejora la clasificación de sentimientos frente a prompts.	Emplea un modelo de lenguaje grande (<i>GPT-4</i>) y compara dos estrategias de prompting: el <i>Simple Prompting</i> y el <i>Chain-of-Thought</i> (<i>CoT</i>) <i>Prompting</i> .	El enfoque de <i>Simple Prompting</i> alcanza una precisión del 84 %, mientras que el método de <i>Chain-of-Thought Prompting</i> mejora el rendimiento hasta un 93 % de precisión.

Referencia	Tamaño del Dataset	Objetivo	Algoritmo	Resultados
(C. Li et al., 2021)	3.000-4.000 reseñas obtenidas del restaurante SemEval-2014 y bases de datos sobre ordenadores portátiles.	Incorporar conocimiento semántico y emocional al LLM a través de técnicas de <i>prompts</i> .	Prompt-tuning con conocimiento de sentimiento para LLMs.	Se determina que el uso de <i>prompts</i> mejora el F1-score entre 2.7 y 3.9 puntos, frente a modelos base. Además la precisión es del 84.5 %. Se obtiene un modelo eficiente y sin necesidad de reentrenamiento.
(T. Li et al., 2025)	4.846 reseñas de productos de Amazon.	Analizar la eficacia de la adaptación de un modelo LLM (<i>LLaMA3</i>) mediante técnicas de fine-tuning eficientes en parámetros (<i>QLoRA</i>).	El estudio emplea el modelo <i>LLaMA3-8B</i> , al que se le aplica cuantización a 4 bits y <i>Low-Rank Adaptation (LoRA)</i> mediante el enfoque <i>QLoRA</i>	Los resultados muestran que el modelo preentrenado sin adaptación obtiene una precisión baja 37 %, mientras que el modelo ajustado con QLoRA alcanza una precisión global de 86 %.
(Ponnusamy et al., 2025)	18.000 reseñas de Amazon y Best-Buy.	Comparar desempeño de LLM con juicios humanos en clasificación de sentimientos.	LLMs: ChatGPT, Claude AI, Bard, Azure AI.	Los investigadores resaltan que los LLMs muestran sesgo en sentimientos neutros, alto costo computacional y menor comprensión contextual en comparación con los humanos.

Capítulo 3

Diseño metodológico

El presente capítulo describe la metodología desarrollada para evaluar el uso de *LLMs* cuantizados en tareas de análisis de sentimientos aplicadas a reseñas de productos en español. El objetivo principal consiste en analizar la capacidad de estos modelos para clasificar percepciones de clientes en contextos de comercio electrónico, considerando tanto la calidad de las predicciones como la viabilidad de su implementación en entornos con recursos computacionales limitados. Asimismo, se estudia el impacto de distintas estrategias de *prompting* sobre el desempeño de los modelos en tareas de clasificación de sentimientos.

Como se observa en la Figura 3.1, la metodología se estructura en tres etapas principales. La primera etapa corresponde a la selección del conjunto de datos de evaluación. En esta fase se identifica un corpus público de reseñas de productos en español, etiquetado por polaridad, y se adapta a una estructura de polaridad compuesta por tres clases: positivo, negativo y neutro. Asimismo, se selecciona un subconjunto de evaluación mediante muestreo aleatorio, con el fin de disponer de una muestra adecuada para el análisis experimental. La segunda etapa comprende la selección del modelo y de las técnicas utilizadas. En este trabajo se emplean *LLMs* cuantizados, con el propósito de reducir el consumo de memoria y facilitar su ejecución en hardware con recursos limitados. En particular, se considera la técnica de PTQ en su variante *Weight-Only Quantization* a 4 bits. La selección del modelo se realiza mediante una evaluación comparativa basada en el consumo de memoria, la latencia y la perplejidad. Además, se implementan distintas técnicas de *prompting*, como *zero-shot*, *one-shot*, *few-shot*, *chain-of-thought*, *role-based* e *iterative*, con el fin de analizar su efecto sobre la clasificación de sentimientos sin necesidad de reentrenar el modelo.

La tercera etapa corresponde a la evaluación del desempeño. En esta fase se analiza la calidad de las clasificaciones producidas por cada combinación de modelo y técnica de *prompting*. Para ello, se emplean matrices de confusión, métricas estándar de clasificación multiclase, como *accuracy*, *precision*, *recall*, *F1-score* y *F1 macro-promediada*, así como intervalos de confianza al 95 % mediante *bootstrap*. Esta evaluación permite comparar el rendimiento global de los modelos y analizar su comportamiento diferenciado según la clase

de sentimiento. A continuación se presentan los detalles del desarrollo de cada una de las etapas resumidas anteriormente.



Figura 3.1: Etapas de la metodología seguida en este estudio

Fuente: Elaboración propia

3.1. Selección del conjunto de datos

La selección del corpus de evaluación constituye una decisión metodológica relevante, ya que condiciona la validez del análisis y la fiabilidad de las comparaciones entre modelos y estrategias de *prompting*. A diferencia de los enfoques supervisados tradicionales, en los que el conjunto de datos se emplea para entrenar el modelo, en este estudio el corpus se utiliza exclusivamente con fines de evaluación. Las reseñas se introducen en los modelos mediante *prompts*, y las predicciones generadas se comparan con las etiquetas reales para estimar el rendimiento del sistema. Por tanto, los criterios de selección del conjunto de datos se orientan a garantizar su adecuación al dominio de aplicación, la disponibilidad de etiquetas de referencia y un tamaño suficiente para realizar una evaluación consistente.

El primer criterio considerado es la disponibilidad de etiquetas de sentimiento asociadas a cada reseña, dado que estas permiten contrastar de forma objetiva las predicciones generadas por los modelos. En ausencia de este etiquetado, la evaluación dependería de juicios manuales adicionales o de procedimientos indirectos, lo que introduciría mayor subjetividad en la comparación. Asimismo, se priorizan conjuntos de datos con un volumen amplio de observaciones, ya que un mayor tamaño muestral reduce la sensibilidad de los resultados a casos particulares y permite obtener estimaciones más estables del desempeño. Dado que el

estudio se centra en reseñas de productos de comercio electrónico en español, también se considera preferible emplear textos originalmente escritos en este idioma, evitando traducciones automáticas que puedan alterar matices léxicos, o sintácticos relevantes para la clasificación de sentimientos.

En relación con la distribución de clases, se reconoce que los corpus de reseñas de comercio electrónico suelen presentar desequilibrios naturales, especialmente por la mayor presencia de valoraciones claramente positivas o negativas frente a opiniones neutras o intermedias (Liu, 2012). En lugar de forzar una distribución equilibrada, se mantiene la estructura original del corpus con el fin de preservar las condiciones propias del dominio de aplicación. Esta decisión permite evaluar el comportamiento de los modelos en un escenario más próximo al uso real, aunque exige emplear métricas que no dependan únicamente del rendimiento global. Por ello, junto con la exactitud, se incorporan medidas como la *F1 macro-promediada*, que asigna el mismo peso a cada clase y resulta adecuada cuando existe desbalance entre categorías (Sokolova y Lapalme, 2009).

Las etiquetas originales del corpus seleccionado se transforman en una estructura de polaridad compuesta por tres clases: positivo, negativo y neutro. Esta codificación permite representar la orientación general de las reseñas sin introducir niveles de granularidad que podrían dificultar la comparación entre modelos o aumentar la ambigüedad de la tarea. A partir del corpus completo, se selecciona un subconjunto de evaluación mediante muestreo aleatorio con semilla fija, garantizando así la reproducibilidad del experimento. El tamaño final de la muestra se define considerando las restricciones computacionales del entorno de ejecución y la necesidad de mantener una base suficiente para comparar de forma razonable las distintas combinaciones de modelos y técnicas de *prompting*.

3.2. Modelo de lenguaje cuantizado y estrategias de *prompting*

En este proyecto se emplea un *LLM* para abordar la tarea de análisis de sentimientos en reseñas de productos. Tal como se ha expuesto en la revisión de la literatura, estos modelos han mostrado una elevada capacidad para procesar lenguaje natural de forma contextual y adaptativa, lo que los hace especialmente adecuados para tareas en las que la interpretación del significado depende del dominio, del tono y de la relación entre las palabras. En este sentido, (Ghatora et al., 2024) señalan que los *LLMs* pueden identificar relaciones semánticas complejas y adaptarse a diferentes contextos de análisis, mientras que (Zhan et al., 2024) muestran que su ajuste a tareas específicas puede mejorar el rendimiento en la clasificación de sentimientos mediante aprendizaje por transferencia.

La elección de un *LLM* como modelo principal responde, por tanto, a su capacidad para generar clasificaciones más sensibles al contexto que las aproximaciones basadas exclusiva-

mente en diccionarios o modelos tradicionales. No obstante, el uso de *LLMs* completos suele implicar elevados requerimientos de memoria, tiempos de inferencia más altos y dependencia de hardware especializado (Jin et al., 2024). Por esta razón, el presente trabajo emplea una versión cuantizada del modelo, con el propósito de reducir las exigencias computacionales y evaluar si es posible mantener un desempeño competitivo en un entorno de ejecución más accesible.

3.2.1. Cuantización de *LLMs*

La cuantización es una técnica de optimización orientada a reducir la precisión numérica con la que se representan los parámetros de un modelo. En lugar de emplear formatos de alta precisión, como *float32*, los pesos y, en algunos casos, las activaciones se representan mediante formatos más compactos, como *int8* o *int4*. Esta transformación permite disminuir el tamaño del modelo, reducir el consumo de memoria y acelerar los tiempos de inferencia, facilitando la ejecución de modelos de gran capacidad en entornos con recursos computacionales limitados. En el caso de los *LLMs*, estas ventajas resultan especialmente relevantes, ya que su despliegue suele estar condicionado por elevados requerimientos de memoria y hardware especializado. Por ello, la cuantización permite combinar la capacidad analítica de estos modelos con una implementación más accesible y eficiente, sin comprometer necesariamente de forma sustancial el rendimiento (González San José et al., 2025; Hermawan et al., 2024).

En este ámbito, los enfoques de cuantización suelen clasificarse en dos grandes familias: *Post-Training Quantization* (PTQ) y *Quantization-Aware Training* (QAT). La PTQ se aplica una vez que el modelo ha sido entrenado en alta precisión, transformando sus parámetros a representaciones de menor precisión sin modificar el proceso original de entrenamiento. Esta característica la convierte en una alternativa sencilla y eficiente cuando se trabaja con modelos ya entrenados y disponibles públicamente. Por el contrario, la QAT incorpora los efectos de la cuantización durante el entrenamiento o ajuste fino del modelo, permitiendo que este se adapte progresivamente a la baja precisión. Aunque este segundo enfoque puede preservar mejor el rendimiento, también requiere mayores recursos computacionales y acceso al proceso de entrenamiento. En este sentido, la PTQ ofrece una mayor flexibilidad para el despliegue, al permitir disponer de una versión original en alta precisión y una versión cuantizada del mismo modelo, que pueden seleccionarse según las necesidades de precisión, memoria o coste computacional (Zhao, Xu, y Guo, 2023).

Dentro de la PTQ pueden distinguirse distintas estrategias en función del componente del modelo sobre el que se aplica la reducción de precisión. La primera es la *Weight-Only Quantization*, que cuantiza únicamente los pesos de las capas del modelo y mantiene las activaciones en mayor precisión. Esta aproximación permite una reducción significativa del tamaño del modelo con un impacto relativamente controlado sobre la calidad de las predic-

ciones. La segunda es la *Weight-Activation Quantization*, que extiende la cuantización tanto a los pesos como a las activaciones, logrando una mayor reducción de memoria y coste de cómputo, aunque con una mayor sensibilidad a errores derivados de la baja precisión. Finalmente, la *KV Cache Quantization* se centra en los tensores de clave y valor almacenados en la memoria caché de los bloques de autoatención. Esta estrategia resulta especialmente útil en modelos autorregresivos, donde el almacenamiento del historial de contexto puede convertirse en un cuello de botella durante la inferencia, sobre todo cuando se procesan secuencias largas (S. Li et al., 2024).

La relevancia práctica de estas técnicas se observa también en su aplicación a familias de modelos ampliamente utilizadas. La literatura reciente ha evaluado de forma sistemática el efecto de la cuantización sobre *LLMs*, analizando el equilibrio entre reducción del coste computacional y preservación del rendimiento en tareas de comprensión y generación de texto (Jin et al., 2024; Jirasky et al., 2024). Estos trabajos muestran que la cuantización no debe entenderse únicamente como una técnica de compresión, sino como una estrategia que facilita el despliegue de *LLMs* en contextos donde la disponibilidad de memoria, capacidad de cómputo o infraestructura especializada puede ser limitada.

En el caso de los modelos de código abierto, la familia *LLaMA*, desarrollada por *Meta AI*, ha sido ampliamente utilizada en estudios sobre cuantización. Diversas evaluaciones indican que la reducción de precisión a 4 bits mediante métodos como *GPTQ* o *AWQ* permite mantener una proporción elevada del rendimiento original en *benchmarks* estándar de comprensión y generación de texto (Touvron et al., 2023). De forma complementaria, (Dettmers et al., 2023) proponen *QLoRA*, una técnica que combina cuantización a 4 bits con (LoRA), permitiendo realizar ajuste fino eficiente de modelos de gran escala con menores requerimientos de hardware. Esta aproximación resulta especialmente pertinente para tareas de dominio específico, como el análisis de sentimientos, en las que puede ser necesario adaptar el comportamiento del modelo sin asumir los costes asociados al reentrenamiento completo.

Asimismo, otros desarrollos recientes han combinado la cuantización con mejoras arquitectónicas orientadas a la eficiencia. En esta línea, *Mistral 7B* incorpora mecanismos como *Grouped-Query Attention* (GQA) y *Sliding Window Attention* (SWA), diseñados para reducir el coste de procesamiento y mejorar la eficiencia en secuencias largas (Jiang et al., 2023). Evaluaciones posteriores con cuantización mediante *AWQ* muestran que modelos arquitectónicamente eficientes pueden mantener un rendimiento competitivo incluso bajo representaciones de menor precisión (Lin et al., 2024). Esto sugiere que la pérdida asociada a la reducción numérica puede mitigarse parcialmente cuando la cuantización se combina con diseños de modelo optimizados, lo que refuerza su interés para aplicaciones empresariales que requieren equilibrio entre precisión, velocidad y consumo de recursos.

De forma paralela, otros proveedores han evaluado la cuantización en familias de modelos con distintas características arquitectónicas y lingüísticas. Desde la perspectiva de *Google DeepMind*, los modelos *Gemma* han sido analizados mediante técnicas híbridas de *post-*

training quantization (PTQ), mostrando que la reducción a 4 bits puede preservar una proporción elevada del rendimiento en tareas de comprensión lectora y razonamiento (Team, 2024). Por su parte, *Alibaba Cloud* ha contribuido al estudio de la cuantización en contextos multilingües a través de la familia *Qwen*, aspecto especialmente relevante para aplicaciones de análisis de sentimientos en mercados internacionales y entornos con diversidad lingüística (Bai et al., 2023). Estos trabajos amplían la evidencia disponible al mostrar que la cuantización no solo resulta aplicable a modelos monolingües o de propósito general, sino también a sistemas diseñados para operar en distintos idiomas y dominios. Finalmente, *Microsoft* ha explorado esta línea en modelos de menor escala orientados a tareas específicas, como *Phi-2*, destacando que la calidad del preentrenamiento y el diseño del corpus pueden contribuir a preservar el rendimiento incluso cuando se reduce la precisión numérica del modelo (Gunasekar et al., 2023). En conjunto, la evidencia disponible sugiere que la cuantización a 4 bits, aplicada mediante técnicas como *GPTQ* o *AWQ*, se ha consolidado como una alternativa razonable para equilibrar eficiencia computacional y conservación de capacidades lingüísticas (Frantar et al., 2022; Lin et al., 2024).

Con el fin de contextualizar la disponibilidad actual de modelos cuantizados, la Tabla 3.1 presenta una síntesis de familias de *LLMs* desarrolladas por distintos proveedores, junto con los métodos de cuantización empleados y el rendimiento reportado tras la reducción de precisión. Esta comparación permite observar que la cuantización a 4 bits se ha consolidado como una alternativa frecuente en modelos de código abierto y propietarios, especialmente cuando se busca reducir el consumo computacional sin deteriorar de forma significativa la calidad del modelo. La Tabla 3.1 no pretende constituir una revisión exhaustiva de todos los modelos cuantizados disponibles, sino ofrecer una visión ilustrativa de las principales alternativas documentadas en la literatura reciente y en familias de modelos ampliamente utilizadas. En conjunto, los resultados recogidos muestran que la cuantización se ha aplicado a modelos de diferentes tamaños, arquitecturas y proveedores, con niveles de preservación del rendimiento generalmente elevados. No obstante, la inclusión de estos modelos en la tabla no implica su selección directa para el desarrollo experimental del presente trabajo. La elección definitiva del modelo se realizará en una fase posterior, a partir de criterios explícitos de eficiencia computacional y calidad textual, aplicados sobre un conjunto de candidatos delimitado por su disponibilidad, compatibilidad técnica y ajuste a los objetivos específicos del estudio.

Proveedor	Modelo	Método	Precisión	Rendimiento tras cuantización
Meta AI	<i>LLaMA</i> 3.2 Instruct (3B–70B)	AWQ / GPTQ	4 bits (INT4)	>95 % rendimiento original (Touvron et al., 2023)
Meta AI	<i>QLoRA</i> (<i>LLaMA</i> -based)	GPTQ + LoRA	4 bits (INT4)	<i>Fine-tuning</i> eficiente en hardware limitado (Dettmers et al., 2023)
Mistral AI	<i>Mistral</i> 7B Instruct	AWQ	4 bits (INT4)	Degradación <2 % en <i>F1-score</i> (Jiang et al., 2023; Lin et al., 2024)
Google DeepMind	<i>Gemma</i> 3 IT (4B–27B)	AWQ / GPTQ	4 bits (INT4)	>94 % capacidad preservada (Team, 2024)
Alibaba Cloud	<i>Qwen</i> 2.5 Instruct (7B–72B)	AWQ / GPTQ	4 bits (INT4)	Degradación <3 % en <i>macro-F1</i> multi-lingüe (Bai et al., 2023)
Microsoft	<i>Phi-2</i> (2.7B)	AWQ	4 bits (INT4)	Calidad de preentrenamiento compensa pérdida representacional (Gunasekar et al., 2023)

Tabla 3.1: Resumen de estudios empíricos sobre cuantización en los principales proveedores de *large language models*

3.2.2. Criterios de selección del *LLM* cuantizado

La selección del modelo principal constituye una decisión metodológica relevante, ya que condiciona tanto la calidad de la clasificación de sentimientos en reseñas en español como la viabilidad práctica del experimento. Dado que el trabajo busca desarrollar una metodología reproducible en un entorno académico sin infraestructura especializada, la comparación se centrará en modelos cuantizados que permitan equilibrar desempeño y eficiencia computacional. Para ello, los modelos candidatos se evaluarán a partir de tres criterios principales: consumo de memoria, latencia y perplejidad. El tamaño del modelo en disco se considerará únicamente como criterio descriptivo, puesto que afecta a la descarga y almacenamiento, pero no será utilizado como factor discriminante si los modelos comparados presentan un número de parámetros y un esquema de cuantización similares (Jin et al., 2024).

La eficiencia computacional se evaluará considerando las restricciones del entorno de ejecución. En este estudio, los experimentos se realizarán en *Google Colab*, una plataforma *cloud* de acceso gratuito que permite ejecutar notebooks con GPU y que resulta habitual en contextos académicos (Bisong, 2019). En particular, se tomará como referencia una GPU *NVIDIA Tesla T4*, con 15 GB de VRAM y aproximadamente 12,7 GB de RAM de sistema. Estas características imponen límites concretos sobre el tamaño del modelo y sobre la configuración de inferencia. En este contexto, la cuantización a 4 bits mediante técnicas como *AWQ* o *GPTQ* permite reducir sustancialmente el consumo de memoria frente a versiones en mayor precisión, facilitando la ejecución de modelos de entre 7 y 8 mil millones de parámetros dentro de las restricciones disponibles (Frantar et al., 2022; Lin et al., 2024).

Como medida operativa de eficiencia se empleará el consumo máximo de memoria de la GPU durante la inferencia, expresado en gigabytes (GB). Esta métrica se registrará tanto durante la carga inicial del modelo como durante el procesamiento de un lote representativo de reseñas, considerando el valor máximo alcanzado (*peak memory usage*) como indicador de referencia (Jin et al., 2024). Se establecerá un umbral máximo admisible de 14 GB de VRAM, con el fin de reservar un margen para operaciones auxiliares del entorno y evitar fallos de ejecución dentro de los 15 GB disponibles en la *Tesla T4*. Los modelos que superen dicho umbral quedarán excluidos de la selección final.

La segunda dimensión de evaluación será la latencia, entendida como el tiempo transcurrido entre la recepción de una reseña y la generación completa de la respuesta del modelo. Esta métrica permite valorar la viabilidad del sistema en escenarios donde debe procesarse un volumen elevado de opiniones en un tiempo razonable (Na et al., 2024). En este estudio, la latencia se medirá a partir de dos indicadores complementarios. En primer lugar, se calculará la velocidad de generación en *tokens* por segundo, registrando el tiempo necesario para producir una cantidad fija de *tokens* a partir de un mismo *prompt* bajo condiciones experimentales constantes. En segundo lugar, se medirá el tiempo transcurrido entre el envío de cada reseña al modelo y la recepción de su clasificación completa, utilizando marcas de

tiempo antes y después de cada llamada de inferencia. A partir de estas mediciones se obtendrán la latencia individual, definida como el tiempo medio de generación por reseña, y el tiempo total de inferencia, correspondiente al tiempo acumulado necesario para procesar el conjunto de evaluación completo (Jin et al., 2024). La disponibilidad de una GPU dedicada en *Google Colab* contribuirá a acelerar la inferencia, mientras que la cuantización permitirá reducir tanto la latencia como el consumo de VRAM (Na et al., 2024).

La tercera dimensión será la perplejidad (*perplexity*), una métrica de evaluación intrínseca ampliamente utilizada en modelos de lenguaje para estimar la calidad generativa del modelo y su capacidad para predecir secuencias de texto (Jelinek, Mercer, Bahl, y Baker, 1977). Formalmente, la perplejidad se define como la exponencial de la entropía cruzada media del modelo sobre un corpus de evaluación. Valores más bajos indican que el modelo asigna mayor probabilidad a las secuencias observadas, lo que suele interpretarse como una mayor coherencia lingüística y una mejor capacidad predictiva (Brown, Della Pietra, Della Pietra, Lai, y Mercer, 1992; Jelinek et al., 1977). En este trabajo, esta métrica se utilizará para valorar si la cuantización a 4 bits introduce una degradación apreciable en la calidad generativa del modelo. Para calcular la perplejidad de forma objetiva y reproducible se empleará el corpus *WikiText-2*, un *benchmark* estándar en la evaluación de modelos de lenguaje (Merity, Xiong, Bradbury, y Socher, 2017). La elección de este corpus, en lugar del conjunto de reseñas utilizado posteriormente para la clasificación de sentimientos, responde a dos razones metodológicas. En primer lugar, *WikiText-2* se utiliza con frecuencia en estudios sobre cuantización de *LLMs*, incluidos los trabajos de (Frantar et al., 2022) sobre *GPTQ* y de (Detters et al., 2023) sobre *QLoRA*, lo que facilita la comparación con valores reportados en la literatura. En segundo lugar, la perplejidad evalúa la capacidad predictiva general del modelo sobre lenguaje natural, por lo que un corpus estandarizado ofrece una referencia más estable que un corpus de dominio específico, cuyas particularidades lingüísticas podrían sesgar la medición.

Como criterio de referencia, se considerará la degradación relativa de la perplejidad respecto al modelo base en alta precisión. En particular, (Frantar et al., 2022) reportan que, cuando la cuantización a 4 bits se aplica con procedimientos de calibración adecuados, el incremento de perplejidad suele ser inferior al 1 % respecto al modelo original en modelos de entre 7 y 70 mil millones de parámetros. En consecuencia, en este trabajo se considerará aceptable un modelo cuantizado cuya perplejidad no supere en más de un 1 % la perplejidad de referencia de su versión no cuantizada, siempre que también cumpla los criterios de memoria y latencia establecidos. Los modelos que presenten una degradación sustancialmente superior serán descartados del proceso de selección.

3.2.3. Estrategias de *prompting* para la clasificación de sentimientos

El *prompting* constituye un mecanismo central para adaptar los *LLMs* a tareas específicas de *NLP*. Se entiende como la formulación de una instrucción textual dirigida a un modelo generativo con el propósito de orientar su respuesta hacia un resultado determinado (Wang y Luo, 2023). A diferencia de los enfoques supervisados tradicionales, que requieren conjuntos amplios de datos etiquetados y procesos de ajuste del modelo, el *prompting* permite guiar el comportamiento de un *LLM* mediante instrucciones en lenguaje natural, sin modificar su arquitectura interna ni realizar procesos de ajuste fino (*fine-tuning*) computacionalmente costosos (Wang y Luo, 2023).

En el análisis de sentimientos, el *prompt* funciona como el principal mecanismo de control de la tarea. A través de su diseño se especifica el rol que debe adoptar el modelo, la clasificación que debe realizar, el formato esperado de la respuesta y, cuando procede, ejemplos de referencia. De este modo, el *prompt* traduce los objetivos metodológicos del estudio en instrucciones comprensibles para el *LLM*, orientando la clasificación de cada reseña en categorías como positiva, negativa o neutra. En esta línea, Wang y Luo (2023) muestran que una formulación estructurada de instrucciones puede mejorar el rendimiento de modelos generativos en tareas de análisis de sentimientos sobre distintos dominios. Esta cuestión resulta especialmente relevante en entornos con recursos computacionales limitados, donde el uso de modelos cuantizados reduce la posibilidad de recurrir a procesos adicionales de ajuste o reentrenamiento.

A partir de esta premisa, las estrategias de *prompting* pueden organizarse en dos grandes categorías metodológicas. La primera corresponde a las técnicas directas, en las que la respuesta del modelo se obtiene en una única interacción a partir de una instrucción inicial. La segunda comprende las técnicas multipaso, que incorporan fases adicionales de razonamiento, revisión, evaluación o refinamiento de la respuesta generada. Esta distinción permite ordenar las estrategias consideradas en el presente estudio según el grado de intervención que se introduce en el proceso de generación. A continuación, se describen las principales técnicas de *prompting* empleadas para la clasificación de sentimientos.

1. **Zero-shot prompting:** El *zero-shot prompting* consiste en proporcionar al modelo únicamente la instrucción de la tarea, sin incluir ejemplos previos que orienten la respuesta. En este enfoque, el *LLM* debe apoyarse en el conocimiento adquirido durante su preentrenamiento para interpretar la consigna, reconocer la estructura del problema y generar una salida coherente con las categorías solicitadas (Nádas, Dheeban, y Tomescu, 2025). Esta estrategia resulta adecuada cuando no se dispone de datos etiquetados de referencia o cuando se busca evaluar la capacidad generalista del modelo sin introducir condicionamientos derivados de ejemplos específicos. En tareas de clasificación de texto, (Chakraborty, Kulkarni, y Li, 2023) muestran que un diseño cuidadoso del *prompt* puede alcanzar un rendimiento competitivo frente a estrategias más comple-

jas, especialmente en dominios donde las opiniones son explícitas y presentan baja ambigüedad. En el contexto del presente trabajo, un ejemplo de *zero-shot prompting* aplicado al análisis de sentimientos de reseñas sería el siguiente:

“Eres un experto en análisis de sentimientos. Clasifica el sentimiento de la siguiente reseña en una de estas tres categorías: positivo, negativo o neutro. Responde únicamente con la etiqueta correspondiente. Reseña: ‘El producto llegó en perfectas condiciones y el servicio de atención al cliente fue excelente, aunque el precio me parece algo elevado.’ Sentimiento:”

En este caso, el *prompt* proporciona la definición de la tarea, delimita las categorías posibles y establece una restricción explícita sobre el formato de salida, pero no ofrece ejemplos de clasificación. Esta simplicidad facilita su implementación y reduce el coste asociado al diseño de instrucciones extensas. No obstante, al depender exclusivamente de la interpretación interna del modelo, puede producir respuestas menos consistentes cuando las reseñas contienen opiniones mixtas, expresiones implícitas o ambigüedades semánticas (Chakraborty et al., 2023). Por ello, el *zero-shot prompting* se considera en este estudio como una estrategia base frente a la cual comparar técnicas de *prompting* más guiadas.

2. ***One-shot y few-shot prompting***: El *one-shot prompting* incorpora un único ejemplo dentro del *prompt* con el propósito de mostrar al modelo el tipo de respuesta esperado, tanto en términos de formato como de criterio de clasificación (Nádas et al., 2025). A diferencia del enfoque *zero-shot*, en el que el modelo debe inferir completamente la tarea a partir de una instrucción general, esta estrategia proporciona una demostración mínima que ayuda a reducir la ambigüedad de la consigna. En tareas de análisis de sentimientos, el ejemplo incluido permite al *LLM* identificar la correspondencia entre una reseña y una etiqueta de polaridad, favoreciendo una salida más alineada con el esquema de clasificación definido. No obstante, al basarse en un único caso de referencia, el modelo puede verse excesivamente condicionado por las características concretas del ejemplo seleccionado, especialmente si este no representa adecuadamente la diversidad lingüística del corpus. Un ejemplo de *one-shot prompting* aplicado al análisis de sentimientos de reseñas sería el siguiente:

“Ejemplo: Reseña: ‘La calidad del material es excelente y el envío fue muy rápido.’ Sentimiento: Positivo. Ahora clasifica el sentimiento de la siguiente reseña: Reseña: ‘El producto no corresponde a la descripción y el servicio postventa fue deficiente.’ Sentimiento:”

El *few-shot prompting* amplía esta lógica mediante la inclusión de varios ejemplos demostrativos en el *prompt*. Esta estrategia permite al modelo observar distintos pa-

trones de asociación entre texto y etiqueta, lo que puede mejorar la consistencia de la clasificación cuando las reseñas presentan variaciones de tono, intensidad emocional o estructura discursiva (Fatemi y Hu, 2023). En este sentido, Fatemi y Hu (2023) muestran que el *few-shot prompting* puede alcanzar resultados cercanos a los de modelos especializados en tareas de clasificación de textos de comercio electrónico. De forma complementaria, Schmitt, Schwerk, y Lempert (2026) señalan que esta técnica puede resultar eficaz incluso en modelos de menor tamaño, al proporcionar referencias que ayudan a interpretar casos ambiguos, opiniones mixtas o expresiones implícitas sin modificar los parámetros internos del sistema. Un ejemplo de *few-shot prompting* para el presente trabajo sería:

“Ejemplos: 1. Reseña: ‘Producto de gran calidad, lo recomiendo sin dudar.’ Sentimiento: Positivo. 2. Reseña: ‘Llegó tarde y el embalaje estaba dañado.’ Sentimiento: Negativo. 3. Reseña: ‘Cumple con lo descrito, sin más.’ Sentimiento: Neutro. Ahora clasifica el sentimiento de la siguiente reseña: Reseña: ‘El diseño es bonito pero la batería dura muy poco.’ Sentimiento:”

Estas estrategias permiten incorporar información contextual adicional sin recurrir al ajuste fino del modelo. Su principal ventaja reside en que orientan la respuesta del *LLM* hacia el formato y los criterios de decisión esperados, mejorando la fidelidad respecto al esquema de clasificación. Sin embargo, su rendimiento depende de la calidad, diversidad y representatividad de los ejemplos incluidos. Si los casos demostrativos son demasiado homogéneos o no reflejan la variedad real de las reseñas, el modelo puede reproducir patrones superficiales y reducir su capacidad de generalización (Nádas et al., 2025).

3. ***Chain-of-thought prompting***: El *chain-of-thought prompting* consiste en guiar al modelo para que produzca una secuencia de razonamiento intermedio antes de emitir la respuesta final. Esta estrategia resulta especialmente adecuada en tareas que requieren interpretar relaciones contextuales, ponderar información contradictoria o analizar distintos elementos dentro de un mismo texto (Miriyyala et al., 2025). En el análisis de sentimientos, su utilidad radica en que permite descomponer la clasificación en pasos sucesivos, como la identificación de expresiones positivas, la detección de elementos negativos y la valoración del sentimiento predominante. De este modo, el modelo no se limita a asignar una etiqueta de forma directa, sino que es orientado hacia una evaluación más estructurada del contenido de la reseña.

La literatura ha mostrado que esta estrategia puede mejorar el desempeño en contextos donde la polaridad no es evidente. En particular, Miriyyala et al. (2025) comparan un *prompt* directo con un enfoque basado en *chain-of-thought* en una muestra de 2.000

reseñas de Amazon. Sus resultados muestran un incremento de precisión del 84 % al 93 %, con mejoras especialmente relevantes en reseñas que contienen opiniones mixtas, múltiples atributos del producto o expresiones ambiguas. Estos resultados sugieren que inducir un razonamiento estructurado puede ayudar al modelo a procesar con mayor precisión textos en los que coexisten valoraciones positivas y negativas. Un ejemplo de *chain-of-thought prompting* aplicado al análisis de sentimientos de reseñas sería el siguiente:

“Analiza el sentimiento de la siguiente reseña paso a paso. Primero, identifica las expresiones positivas que aparecen. Segundo, identifica las expresiones negativas. Tercero, valora cuál predomina en el conjunto del texto. Finalmente, clasifica el sentimiento global como positivo, negativo o neutro. Reseña: ‘La cámara hace fotos increíbles y el diseño es elegante, pero el precio es desorbitado y la batería dura muy poco.’ Sentimiento:”

En el presente trabajo, esta estrategia se incorpora con el propósito de evaluar si una instrucción basada en razonamiento explícito mejora la clasificación de reseñas con polaridad mixta o ambigua frente a técnicas más directas como el *zero-shot* o el *few-shot prompting*. No obstante, también se tendrá en cuenta que este enfoque puede incrementar la longitud de la respuesta generada y, por tanto, aumentar el tiempo de inferencia. Esta consideración resulta relevante en el marco de modelos cuantizados, donde se busca equilibrar calidad clasificatoria y eficiencia computacional.

4. ***Role-based prompting***: El *role-based prompting* consiste en asignar al modelo un rol específico antes de formular la tarea, con el fin de orientar su respuesta hacia un marco de interpretación determinado. Esta estrategia parte de la premisa de que el contexto inicial proporcionado al *LLM* puede influir en el tipo de razonamiento, el vocabulario utilizado y los criterios que el modelo prioriza durante la generación (Zheng, Pei, Logeswaran, Lee, y Jurgens, 2024). En tareas de análisis de sentimientos, la asignación de un rol vinculado a la experiencia del consumidor puede ayudar a focalizar la clasificación en aspectos como la valoración subjetiva, la carga emocional y la percepción global del producto.

La literatura muestra que la efectividad de esta técnica depende del rol seleccionado y de su adecuación a la tarea. (Zheng et al., 2024) evalúan un amplio conjunto de roles y observan que su impacto sobre el rendimiento del modelo varía de forma considerable según el contexto de aplicación. De forma similar, (Diep, 2025) señalan que diferentes formulaciones de rol pueden producir variaciones relevantes en la calidad de las respuestas, lo que evidencia la necesidad de diseñar instrucciones alineadas con el objetivo analítico. En este trabajo, el rol se formula de manera específica para el

dominio de reseñas de comercio electrónico en español, evitando descripciones demasiado generales que puedan generar respuestas poco orientadas a la tarea. Un ejemplo de *role-based prompting* orientado al presente trabajo sería:

“Actúa como un analista experto en experiencia del consumidor con amplia experiencia en el análisis de reseñas de productos de comercio electrónico en español. Tu tarea es clasificar el sentimiento de la siguiente reseña en una de estas categorías: positivo, negativo o neutro, prestando especial atención a la carga emocional y a la valoración global del producto. Reseña: ‘Esperaba más por el precio que pagué, aunque la entrega fue rápida y el producto llegó bien embalado.’ Sentimiento:”

En este caso, el rol asignado delimita el marco interpretativo del modelo y refuerza los criterios relevantes para la clasificación. A diferencia de un *prompt* meramente instructivo, esta estrategia introduce un contexto profesional que puede favorecer respuestas más consistentes con la tarea de análisis de sentimientos. No obstante, su utilidad dependerá de que el rol no introduzca sesgos innecesarios ni desplace la atención hacia aspectos secundarios del texto.

5. *Iterative prompting*:

El *iterative prompting* es una estrategia multipaso en la que la clasificación se organiza en ciclos sucesivos de generación, revisión y refinamiento (Nádas et al., 2025). A diferencia de las técnicas directas, donde el modelo produce una respuesta en una única interacción, este enfoque incorpora una fase posterior de revisión en la que el *LLM* evalúa críticamente su propia salida y, si procede, ajusta la clasificación inicial. En el análisis de sentimientos, esta estrategia resulta especialmente pertinente cuando las reseñas contienen valoraciones mixtas, ironía, sarcasmo o ambigüedades semánticas, ya que una única respuesta puede no captar adecuadamente la polaridad global del texto (Miriayala et al., 2025).

La literatura ha documentado el potencial de este tipo de enfoques autorreflexivos. (Madaan et al., 2023) proponen el marco *Self-Refine*, en el que el modelo genera una respuesta inicial, produce retroalimentación sobre posibles errores y elabora una versión revisada de manera iterativa. Sus experimentos en tareas de clasificación, razonamiento y generación de texto muestran mejoras frente a estrategias de una sola pasada, especialmente en problemas que requieren análisis semántico o evaluación de información contradictoria. Trasladado al análisis de sentimientos, este procedimiento permite que el modelo reconsidere clasificaciones iniciales que podrían haber sido influenciadas por elementos aislados de la reseña, incorporando una evaluación más equilibrada de los aspectos positivos y negativos del texto.

En el presente trabajo, el *iterative prompting* se implementa en dos fases. En la primera, el modelo recibe la reseña y genera una clasificación inicial de sentimiento. En la segunda, se le presenta dicha clasificación junto con una instrucción explícita de revisión, solicitándole que reconsidere la respuesta teniendo en cuenta los matices positivos y negativos del texto. No obstante, para mantener la comparabilidad entre técnicas y facilitar la evaluación automática, la salida final se restringe exclusivamente a una de las tres etiquetas posibles: positivo, negativo o neutro.

Fase 1: “Clasifica el sentimiento de la siguiente reseña en una de estas categorías: positivo, negativo o neutro. Reseña: ‘El producto tiene un diseño bonito y llegó antes de lo esperado, pero se rompió a los tres días de uso y el servicio de atención al cliente no respondió a mi reclamación.’ Sentimiento:”

Fase 2: “Has clasificado la reseña anterior como [respuesta de la Fase 1]. Revisa tu clasificación teniendo en cuenta los siguientes aspectos: (1) ¿Has considerado todos los elementos positivos y negativos mencionados? (2) ¿Algún elemento tiene mayor peso emocional que los demás? (3) ¿La clasificación refleja fielmente la valoración global del consumidor? Responde únicamente con una de estas tres etiquetas: positivo, negativo o neutro. Sentimiento final:”

Este procedimiento permite incorporar una segunda instancia de evaluación sobre la respuesta generada, reduciendo la probabilidad de clasificaciones basadas en señales superficiales o en fragmentos aislados del texto. No obstante, su aplicación implica un mayor coste de inferencia, ya que requiere al menos dos interacciones con el modelo por cada reseña analizada. Por ello, en el diseño experimental se considerará tanto su posible mejora en calidad clasificatoria como el incremento asociado en tiempo de procesamiento y consumo computacional.

3.3. Evaluación del desempeño de los modelos

La evaluación del desempeño tiene como propósito valorar la capacidad de las distintas combinaciones de modelo y estrategia de *prompting* para clasificar el sentimiento expresado en reseñas de productos de comercio electrónico en español. A diferencia de la selección del modelo, centrada en criterios de eficiencia computacional, esta fase se orienta a analizar la calidad de las predicciones obtenidas. Para ello, se examina tanto el rendimiento global del sistema como su comportamiento por clase de sentimiento y por técnica de *prompting*, lo que permite identificar patrones de error y condiciones bajo las cuales la clasificación resulta menos precisa (Sokolova y Lapalme, 2009).

3.3.1. Matriz de confusión

La matriz de confusión se emplea como herramienta de análisis para examinar la relación entre las etiquetas reales del conjunto de evaluación y las predicciones generadas por el modelo (Fawcett, 2006). En el presente trabajo, la tarea corresponde a una clasificación de sentimientos en tres clases: positivo, negativo y neutro. Por ello, la matriz adopta una estructura 3×3 , en la que las filas representan la clase real de cada reseña y las columnas la clase predicha por el modelo. Esta representación permite observar no solo el número total de aciertos, sino también la dirección de los errores cometidos entre clases.

Tabla 3.2: Estructura de la matriz de confusión para clasificación ternaria de sentimientos.

Real \ Predicho	Positivo	Negativo	Neutro
Positivo	TP_{pos}	$E_{\text{pos} \rightarrow \text{neg}}$	$E_{\text{pos} \rightarrow \text{neu}}$
Negativo	$E_{\text{neg} \rightarrow \text{pos}}$	TP_{neg}	$E_{\text{neg} \rightarrow \text{neu}}$
Neutro	$E_{\text{neu} \rightarrow \text{pos}}$	$E_{\text{neu} \rightarrow \text{neg}}$	TP_{neu}

En la Tabla 3.2, los elementos situados en la diagonal principal, denotados como TP_i , representan las instancias de la clase i correctamente clasificadas por el modelo. Así, TP_{pos} corresponde a las reseñas positivas predichas como positivas, TP_{neg} a las reseñas negativas predichas como negativas y TP_{neu} a las reseñas neutras predichas como neutras. Por el contrario, los elementos fuera de la diagonal, denotados como $E_{i \rightarrow j}$, representan errores de clasificación, es decir, reseñas cuya clase real es i que han sido asignadas incorrectamente a la clase j . Por ejemplo, $E_{\text{pos} \rightarrow \text{neg}}$ indica el número de reseñas positivas clasificadas como negativas, mientras que $E_{\text{neu} \rightarrow \text{pos}}$ recoge las reseñas neutras clasificadas como positivas.

A partir de esta matriz se obtienen los valores necesarios para calcular las métricas de evaluación por clase. Para una clase determinada i , los verdaderos positivos (TP_i) corresponden a las instancias correctamente asignadas a dicha clase. Los falsos positivos (FP_i) representan las reseñas pertenecientes a otras clases que el modelo clasifica erróneamente como i , mientras que los falsos negativos (FN_i) recogen las instancias de la clase i que el modelo asigna a una categoría distinta. Finalmente, los verdaderos negativos (TN_i) corresponden a las observaciones que no pertenecen a la clase i y que tampoco han sido clasificadas como tal. Esta formulación permite analizar el rendimiento de cada categoría de sentimiento de manera independiente, aspecto especialmente relevante en contextos multiclase y con posible desbalance entre etiquetas.

El análisis de la matriz de confusión resulta especialmente informativo en tareas de análisis de sentimientos, ya que no todos los errores tienen la misma interpretación semántica. La confusión directa entre positivo y negativo, representada por $E_{\text{pos} \rightarrow \text{neg}}$ y $E_{\text{neg} \rightarrow \text{pos}}$, implica una

inversión completa de la polaridad y constituye el error más severo desde el punto de vista interpretativo. En cambio, la confusión entre la clase neutra y las clases polares suele asociarse a reseñas con valoraciones ambiguas, moderadas o mixtas, en las que coexisten elementos positivos y negativos (Liu, 2012). Asimismo, una frecuencia elevada de predicciones neutras en reseñas originalmente positivas o negativas puede indicar una tendencia conservadora del modelo, mientras que una sobreasignación de clases polares a reseñas neutras puede reflejar una sensibilidad excesiva ante expresiones aisladas de valoración.

3.3.2. Métricas de clasificación

A partir de la matriz de confusión se calculan las métricas empleadas para evaluar el rendimiento del sistema de clasificación multiclase. En este trabajo, cada clase de sentimiento se analiza de forma independiente, de modo que las categorías positivo, negativo y neutro se tratan sucesivamente como la clase de interés frente al resto de categorías. Esta formulación permite obtener medidas específicas por clase y resulta adecuada cuando el objetivo no es solo conocer el rendimiento global del modelo, sino también identificar diferencias en su comportamiento según la polaridad analizada (Sokolova y Lapalme, 2009). La primera métrica considerada es la exactitud o *accuracy*, que mide la proporción de instancias correctamente clasificadas sobre el total de observaciones evaluadas. Se define como:

$$\text{Accuracy} = \frac{\sum_{i=1}^C TP_i}{N} \quad (3.1)$$

donde C representa el número total de clases, TP_i corresponde al número de verdaderos positivos de la clase i y N indica el número total de reseñas evaluadas. En el presente caso, $C = 3$, dado que se distinguen las clases positivo, negativo y neutro. La exactitud proporciona una medida general del rendimiento del modelo; sin embargo, puede resultar insuficiente cuando existe desequilibrio entre clases, ya que un clasificador puede alcanzar un valor elevado si predice correctamente la clase mayoritaria, aun cuando su desempeño sea limitado en categorías menos frecuentes (Sokolova y Lapalme, 2009).

Para complementar esta visión global, se calcula la precisión o *precision* de cada clase i , definida como la proporción de predicciones asignadas a dicha clase que son realmente correctas:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (3.2)$$

En esta expresión, FP_i representa los falsos positivos de la clase i , es decir, las reseñas que pertenecen a otras categorías, pero que el modelo ha clasificado incorrectamente como i . En el contexto del análisis de sentimientos, una precisión baja para una clase indica que el modelo tiende a sobreasignar dicha etiqueta a reseñas que no pertenecen realmente a ella. Por ejemplo, una baja precisión en la clase neutra podría sugerir que el modelo clasifica como

neutras reseñas que expresan una polaridad positiva o negativa.

También se calcula la exhaustividad o *recall*, que mide la proporción de instancias reales de la clase i que el modelo identifica correctamente:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (3.3)$$

donde FN_i corresponde a los falsos negativos de la clase i , es decir, reseñas cuya etiqueta real es i , pero que han sido clasificadas como otra categoría. Esta métrica permite evaluar la capacidad del modelo para recuperar todos los casos pertenecientes a una clase determinada. En este estudio, un valor bajo de *recall* para la clase negativa, por ejemplo, indicaría que el modelo no detecta adecuadamente reseñas con insatisfacción, lo que puede tener implicaciones relevantes en aplicaciones empresariales orientadas a identificar problemas en la experiencia del cliente.

La puntuación F1 o *F1-score* combina precisión y exhaustividad mediante su media armónica:

$$F1_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (3.4)$$

Esta métrica resulta útil cuando se busca equilibrar la capacidad del modelo para evitar falsas asignaciones a una clase y, al mismo tiempo, recuperar correctamente las instancias reales de dicha categoría. Al utilizar la media armónica, la *F1-score* penaliza los casos en los que una de las dos métricas es elevada y la otra reducida, proporcionando así una medida más exigente del rendimiento por clase.

Dado que el corpus de evaluación puede presentar desequilibrio entre las clases positivo, negativo y neutro, se reporta adicionalmente la *F1 macro-promediada*. Esta métrica se calcula como la media aritmética de las puntuaciones F1 obtenidas para cada clase:

$$F1_{\text{macro}} = \frac{1}{C} \sum_{i=1}^C F1_i \quad (3.5)$$

donde $C = 3$ corresponde al número de clases evaluadas. A diferencia de métricas ponderadas por frecuencia, la *F1 macro-promediada* asigna el mismo peso a cada categoría, con independencia del número de observaciones que contenga. Por ello, se adopta como indicador principal de rendimiento en este trabajo, ya que permite valorar de forma equilibrada el desempeño del modelo en las tres clases de sentimiento y evita que los resultados queden dominados por la clase mayoritaria (Sokolova y Lapalme, 2009).

Todas las métricas descritas se calcularán de forma independiente para cada una de las estrategias de *prompting* evaluadas: *zero-shot*, *one-shot*, *few-shot*, *chain-of-thought*, *role-based* e *iterative prompting*. Esta comparación permitirá identificar qué técnica genera clasificaciones más precisas, así como determinar si existen diferencias en los tipos de error cometidos

por cada estrategia (Min et al., 2022). Además, permitirá valorar si las aproximaciones más complejas, como el *iterative prompting*, justifican su mayor coste computacional mediante una mejora suficiente del rendimiento clasificatorio.

3.3.3. Intervalos de confianza y significación estadística

Para evaluar la estabilidad de los resultados, se calcularán intervalos de confianza al 95 % para la *F1 macro-promediada* obtenida por cada combinación de modelo y estrategia de *prompting*. Para ello, se empleará el método de *bootstrap* con 1.000 remuestreos, una técnica de inferencia no paramétrica que permite estimar la distribución muestral de un estadístico mediante remuestreo con reemplazo a partir de los datos observados (Efron y Tibshirani, 1994). En cada iteración, se generará una muestra aleatoria del mismo tamaño que el conjunto de evaluación original, extrayendo observaciones con reemplazo, y se calculará la *F1 macro-promediada* sobre dicha muestra. Tras completar las 1.000 iteraciones, se obtendrá una distribución empírica del estadístico, cuyos percentiles 2,5 y 97,5 definirán los límites inferior y superior del intervalo de confianza al 95 %.

El uso de *bootstrap* resulta adecuado en este contexto porque métricas como la *F1-score* no suelen seguir una distribución normal, especialmente cuando el tamaño muestral es reducido o existe desequilibrio entre clases (Sokolova y Lapalme, 2009). A diferencia de los métodos paramétricos, este procedimiento no requiere asumir una forma específica para la distribución subyacente, lo que lo hace apropiado para evaluar métricas de clasificación en conjuntos de datos de tamaño limitado (Efron y Tibshirani, 1994). La interpretación de los intervalos se realizará considerando su grado de solapamiento. Cuando los intervalos de confianza de dos combinaciones se solapen ampliamente, no se afirmará que una estrategia es superior a otra con suficiente respaldo estadístico. En cambio, diferencias consistentes acompañadas de intervalos claramente separados se interpretarán como indicios de un mejor desempeño relativo.

Capítulo 4

Resultados

En este capítulo se presentan los resultados obtenidos a lo largo del proceso experimental. En primer lugar, se expone el procedimiento seguido para la selección del modelo de lenguaje empleado en el estudio, considerando los criterios de eficiencia computacional y calidad textual definidos en el Capítulo 3. A continuación, se analizan los resultados de clasificación de sentimientos obtenidos mediante las distintas estrategias de *prompting* evaluadas, así como las diferencias observadas en términos de rendimiento global y comportamiento por clase. Finalmente, se evalúan los patrones de error identificados a partir de las matrices de confusión y las métricas de evaluación utilizadas. Todos los experimentos fueron ejecutados en el entorno gratuito de *Google Colab* (Bisong, 2019), utilizando una GPU *NVIDIA Tesla T4* con 15 GB de VRAM y aproximadamente 12,7 GB de memoria RAM. Con el propósito de garantizar la transparencia y reproducibilidad del estudio, el código fuente desarrollado para la recopilación de resultados, la ejecución de los modelos, la implementación de las estrategias de *prompting* y el cálculo de las métricas de evaluación se encuentra disponible en un repositorio público de GitHub.¹

4.1. Caracterización de los *LLMs* cuantizados

La selección del modelo se realizó siguiendo los criterios metodológicos definidos en el Capítulo 3. En primer lugar, se establecieron tres condiciones de inclusión. El modelo debía disponer de pesos abiertos (*open-weight*), contar con una versión ajustada para instrucciones (*Instruct-tuned*) y poder ejecutarse en cuantización a 4 bits dentro del límite de 15 GB de VRAM disponible en el entorno experimental. Estos criterios excluyeron los modelos propietarios accesibles únicamente mediante API, como *GPT-4*, *Claude* o *Gemini*, debido a que no permiten despliegue local, limitan el control sobre los datos procesados, dificultan la medición directa de variables como consumo de memoria y latencia en el entorno experimental, y generan costes por uso que no se ajustan al objetivo de evaluar una solución reproducible

¹<https://github.com/123Iren/TFG-LLM-Sentiment-Analysis-Irene-Barrena->

en hardware convencional.

Tras esta primera delimitación, se acotó el tamaño de los modelos candidatos. El rango seleccionado se situó entre 7 y 8 mil millones de parámetros, al representar un equilibrio razonable entre capacidad semántica y viabilidad computacional. Los modelos de menor tamaño, como *Phi-2* de 2,7B o *Gemma 3* de 4B, fueron descartados por su menor capacidad esperada para capturar matices lingüísticos relevantes en tareas de análisis de sentimientos, especialmente en reseñas con opiniones mixtas, ironía o ambigüedad. En el extremo opuesto, modelos de mayor tamaño, como *Gemma 3* de 12B o *LLaMA 3.1* de 70B, resultaban poco adecuados para el entorno disponible, ya que su consumo de VRAM podía aproximarse o superar el límite operativo y reducir el margen necesario para la *KV cache* durante la inferencia. Por ello, el análisis se concentró en modelos de tamaño intermedio, suficientemente expresivos para la tarea y compatibles con las restricciones del entorno de ejecución.

Respecto al método de cuantización, se consideraron inicialmente técnicas de PTQ a 4 bits ampliamente utilizadas, como *AWQ* y *GPTQ*. Sin embargo, durante las pruebas preliminares en *Google Colab* se identificaron incompatibilidades entre algunas bibliotecas requeridas para la inferencia, como *autoawq* y *gptqmodel*, y dependencias empleadas para la carga y gestión de datos. Este tipo de conflictos ha sido señalado por Jin et al. (2024) como una de las dificultades prácticas asociadas al despliegue de modelos cuantizados en entornos reproducibles. Ante esta restricción, se optó por utilizar *bitsandbytes* (Dettmers, Lewis, Belkada, y Zettlemoyer, 2022), una alternativa compatible con *Transformers* de *Hugging Face* (Wolf, Debut, Sanh, et al., 2020) que permite aplicar cuantización a 4 bits directamente durante la carga del modelo. La elección de *bitsandbytes* permitió trabajar con las versiones oficiales de los repositorios de cada proveedor sin depender de variantes pre-cuantizadas. En particular, se empleó el formato NF4, diseñado para representar de forma eficiente pesos con distribuciones aproximadamente normales. Según Dettmers et al. (2023), esta representación permite reducir de manera sustancial el consumo de memoria manteniendo un rendimiento cercano al de los modelos en mayor precisión, siempre que la cuantización se aplique bajo una configuración adecuada. Esta decisión metodológica favoreció la reproducibilidad del experimento, redujo los problemas de compatibilidad del entorno y permitió comparar los modelos candidatos bajo una misma configuración de cuantización.

Tras aplicar los criterios anteriores, se seleccionaron tres modelos procedentes de proveedores distintos, con el fin de que la comparación no se limitara a variaciones de tamaño dentro de una misma familia, sino que incorporara diferencias arquitectónicas y de diseño relevantes. Como se muestra en la Tabla 4.1, los modelos candidatos fueron *LLaMA 3.1 Instruct*, *Qwen 2.5 Instruct* y *Mistral 7B Instruct*. La elección de estas alternativas responde a su disponibilidad como modelos de pesos abiertos, su ajuste instruccional y su tamaño compatible con una evaluación en entornos de recursos limitados. Además, cada uno de ellos presenta características diferenciadas que pueden influir en su comportamiento en tareas de análisis de sentimientos, como la orientación multilingüe, el uso de mecanismos de atención

eficientes o su posición como modelo de referencia dentro del ecosistema *open-weight*.

Modelo	Proveedor	Parámetros	Características principales
<i>LLaMA 3.1 Instruct</i>	<i>Meta AI</i>	8B	Referencia del ecosistema <i>open-weight</i> . Preentrenado con más de 15 billones de tokens multilingües. Incorpora GQA (Grattafiori, Dubey, Jauhri, et al., 2024).
<i>Qwen 2.5 Instruct</i>	<i>Alibaba Cloud</i>	7B	Buen rendimiento en <i>benchmarks</i> multilingües, especialmente relevante para textos en español (Yang, Yang, Zhang, et al., 2024).
<i>Mistral 7B Instruct</i>	<i>Mistral AI</i>	7B	Usa SWA y GQA. Supera a modelos más grandes en varias tareas (Jiang et al., 2023).

Tabla 4.1: Modelos candidatos seleccionados para la evaluación comparativa.

La evaluación empírica de los modelos candidatos se organizó en torno a cuatro dimensiones: tamaño en disco, consumo de memoria, latencia y perplejidad. El tamaño en disco se consideró como un criterio descriptivo, dado que permite estimar los requisitos de descarga y almacenamiento de cada modelo en su versión original, antes de aplicar la cuantización. Aunque esta dimensión no determina por sí sola la selección final, resulta útil para contextualizar la carga computacional de partida y las diferencias iniciales entre modelos de tamaño similar. La Tabla 4.2 presenta los tamaños observados para las versiones originales en *float16*.

Modelo	Tamaño en disco (GB)
<i>LLaMA 3.1 Instruct</i> (8B)	16,06
<i>Qwen 2.5 Instruct</i> (7B)	15,23
<i>Mistral 7B Instruct</i>	14,50

Tabla 4.2: Tamaño en disco de los modelos candidatos en su versión original (*float16*).

Como se observa en la Tabla 4.2, los tres modelos presentan tamaños en disco relativamente próximos, situados entre 14,50 GB y 16,06 GB en su versión original. *LLaMA 3.1 Instruct* es el modelo de mayor tamaño, coherente con sus 8 mil millones de parámetros, mientras que *Mistral 7B Instruct* presenta el menor tamaño de almacenamiento. Estas diferencias iniciales son relevantes para describir el punto de partida de la evaluación, aunque no constituyen un criterio suficiente para determinar la selección final. Por ello, el análisis se complementa con medidas de consumo de memoria durante la inferencia, latencia y perplejidad.

La cuantización a 4 bits permitió ejecutar los tres modelos dentro de las restricciones de memoria impuestas por la GPU *NVIDIA Tesla T4*. Como se observa en la Figura 4.1,

Mistral 7B presentó el menor consumo de memoria durante la inferencia, alcanzando un pico de 12,38 GB de VRAM. *Qwen 2.5 Instruct* registró un consumo máximo de 12,96 GB, mientras que *LLaMA 3.1 Instruct* alcanzó 13,91 GB, situándose próximo al límite operativo de 15 GB disponible en el entorno experimental. Aunque ninguno de los modelos superó dicho umbral, los resultados sugieren que *Mistral 7B* dispone de un margen operativo más amplio para procesar secuencias largas o ejecutar tareas adicionales sin riesgo de saturación de memoria.

Las diferencias observadas en latencia muestran un comportamiento más complejo. Si bien *Qwen 2.5 Instruct* y *LLaMA 3.1 Instruct* alcanzaron mayores velocidades de generación en términos de tokens por segundo, el indicador más relevante para la tarea planteada es el tiempo necesario para producir una clasificación completa por reseña. Bajo este criterio, *Mistral 7B* obtuvo el mejor resultado, con un tiempo medio de 0,564 segundos por reseña, frente a 2,053 segundos para *Qwen 2.5 Instruct* y 2,59 segundos para *LLaMA 3.1 Instruct*. Esta diferencia se explica por la longitud de las respuestas generadas. Dado que la tarea consiste en asignar una única etiqueta de sentimiento, los modelos que producen salidas más concisas reducen significativamente el tiempo total de inferencia.

La tercera dimensión analizada fue la perplejidad, calculada sobre el corpus *WikiText-2* (Merity et al., 2017). Los resultados muestran nuevamente una ventaja para *Mistral 7B*, que obtuvo un valor de 8,08, frente a 10,28 para *Qwen 2.5 Instruct* y 10,29 para *LLaMA 3.1 Instruct*. Dado que la perplejidad mide la capacidad del modelo para asignar probabilidad a secuencias lingüísticas observadas, valores más bajos indican una mejor capacidad predictiva y una menor degradación tras la cuantización. Los tres modelos se sitúan dentro de los rangos reportados en la literatura para arquitecturas cuantizadas a 4 bits (Frantar et al., 2022), lo que sugiere que la cuantización NF4 preservó adecuadamente sus capacidades lingüísticas. No obstante, la diferencia observada a favor de *Mistral 7B* apunta a una mayor robustez frente a la reducción de precisión numérica, posiblemente asociada a las optimizaciones arquitectónicas incorporadas en el modelo, entre ellas GQA y SWA (Jiang et al., 2023).

Los resultados de la Figura 4.1 muestran que *Mistral 7B* presenta el mejor comportamiento en las tres dimensiones principales evaluadas, al combinar el menor consumo de memoria, la menor latencia efectiva por reseña y la menor perplejidad. No obstante, dado que los tres modelos cumplieron los requisitos mínimos de viabilidad computacional y calidad lingüística, ninguno fue descartado en esta fase. Por tanto, los tres candidatos avanzaron a la etapa siguiente, en la que su desempeño se evaluó directamente en la tarea de clasificación de sentimientos mediante las distintas estrategias de *prompting*.

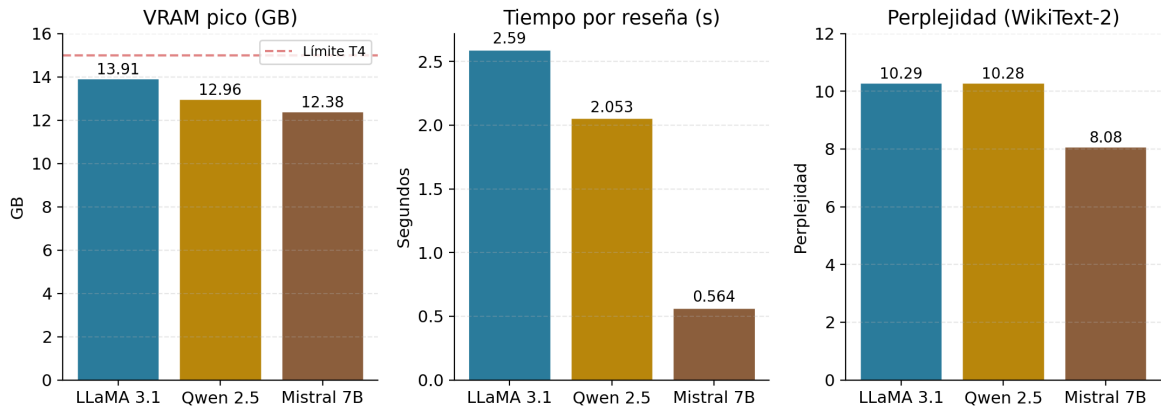


Figura 4.1: Comparativa de eficiencia de los tres modelos cuantizados. Mistral 7B destaca en las tres dimensiones, con el menor consumo de VRAM, la menor latencia por reseña y la perplejidad más baja.

4.2. Clasificación de sentimientos mediante técnicas de *prompting*

Tras la fase de selección de modelos, se evaluó el desempeño de las 18 combinaciones resultantes de cruzar los tres modelos candidatos con las seis estrategias de *prompting* consideradas. Todas las configuraciones se aplicaron sobre el mismo conjunto de evaluación y bajo idénticas condiciones de inferencia, permitiendo una comparación directa tanto del rendimiento clasificatorio como del coste computacional asociado a cada estrategia. El conjunto de evaluación se construyó a partir del *split* de prueba de *Amazon Reviews Multi* en español (Keung, Lu, Szarvas, y Smith, 2020). Las valoraciones originales se agruparon en tres categorías de sentimiento, asociando las puntuaciones de una y dos estrellas a la clase negativa, las de tres estrellas a la clase neutra y las de cuatro y cinco estrellas a la clase positiva (Liu, 2012). A partir de este corpus se seleccionaron 200 reseñas mediante muestreo aleatorio con semilla fija. La distribución obtenida estuvo compuesta por 91 reseñas negativas (45,5 %), 78 positivas (39,0 %) y 31 neutras (15,5 %). La menor representación de la clase neutra reproduce el patrón observado en el corpus original y proporciona un escenario de evaluación caracterizado por una distribución no uniforme entre categorías. El tamaño de la muestra respondió a las restricciones computacionales del entorno de ejecución, dado que cada combinación modelo-técnica requería inferencia secuencial sobre el conjunto completo de reseñas. Al mismo tiempo, se mantuvo un número suficiente de observaciones para comparar de forma homogénea las estrategias evaluadas bajo las mismas condiciones experimentales.

La Figura 4.2 presenta el tiempo total acumulado necesario para procesar las 200 reseñas mediante las seis estrategias de *prompting* en cada modelo. Se observan diferencias sustanciales entre arquitecturas. *Mistral 7B* completó el conjunto de experimentos en 2.129 segundos, mientras que *Qwen 2.5* requirió 5.202 segundos y *LLaMA 3.1* alcanzó 7.022 segundos. Estas

diferencias se mantuvieron de forma consistente a lo largo de las distintas estrategias evaluadas, reflejando el comportamiento observado previamente en las mediciones individuales de latencia.

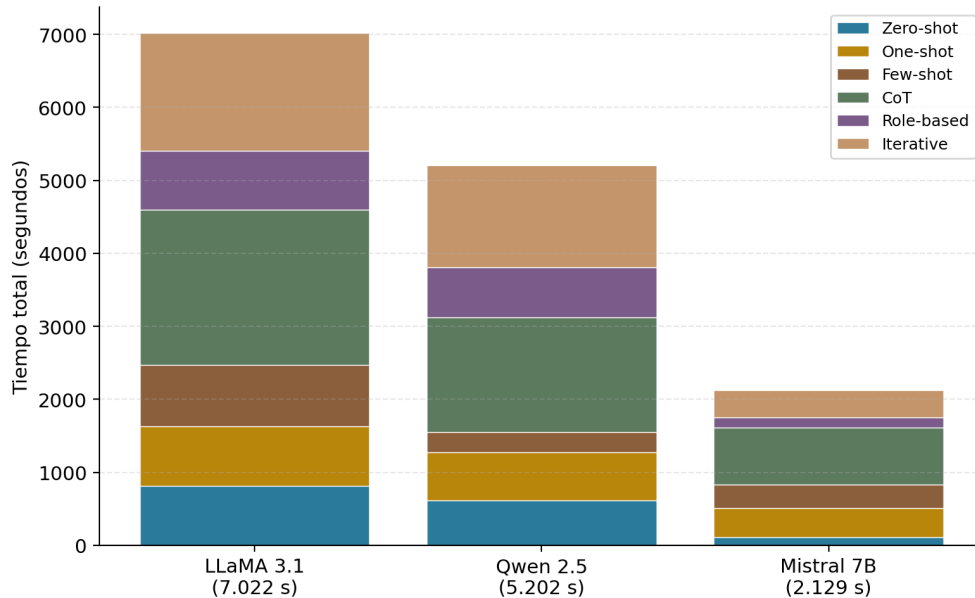


Figura 4.2: Tiempo de ejecución total acumulado por modelo. Mistral 7B completa las 6 técnicas en 2.129 s (35 min), frente a los 5.202 s de Qwen 2.5 (87 min) y los 7.022 s de LLaMA 3.1 (117 min). Las técnicas *CoT* e *iterative* concentran la mayor proporción del tiempo.

La distribución interna del tiempo de ejecución muestra que el coste computacional no depende únicamente del modelo, sino también del tipo de interacción que exige cada estrategia de *prompting*. En los tres modelos, las técnicas *Chain-of-Thought* e *iterative prompting* concentran la mayor proporción del tiempo total acumulado, lo que resulta coherente con su mayor demanda de generación. La primera induce respuestas más extensas al solicitar un razonamiento intermedio, mientras que la segunda requiere al menos dos llamadas de inferencia por reseña. En cambio, las estrategias directas, especialmente *zero-shot* y *one-shot*, presentan un coste temporal menor al limitar la salida esperada a una clasificación más breve. Este resultado sugiere que la complejidad del *prompt* introduce un coste adicional observable incluso bajo cuantización, por lo que la comparación entre técnicas debe considerar simultáneamente el rendimiento clasificatorio y el tiempo necesario para obtener cada predicción.

Antes de presentar las métricas de clasificación, se verificó la consistencia operativa de los *prompts* y del procedimiento de extracción de etiquetas. Todas las estrategias partieron de una instrucción común que especificaba la tarea, las tres categorías posibles y la restricción de responder con una única palabra. En las estrategias *one-shot* y *few-shot*, los ejemplos procedieron del mismo corpus, aunque fueron excluidos del conjunto de evaluación para evitar solapamiento entre demostraciones y observaciones evaluadas (Min et al., 2022). Las

predicciones se extrajeron mediante un *parser* diseñado para identificar la última aparición de las etiquetas *positivo*, *negativo* o *neutro* en la respuesta del modelo. Este procedimiento permitió clasificar 3.599 de las 3.600 inferencias generadas, con una única respuesta no identificable. El código completo utilizado para la generación de respuestas, extracción de etiquetas y cálculo de métricas se encuentra disponible en el repositorio público indicado al inicio del capítulo. Las Tablas 4.3 y 4.4 recogen los *prompts* utilizados en cada estrategia, ilustrados mediante una misma reseña del conjunto de evaluación. Esta presentación permite comparar la estructura de las instrucciones bajo condiciones textuales equivalentes. Las técnicas directas se diferencian por la ausencia o presencia de ejemplos, mientras que *chain-of-thought* incorpora una secuencia explícita de análisis, *role-based* añade un marco profesional de interpretación e *iterative prompting* introduce una segunda fase de revisión. Aunque en todos los casos se mantuvo la restricción de respuesta en una única palabra, la inclusión de ejemplos en *one-shot* y *few-shot* generó una estructura más propensa a continuaciones no solicitadas, un comportamiento relacionado con los efectos de demostración descritos por Min et al. (2022).

La Tabla 4.5 presenta las respuestas reales generadas por los tres modelos para la misma reseña negativa empleada en las tablas anteriores. En este ejemplo, todos identificaron correctamente la polaridad negativa, aunque con distinto grado de adherencia al formato solicitado. *Mistral 7B* produjo salidas más concisas en la mayoría de las estrategias, mientras que *LLaMA 3.1* y *Qwen 2.5* añadieron explicaciones, repeticiones de la etiqueta o ejemplos adicionales en varios casos. En particular, las estrategias *one-shot* y *few-shot* generaron continuaciones con reseñas ficticias en algunos modelos, lo que puede afectar la extracción automática si aparecen etiquetas posteriores no asociadas a la reseña original. Asimismo, el *role-based prompting* produjo en *LLaMA 3.1* una repetición de la etiqueta, compatible con artefactos de generación que pueden producirse en modelos cuantizados a baja precisión. Este comportamiento justifica el uso de un *parser* basado en la última aparición de una etiqueta válida y contribuye a interpretar las diferencias de latencia y rendimiento entre estrategias.

Técnica	Prompt real empleado	Justificación
Zero-shot	Clasifica el sentimiento de la siguiente reseña como positivo, negativo o neutro. Responde únicamente con una palabra. Reseña: mala calidad. Lo compramos hace pocos meses, se ha utilizado practicamente nada y ha dejado de funcionar sin mas, no se enciende. Sentimiento:	Línea base sin ejemplos.
One-shot	Clasifica el sentimiento de la siguiente reseña como positivo, negativo o neutro. Responde únicamente con una palabra. Ejemplo: Reseña: Buena compra. Está muy bien. No tiene un precio excesivo y te ofrece unas buenas prestaciones. Os lo recomiendo Sentimiento: positivo Reseña: mala calidad. Lo compramos hace pocos meses, se ha utilizado practicamente nada y ha dejado de funcionar sin mas, no se enciende. Sentimiento:	Un ejemplo real (clase positiva).
Few-shot	Clasifica el sentimiento de la siguiente reseña como positivo, negativo o neutro. Responde únicamente con una palabra. Ejemplo 1: Reseña: Buena compra. Está muy bien. No tiene un precio excesivo y te ofrece unas buenas prestaciones. Os lo recomiendo — Sentimiento: positivo Ejemplo 2: Reseña: no me llego. no me llego el articulo me lo mando por correos normal sin seguimiento y nunca me llego tota un desastre — Sentimiento: negativo Ejemplo 3: Reseña: Barandillas horizontales no funciona. No he conseguido que queden niveladas. Al tener barandilla con barras horizontales, el soporte no llega ha realizar su función. Aparte de eso, es lo que estaba buscando. Seguiré mirando algo que adapte mejor. — Sentimiento: neutro Reseña: mala calidad. Lo compramos hace pocos meses, se ha utilizado practicamente nada y ha dejado de funcionar sin mas, no se enciende. Sentimiento:	Tres ejemplos reales (uno por clase), excluidos de la evaluación.

Tabla 4.3: *Prompts* reales empleados (parte 1 de 2), ilustrados con la misma reseña del corpus.

Técnica	Prompt real empleado	Justificación
<i>Chain-of-thought</i>	Clasifica el sentimiento de la siguiente reseña como positivo, negativo o neutro. Piensa paso a paso: 1. Identifica las palabras clave que expresan sentimiento. 2. Determina si el tono general es positivo, negativo o neutro. 3. Responde con una única palabra: positivo, negativo o neutro. Reseña: mala calidad. Lo compramos hace pocos meses, se ha utilizado practicamente nada y ha dejado de funcionar sin mas, no se enciende. Análisis y sentimiento:	Razonamiento paso a paso (max_new_tokens=150).
<i>Role-based</i>	Eres un experto en análisis de sentimientos con amplia experiencia en reseñas de productos de comercio electrónico. Clasifica el sentimiento de la siguiente reseña como positivo, negativo o neutro. Responde únicamente con una palabra. Reseña: mala calidad. Lo compramos hace pocos meses, se ha utilizado practicamente nada y ha dejado de funcionar sin mas, no se enciende. Sentimiento:	Asigna un rol de dominio (Zheng et al., 2024).
<i>Iterative</i> (fase 1)	Clasifica el sentimiento de la siguiente reseña como positivo, negativo o neutro. Responde únicamente con una palabra. Reseña: mala calidad. Lo compramos hace pocos meses, se ha utilizado practicamente nada y ha dejado de funcionar sin mas, no se enciende. Sentimiento:	Fase inicial, idéntica a <i>zero-shot</i> .
<i>Iterative</i> (fase 2)	Has clasificado la siguiente reseña como “negativo”. Reseña: mala calidad. Lo compramos hace pocos meses, se ha utilizado practicamente nada y ha dejado de funcionar sin mas, no se enciende. Revisa tu clasificación. ¿Es correcta? Si no, corrígela. Responde únicamente con una palabra: positivo, negativo o neutro. Sentimiento final:	El modelo revisa su clasificación previa (Madaan et al., 2023).

Tabla 4.4: *Prompts* reales empleados (parte 2 de 2). Fase 2 de *iterative* con clasificación previa “negativo”.

Técnica	LLaMA 3.1 8B	Qwen 2.5 7B	Mistral 7B
Zero-shot	“negativo. La razón es que la reseña menciona que el producto dejó de funcionar sin más [...]”	“Negativo”	“negativo”
One-shot	“negativo \n\nReseña: Buen producto [...] Sentimiento: positivo [...]” (genera reseñas ficticias)	“negativo \n\nReseña: El producto es correcto [...] Sentimiento: neutro” (genera ejemplos adicionales)	“negativo”
Few-shot	“negativo \n\nReseña: muy buena compra [...]” (genera reseñas ficticias)	“negativo”	“negativo \n\nReseña: muy buena compra [...] Sentimiento: positivo” (genera ejemplos adicionales)
CoT	“negativo. ‘ha dejado de funcionar sin más’ y ‘no se enciende’ indican experiencia desagradable.”	“Negativo. ‘mala calidad’, ‘ha dejado de funcionar’ y ‘no se enciende’ indican tono negativo.”	“1. Palabras clave: mala calidad, dejado de funcionar, no se enciende. 2. Tono: negativo. 3. Respuesta: negativo.”
Role-based	“negativo. La respuesta final es negativo. [...]” (repite la clasificación)	“Negativo”	“negativo”
Iterative (fase 2)	“negativo. El producto no ha funcionado correctamente. La clasificación es correcta.”	“negativo. La reseña es claramente negativa [...]”	“negativo”

Tabla 4.5: Respuestas reales de cada modelo para la reseña “mala calidad. Lo compramos hace pocos meses [...]”. En rojo los comportamientos problemáticos. Etiqueta real: *negativo*.

4.2.1. Resultados globales

La Figura 4.3 y la Tabla 4.6 presentan el desempeño global de las 18 combinaciones evaluadas. La representación de la *F1 macro-promediada* con intervalos de confianza permite distinguir dos patrones principales. En primer lugar, la técnica *one-shot* obtuvo los valores más bajos en las tres arquitecturas, con puntuaciones comprendidas entre 0,3052 y 0,4779. En todos los casos, estos valores se situaron por debajo de los obtenidos mediante *zero-shot*, tanto en *F1 macro-promediada* como en *accuracy*. En este experimento, la inclusión de un único ejemplo no se asoció con una mejora del rendimiento y mostró un comportamiento compatible con los efectos de demostración descritos por Min et al. (2022). En segundo lugar, el comportamiento de *few-shot* fue marcadamente dependiente del modelo. En *Qwen 2.5*, esta estrategia alcanzó el valor más alto del experimento, con una *F1 macro-promediada* de 0,7216 y una *accuracy* de 0,8400. En cambio, en *LLaMA 3.1* obtuvo una *F1 macro-promediada* de 0,3285 y una *accuracy* de 0,3350, valores próximos a los observados con *one-shot*. *Mistral 7B* ocupó una posición intermedia, con una *F1 macro-promediada* de 0,6335 y una *accuracy* de 0,7050. Esta dispersión indica que la incorporación de ejemplos en contexto no produjo un rendimiento homogéneo entre arquitecturas.

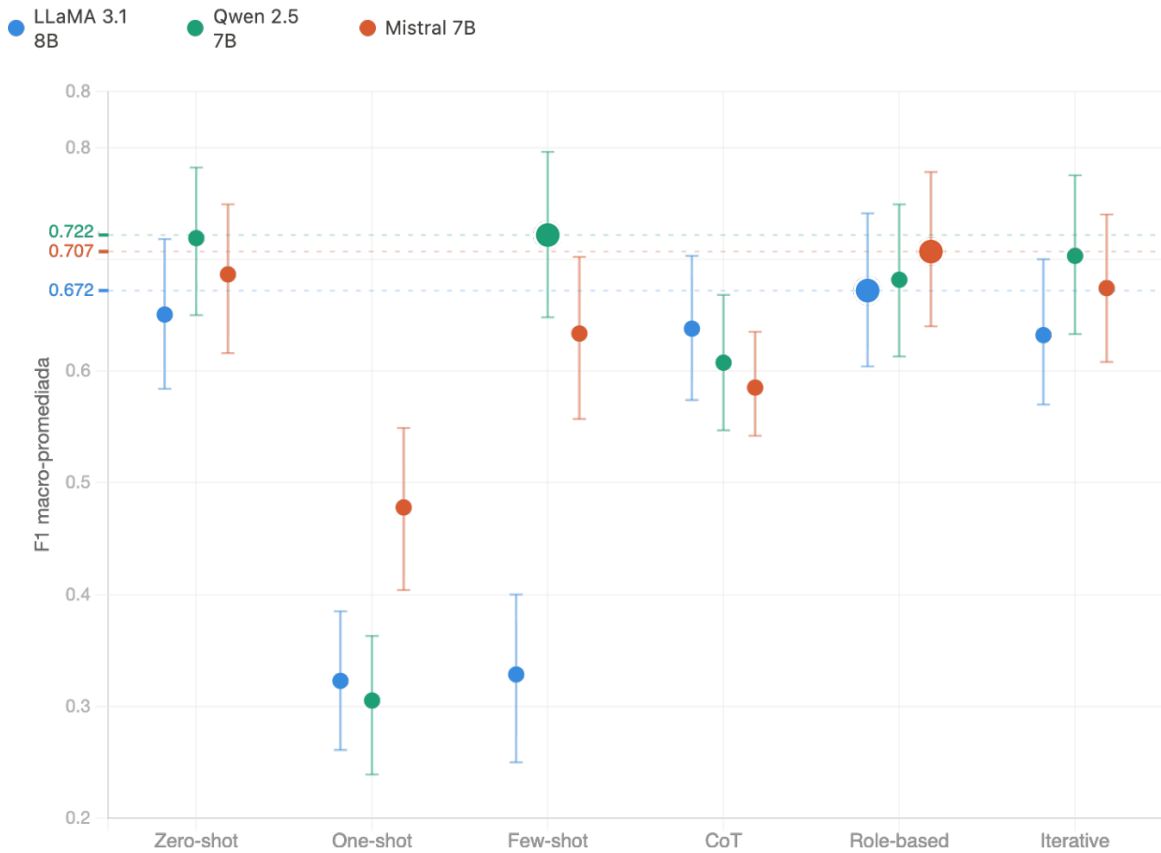


Figura 4.3: *F1 macro-promediada* por modelo y técnica.

Técnica	LLaMA 3.1 8B	Qwen 2.5 7B	Mistral 7B
Zero-shot	0,7150	0,8000	0,7950
One-shot	0,3200	0,3300	0,5650
Few-shot	0,3350	0,8400	0,7050
CoT	0,6800	0,7350	0,7750
Role-based	0,7850	0,7800	0,8100
Iterative	0,7050	0,7850	0,7900

Tabla 4.6: *Accuracy* por modelo y técnica, sobre 200 reseñas.

La estrategia *role-based* presentó el comportamiento más estable entre modelos, con valores de *F1 macro-promediada* entre 0,6719 y 0,7068. Además, obtuvo el mejor resultado dentro de *LLaMA 3.1* y *Mistral 7B*, junto con valores de *accuracy* elevados en ambas arquitecturas. Frente a la variabilidad observada en *few-shot*, esta técnica mantuvo un rango de rendimiento más estrecho, lo que sugiere una menor sensibilidad al modelo empleado. Si bien Zheng et al. (2024) no observan mejoras consistentes asociadas a la asignación de roles en tareas generales, en este experimento la técnica mostró el comportamiento más estable, aunque los datos no permiten atribuir dicho comportamiento exclusivamente a la asignación del rol. Por su parte, las estrategias *chain-of-thought* e *iterative prompting* no mostraron una ventaja sistemática frente a las técnicas directas. *Zero-shot* alcanzó valores de *F1 macro-promediada* entre 0,6504 y 0,7188, superiores a los de *chain-of-thought* en los tres modelos, pese al mayor tiempo de ejecución de esta última estrategia. Por su parte, *iterative prompting* obtuvo valores entre 0,6321 y 0,7029, sin superar de forma consistente a *zero-shot* o *role-based*. En los datos analizados, una mayor complejidad de la instrucción no se tradujo necesariamente en mejores métricas globales de clasificación. Las cinco combinaciones con mayor *F1 macro-promediada* fueron *Qwen 2.5 few-shot* (0,7216), *Qwen 2.5 zero-shot* (0,7188), *Mistral 7B role-based* (0,7068), *Qwen 2.5 iterative* (0,7029) y *Mistral 7B zero-shot* (0,6864). Sin embargo, la Figura 4.3 muestra un solapamiento amplio entre sus intervalos de confianza. Por tanto, no se dispone de evidencia estadística suficiente para afirmar la superioridad de una de estas configuraciones sobre las demás al nivel de confianza considerado.

Para precisar el origen de las diferencias observadas en las métricas globales, se analizó el rendimiento de cada combinación por clase de sentimiento. La Tabla 4.7 desagrega la *F1-score* para las categorías positiva, negativa y neutra en las 18 configuraciones evaluadas. Los resultados muestran una asimetría persistente entre clases. Las categorías positiva y negativa alcanzan valores superiores a 0,80 en varias de las combinaciones con mejor desempeño, mientras que la clase neutra se mantiene por debajo de 0,42 en todos los modelos y estrategias. Este patrón aparece de forma transversal, por lo que no se restringe a una arquitectura ni a una técnica de *prompting* específica.

Técnica	LLaMA 3.1 8B			Qwen 2.5 7B			Mistral 7B		
	Pos.	Neg.	Neu.	Pos.	Neg.	Neu.	Pos.	Neg.	Neu.
Zero-shot	0,711	0,860	0,380	0,867	0,877	0,413	0,829	0,897	0,333
One-shot	0,264	0,349	0,355	0,233	0,400	0,283	0,495	0,684	0,254
Few-shot	0,390	0,269	0,327	0,897	0,896	0,372	0,737	0,756	0,408
CoT	0,766	0,774	0,374	0,803	0,831	0,189	0,844	0,860	0,051
Role-based	0,850	0,875	0,291	0,818	0,888	0,339	0,859	0,904	0,357
Iterative	0,713	0,850	0,333	0,829	0,859	0,421	0,824	0,897	0,302

Tabla 4.7: *F1-score* por clase para cada combinación.

La diferencia entre clases ayuda a interpretar la distancia entre *accuracy* y *F1 macro-promediada* observada anteriormente. En varias configuraciones, los valores de *accuracy* son relativamente elevados debido al buen rendimiento en las clases mayoritarias, especialmente negativa y positiva. Sin embargo, la *F1 macro-promediada* reduce esa ventaja al asignar el mismo peso a las tres categorías. Por ello, el bajo rendimiento en la clase neutra afecta de manera visible al indicador macro, incluso en combinaciones con buenos resultados agregados en términos de exactitud. La clase neutra presenta los valores más bajos en todas las configuraciones. Esta diferencia puede relacionarse con dos características observables del diseño experimental. En primer lugar, la clase neutra cuenta con menor número de instancias en el conjunto de evaluación, con 31 reseñas frente a 91 negativas y 78 positivas. En segundo lugar, las reseñas neutras suelen incluir valoraciones menos polarizadas o combinaciones de elementos positivos y negativos, lo que reduce la separación semántica respecto a las clases polares (Liu, 2012). Los resultados son consistentes con lo reportado por Ponnusamy et al. (2025), quienes identifican la clasificación de sentimientos neutros como una fuente recurrente de dificultad en modelos generativos aplicados a opiniones de consumidores. En este experimento, el bajo rendimiento de la clase neutra se mantiene incluso en las combinaciones con mayor *F1 macro-promediada*, lo que indica que el buen desempeño global se concentra principalmente en la detección de polaridades positivas y negativas.

Con el fin de evaluar la naturaleza de los errores cometidos por las configuraciones con mejor desempeño global, se analizaron las matrices de confusión correspondientes a *Qwen 2.5 few-shot*, *Qwen 2.5 zero-shot* y *Mistral 7B role-based* (Tablas 4.8, 4.9 y 4.10). En los tres casos, las confusiones directas entre las clases positiva y negativa fueron reducidas. *Qwen 2.5 few-shot* registró cinco inversiones de polaridad, *Qwen 2.5 zero-shot* tres y *Mistral 7B role-based* únicamente dos. La mayor parte de los errores se concentró en la clase neutra. Mientras que las categorías positiva y negativa alcanzaron entre 65 y 86 clasificaciones correctas sobre un total de 78 y 91 observaciones, respectivamente, la clase neutra obtuvo entre 8 y 13 aciertos sobre 31 instancias. Además, las observaciones neutras mal clasificadas se distribuyeron entre las clases positiva y negativa sin una dirección predominante, lo que indica que la pérdida de rendimiento no se debe a una tendencia sistemática hacia una polaridad concreta, sino a la dificultad de separar las valoraciones intermedias de las opiniones claramente favorables o desfavorables.

Real \ Predicho	Positivo	Negativo	Neutro
Positivo	74	3	1
Negativo	2	86	3
Neutro	11	12	8

Tabla 4.8: Matriz de confusión: *Qwen 2.5* con *few-shot* (F1-macro = 0,7216).

La Tabla 4.11 confirma este comportamiento. En las clases positiva y negativa, los valores de *precision* y *recall* se mantuvieron elevados y relativamente equilibrados en las tres

Real \ Predicho	Positivo	Negativo	Neutro
Positivo	65	2	11
Negativo	1	82	8
Neutro	6	12	13

Tabla 4.9: Matriz de confusión: *Qwen 2.5* con *zero-shot* (F1-macro = 0,7188).

Real \ Predicho	Positivo	Negativo	Neutro
Positivo	67	2	9
Negativo	0	85	6
Neutro	11	10	10

Tabla 4.10: Matriz de confusión: *Mistral 7B* con *role-based* (F1-macro = 0,7068).

configuraciones. Por el contrario, la clase neutra presentó un patrón más heterogéneo. *Qwen 2.5 few-shot* alcanzó la mayor *precision* para esta categoría (0,667), pero también el menor *recall* (0,258), lo que refleja una identificación muy selectiva de las observaciones neutras. En el extremo opuesto, *Qwen 2.5 zero-shot* obtuvo el mayor *recall* (0,419), acompañado de una *precision* inferior (0,406), capturando una proporción más amplia de reseñas neutras a costa de incrementar las asignaciones incorrectas. *Mistral 7B role-based* mostró valores intermedios en ambas métricas, con un comportamiento más equilibrado entre capacidad de detección y fiabilidad de las predicciones.

Combinación	Positivo		Negativo		Neutro	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
<i>Qwen 2.5 few-shot</i>	0,851	0,949	0,852	0,945	0,667	0,258
<i>Qwen 2.5 zero-shot</i>	0,903	0,833	0,854	0,901	0,406	0,419
<i>Mistral 7B role-based</i>	0,859	0,859	0,876	0,934	0,400	0,323

Tabla 4.11: *Precision* y *recall* por clase para las tres mejores combinaciones. En negrita se marca el mejor valor de cada métrica por clase.

El desglose de la clase neutra muestra diferencias relevantes entre *precision* y *recall*. *Qwen 2.5 few-shot* obtuvo la mayor *precision* para esta categoría (0,667), pero el menor *recall* entre las tres combinaciones (0,258), lo que indica que sus predicciones neutras fueron más selectivas, aunque dejó sin identificar una proporción elevada de reseñas realmente neutras. *Qwen 2.5 zero-shot* presentó el patrón contrario, con el mayor *recall* para la clase neutra (0,419) y una *precision* menor (0,406), por lo que recuperó más instancias neutras, aunque con mayor número de asignaciones incorrectas. *Mistral 7B role-based* se situó en una posición intermedia en *recall* y mostró una *precision* similar a la de *Qwen 2.5 zero-shot*. Estos resultados indican que el error en la clase neutra no adopta el mismo perfil en todas las configuraciones. El comportamiento de *chain-of-thought* refuerza la dificultad observada en la clase neutra. En particular, *Mistral 7B* con esta estrategia obtuvo una *F1-score* de 0,051 para

dicha categoría, correspondiente a un único acierto sobre 31 instancias neutras. Este resultado contrasta con sus valores más elevados en las clases positiva y negativa, por lo que la pérdida de rendimiento se concentró en la identificación de valoraciones intermedias. En los datos analizados, la incorporación de razonamiento explícito no mejoró la detección de neutralidad y se asoció con una mayor asignación hacia clases polares.

En el caso de *LLaMA 3.1*, las estrategias *one-shot* y *few-shot* mostraron una distribución del error distinta a la observada en *zero-shot* y *role-based*. Estas dos estrategias acumularon 67 y 100 inversiones de polaridad, respectivamente, sobre las 169 instancias polares del conjunto de evaluación, frente a un rango de 4 a 8 inversiones en las técnicas directas. Este patrón es coherente con las salidas cualitativas presentadas en la Tabla 4.5, donde las estrategias basadas en ejemplos generaron continuaciones no solicitadas en algunos casos. Por tanto, en esta arquitectura, la incorporación de ejemplos no solo redujo el rendimiento agregado, sino que también alteró la separación entre polaridades opuestas.

La comparación conjunta de las matrices de confusión, las métricas por clase y los intervalos de confianza permite matizar la selección de la configuración más adecuada. Aunque *Qwen 2.5 few-shot* obtuvo la mayor *F1 macro-promediada*, su ventaja frente a *Qwen 2.5 zero-shot* fue muy reducida y no estuvo acompañada de una mejora clara en la clase neutra. De hecho, *Qwen 2.5 zero-shot* presentó un número menor de inversiones de polaridad y el mayor *recall* para la categoría neutra entre las tres mejores combinaciones. Por ello, si se prioriza un equilibrio entre rendimiento global, simplicidad del *prompt*, estabilidad del formato de salida y cobertura de la clase neutra, *Qwen 2.5 zero-shot* constituye la alternativa más defendible dentro del experimento. En cambio, si el criterio principal es reducir al mínimo las inversiones entre polaridades opuestas, *Mistral 7B role-based* ofrece un perfil competitivo, aunque con una *F1 macro-promediada* ligeramente inferior.

Capítulo 5

Conclusiones

La creciente importancia de las percepciones del consumidor en la toma de decisiones empresariales ha convertido el análisis de sentimientos en una herramienta relevante para organizaciones que buscan diferenciarse mediante la comprensión de las opiniones de sus clientes. Sin embargo, la implementación práctica de sistemas de análisis de sentimientos basados en *LLMs* ha estado condicionada por sus elevados requerimientos computacionales, la dependencia de APIs propietarias con costes por uso y la necesidad de conjuntos etiquetados para el ajuste fino del modelo. Estas barreras limitan la adopción de este tipo de soluciones en entornos académicos y en organizaciones con infraestructura limitada, a pesar de que la evidencia disponible muestra que los *LLMs* ofrecen una mayor capacidad contextual que los enfoques tradicionales basados en diccionarios o modelos supervisados clásicos.

En este contexto, la revisión del estado del arte permitió identificar una evolución metodológica progresiva en el análisis de sentimientos. Los enfoques léxicos destacan por su interpretabilidad y bajo coste computacional, pero presentan dificultades para capturar matices contextuales, ambigüedad, ironía o variaciones propias de cada dominio. Los modelos supervisados y de aprendizaje profundo mejoran la capacidad de clasificación al aprender patrones directamente a partir de los datos, aunque requieren conjuntos etiquetados, mayor preprocesamiento y recursos computacionales más exigentes. Posteriormente, las arquitecturas *transformer* y los *LLMs* ampliaron la capacidad de comprensión contextual del lenguaje, mientras que las estrategias de *prompting* ofrecieron una vía flexible para adaptar el comportamiento del modelo sin modificar sus parámetros internos. Finalmente, los modelos cuantizados surgieron como una alternativa orientada a reducir las barreras de despliegue, al disminuir el consumo de memoria y facilitar la ejecución de *LLMs* en entornos con recursos limitados.

A partir de esta evolución, el presente Trabajo de Fin de Grado tuvo como objetivo evaluar el uso de *LLMs* cuantizados en tareas de análisis de sentimientos aplicadas a reseñas de productos en español, atendiendo tanto a la calidad de las predicciones como a la viabilidad de su ejecución en entornos con recursos computacionales limitados. Para ello, se

diseñó una metodología estructurada en tres etapas. En primer lugar, se seleccionó un corpus público de reseñas de comercio electrónico en español y se adaptaron sus etiquetas a una estructura de polaridad formada por las clases positivo, negativo y neutro. En segundo lugar, se caracterizaron distintos *LLMs* cuantizados a 4 bits, considerando criterios de eficiencia computacional, latencia y perplejidad, y se evaluaron distintas estrategias de *prompting* sin recurrir al ajuste fino del modelo. Finalmente, se analizó el desempeño de cada combinación modelo-técnica mediante matrices de confusión, métricas de clasificación multiclase e intervalos de confianza obtenidos mediante *bootstrap*. De este modo, el trabajo permitió valorar si los *LLMs* cuantizados pueden constituir una alternativa eficiente, reproducible y técnicamente accesible para el análisis de percepciones de clientes.

En relación con los resultados principales, las mejores combinaciones alcanzaron valores de *F1 macro-promediada* superiores a 0,70 y valores de *accuracy* entre 0,80 y 0,84. Estos resultados muestran que los *LLMs* cuantizados a 4 bits pueden ofrecer un rendimiento competitivo en la clasificación de sentimientos sin recurrir a ajuste fino, APIs propietarias ni infraestructura especializada. Además, los valores obtenidos se sitúan en un rango comparable al reportado por estudios recientes con modelos no cuantizados y por encima de algunos resultados documentados para enfoques léxicos tradicionales, aunque esta comparación debe interpretarse con cautela debido a las diferencias entre corpus, idiomas y diseños experimentales. El análisis comparativo permitió identificar perfiles diferenciados entre modelos y estrategias de *prompting*. *Mistral 7B* presentó el mejor comportamiento computacional, con menor consumo de VRAM, menor latencia por reseña y menor perplejidad, lo que lo convierte en una opción especialmente atractiva cuando se prioriza la eficiencia operativa. *Qwen 2.5*, por su parte, alcanzó los mejores valores clasificatorios, especialmente bajo las estrategias *few-shot* y *zero-shot*. Sin embargo, dado que *Qwen 2.5 few-shot* no mostró una ventaja estadísticamente clara frente a *Qwen 2.5 zero-shot*, y considerando que esta última estrategia presentó menor complejidad, menos inversiones de polaridad y mayor recuperación de la clase neutra, *Qwen 2.5 zero-shot* se posiciona como la alternativa más equilibrada dentro del experimento.

En cuanto a las técnicas de *prompting*, los resultados muestran que una mayor complejidad instruccional no implicó necesariamente un mejor desempeño. *One-shot* degradó el rendimiento en los tres modelos, mientras que *few-shot* presentó una alta sensibilidad a la arquitectura empleada. *Role-based prompting* destacó por su estabilidad entre modelos, aunque sin superar de forma estadísticamente diferenciada a las mejores configuraciones. Por el contrario, *chain-of-thought* e *iterative prompting* no aportaron mejoras sistemáticas y aumentaron el coste temporal de inferencia. Estos resultados sugieren que, en modelos cuantizados a 4 bits, las estrategias simples y directas pueden ofrecer un equilibrio más favorable entre rendimiento, coste computacional y estabilidad de salida. La principal dificultad empírica se concentró en la clase neutra, cuyo *F1-score* fue claramente inferior al observado en las clases positiva y negativa. Este patrón coincide con la literatura previa sobre análisis de

sentimientos, donde la neutralidad suele ser una categoría problemática por su menor separación semántica respecto a las clases polares y por la presencia de valoraciones mixtas o ambiguas. No obstante, las inversiones directas entre polaridades opuestas fueron escasas en las mejores configuraciones, lo que resulta relevante desde una perspectiva empresarial, ya que indica que el sistema distingue con mayor fiabilidad las opiniones claramente favorables y desfavorables que las percepciones intermedias.

El estudio presenta varias limitaciones que deben considerarse al interpretar los resultados. En primer lugar, el conjunto de evaluación se limitó a 200 reseñas, debido a las restricciones computacionales del entorno gratuito de *Google Colab*. Aunque este tamaño permitió comparar todas las combinaciones bajo condiciones homogéneas, una muestra más amplia permitiría obtener intervalos de confianza más estrechos y estimaciones más estables, especialmente para la clase neutra, que contó con menor representación. En segundo lugar, no se evaluó de forma específica el comportamiento de los modelos ante fenómenos lingüísticos complejos como ironía, sarcasmo o polaridad mixta, cuya identificación requeriría un corpus de mayor tamaño y anotación manual especializada. Asimismo, la cuantización a 4 bits pudo haber afectado algunas capacidades de generación o razonamiento secuencial, aspecto especialmente relevante para interpretar el bajo rendimiento observado en *chain-of-thought*. También debe señalarse que los *prompts* se formularon en texto plano, sin emplear los *chat templates* específicos de cada arquitectura, lo que podría haber limitado el aprovechamiento de formatos instruccionales propios de cada modelo. Finalmente, la evaluación se realizó exclusivamente sobre reseñas de comercio electrónico en español, por lo que los patrones observados no pueden generalizarse directamente a otros idiomas, dominios o tipos de texto.

Estas limitaciones abren varias líneas de trabajo futuro. Una primera extensión consistiría en ampliar el corpus de evaluación y equilibrar mejor la representación de las clases, con el fin de analizar de forma más precisa el comportamiento ante opiniones neutras o ambiguas. También resultaría pertinente evaluar los modelos utilizando los *chat templates* oficiales de cada arquitectura, así como comparar los resultados con modelos supervisados ajustados específicamente para español, como variantes de *BERT* o *RoBERTa*. De igual forma, la comparación entre cuantización a 4 y 8 bits permitiría analizar si una mayor precisión numérica mejora el rendimiento de estrategias que requieren razonamiento explícito, como *chain-of-thought*. Finalmente, la aplicación de la metodología a otros idiomas, sectores y tipos de reseñas permitiría valorar la robustez de los patrones identificados, en particular la estabilidad del *role-based prompting* y la sensibilidad de algunos modelos a las demostraciones incluidas en contexto.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que *ChatGPT* u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Irene Barrena, estudiante de Administración y Dirección de Empresas de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado “Análisis de percepciones de clientes con modelos de lenguaje de gran escala (LLMs) cuantizados”, declaro que he utilizado la herramienta de Inteligencia Artificial Generativa *ChatGPT* u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. Brainstorming de ideas de investigación: Utilizado para idear y esbozar posibles áreas de investigación.
2. Crítico: Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. Referencias: Usado conjuntamente con otras herramientas para identificar referencias preliminares que luego he contrastado y validado.
4. Metodólogo: Para descubrir métodos aplicables a problemas específicos de investigación.
5. Interpretador de código: Para realizar análisis de datos preliminares.
6. Constructor de plantillas: Para diseñar formatos específicos para secciones del trabajo.
7. Corrector de estilo literario y de lenguaje: Para mejorar la calidad lingüística y estilística del texto.
8. Generador previo de diagramas de flujo y contenido: Para esbozar diagramas iniciales.
9. Sintetizador y divulgador de libros complicados: Para resumir y comprender literatura

compleja.

10. Revisor: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
11. Traductor: Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado *ChatGPT* u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: Junio 2026

Firma: Irene Barrena

Referencias

- Abraham, M., Geng, T., Kogler, F., y Taylor, L. (2024). *What consumers want from personalization*. Descargado de <https://www.bcg.com/publications/2024/what-consumers-want-from-personalization> (Accessed: 2025-10-06)
- Aguilar-Moreno, J. A., Palos-Sanchez, P. R., y Pozo-Barajas, R. d. (2024). Sentiment analysis to support business decision-making. a bibliometric study. *AIMS Mathematics*, 9(2), 4337–4375.
- Ain, Q. T., Ali, M., Riaz, A., Noureen, A., Kamran, M., Hayat, B., y Rehman, A.-u. (2017). Sentiment analysis using deep learning techniques: a review. *International Journal of Advanced Computer Science and Applications*, 8(6).
- Arora, N., Ensslen, D., Fiedler, L., Liu, W. W., Robinson, K., Stein, E., y Schöler, G. (2021). *The value of getting personalization right—or wrong—is multiplying*. Descargado de <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying> (Accessed: 2025-10-06)
- Bai, J., y cols. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*. Descargado de <https://arxiv.org/abs/2309.16609>
- Bisong, E. (2019). Google colaboratory. En *Building machine learning and deep learning models on google cloud platform* (pp. 59–64). Apress.
- Brown, P. F., Della Pietra, V. J., Della Pietra, S. A., Lai, J. C., y Mercer, R. L. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1), 31–40. Descargado de <https://aclanthology.org/J92-1002>
- Chakraborty, M., Kulkarni, A., y Li, Q. (2023). Zero-shot approach to overcome perturbation sensitivity of prompts. *arXiv preprint arXiv:2305.15689*.
- Cho, H., Kim, S., Lee, J., y Lee, J.-S. (2014). Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems*, 71, 61–71.
- Christensen, C. M., Raynor, M. E., y McDonald, R. (2015). What is disruptive innovation? *Harvard Business Review*, 93(12), 44–53. Descargado de <https://hbr.org/2015/12/what-is-disruptive-innovation> (Accessed: 2025-10-06)
- Dettmers, T., Lewis, M., Belkada, Y., y Zettlemoyer, L. (2022). GPT3.int8(): 8-bit Matrix

- Multiplication for Transformers at Scale. En *Advances in neural information processing systems (neurips)* (Vol. 35).
- Dettmers, T., y cols. (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*. Descargado de <https://arxiv.org/abs/2305.14314>
- Diep, A. (2025). Self-transparency failures in expert-persona llms: How instruction-following overrides disclosure. *arXiv preprint arXiv:2511.21569*.
- Efron, B., y Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman & Hall/CRC. doi: 10.1201/9780429246593
- Fatemi, S., y Hu, Y. (2023). A comparative analysis of fine-tuned llms and few-shot learning of llms for financial sentiment analysis. *arXiv preprint arXiv:2312.08725*.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. doi: 10.1016/j.patrec.2005.10.010
- Frantar, E., y cols. (2022). Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*. Descargado de <https://arxiv.org/abs/2210.17323>
- Ghatora, P. S., Hosseini, S. E., Pervez, S., Iqbal, M. J., y Shaukat, N. (2024). Sentiment analysis of product reviews using machine learning and pre-trained llm. *Big Data and Cognitive Computing*, 8(12), 199.
- González San José, P., y cols. (2025). Análisis de flexibilidad de los modelos llm cuantizados.
- Grattafiori, A., Dubey, A., Jauhri, A., y cols. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Gunasekar, S., y cols. (2023). Textbooks are all you need. *arXiv preprint arXiv:2306.11644*. Descargado de <https://arxiv.org/abs/2306.11644>
- Hermawan, D., y cols. (2024). Financial sentiment analysis using quantized large language models: a comparative study. *Journal of Financial Data Science*. (Preprint)
- Hernández, S. D. R. P. (2020). La fidelización del cliente y retención del cliente: Tendencia que se exige hoy en día/customer loyalty and customer retention: Trend required today. *Gestión en el tercer milenio*, 23(45), 5–14.
- Jelinek, F., Mercer, R. L., Bahl, L. R., y Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1), S63. Descargado de <https://doi.org/10.1121/1.2016299> doi: 10.1121/1.2016299
- Jiang, A. Q., y cols. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*. Descargado de <https://arxiv.org/abs/2310.06825>
- Jin, Z., y cols. (2024). A comprehensive evaluation of quantized instruction-tuned large language models. *arXiv preprint arXiv:2408.09126*. Descargado de <https://arxiv.org/abs/2408.09126>
- Jirasky, P., y cols. (2024). Quantization of large language models: A survey. *arXiv preprint*.
- Keung, P., Lu, Y., Szarvas, G., y Smith, N. A. (2020). The multilingual amazon reviews cor-

- pus. En *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 4563–4568).
- Li, C., Gao, F., Bu, J., Xu, L., Chen, X., Gu, Y., ... others (2021). Sentiprompt: Sentiment knowledge enhanced prompt-tuning for aspect-based sentiment analysis. *arXiv preprint arXiv:2109.08306*.
- Li, S., Ning, X., Wang, L., Liu, T., Shi, X., Yan, S., ... Wang, Y. (2024). Evaluating quantized large language models. *arXiv preprint arXiv:2402.18158*.
- Li, T., Li, H., y Zhou, Y. (2025). E-commerce sentiment analysis using fine-tuned llama3 models: A qlora-based approach. *Journal of Technology Innovation and Engineering*, 1(4).
- Lin, J., y cols. (2024). Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*. Descargado de <https://arxiv.org/abs/2306.00978>
- Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers. doi: 10.2200/S00416ED1V01Y201204HLT016
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., ... Hajishirzi, H. (2023). Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*. Descargado de <https://arxiv.org/abs/2303.17651>
- Merity, S., Xiong, C., Bradbury, J., y Socher, R. (2017). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., y Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? En *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 11048–11064). Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759
- Miriyala, V., Bukkapatnam, S., y Prahallad, L. (2025). Enhancing granular sentiment classification with chain-of-thought prompting in large language models. *arXiv preprint arXiv:2505.04135*.
- Na, S., y cols. (2024). Inference optimization of large language models on cpus. *arXiv preprint arXiv:2407.07304*. (Verificar autores completos y datos de publicación en Google Scholar)
- Nguyen, H., Veluchamy, A., Diop, M., y Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, 1(4), 7.
- Nádas, T., Dheeban, J., y Tomescu, V. (2025). A systematic evaluation of prompting techniques for synthetic data generation. *arXiv preprint*.
- Panthati, J., Bhaskar, J., Ranga, T. K., y Challa, M. R. (2018). Sentiment analysis of product reviews using deep learning. En *2018 international conference on advances in computing, communications and informatics (icacci)* (pp. 2408–2414).

- Pathak, S., y cols. (2023). Decision support system for corporate reputation based social media listening using a cross-source sentiment analysis engine. En *12th international conference on data science, technology and applications* (pp. 559–567).
- Perikos, I., y Diamantopoulos, A. (2024). Explainable aspect-based sentiment analysis using transformer models. *Big Data and Cognitive Computing*, 8(11), 141.
- Ponnusamy, M., Sonvani, T., Traylor, B., y Nanda, G. (2025). Comparing llm-based topic modeling and sentiment analysis with human evaluation of consumer insights. *Available at SSRN 5362200*.
- Schmitt, M., Schwerk, A., y Lempert, S. (2026). Enhancing sentiment classification and irony detection in large language models through advanced prompt engineering techniques. *arXiv preprint arXiv:2601.08302*.
- Sokolova, M., y Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. doi: 10.1016/j.ipm.2009.03.002
- Tan, K. L., Lee, C. P., Anbananthen, K. S. M., y Lim, K. M. (2022). Roberta-lstm: a hybrid model for sentiment analysis with transformer and recurrent neural network. *Ieee Access*, 10, 21517–21525.
- Team, G. (2024). Gemma: Open models based on gemini research and technology. *Google DeepMind Technical Report*. Descargado de <https://arxiv.org/abs/2403.08295>
- Touvron, H., y cols. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. Descargado de <https://arxiv.org/abs/2302.13971>
- Wang, Y., y Luo, Z. (2023). Enhance multi-domain sentiment analysis of review texts through prompting strategies. En *2023 international conference on high performance big data and intelligent systems (hdis)* (pp. 1–7).
- Wolf, T., Debut, L., Sanh, V., y cols. (2020). Transformers: State-of-the-art natural language processing. En *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations (emnlp)* (pp. 38–45).
- Yan, H., y Li, Z. (2022). Review of sentiment analysis: An emotional product development view. *frontiers of engineering management*, 9 (4), 592–609.
- Yang, A., Yang, B., Zhang, B., y cols. (2024). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Zhan, T., Shi, C., Shi, Y., Li, H., y Lin, Y. (2024). Optimization techniques for sentiment analysis based on llm (gpt-3). *arXiv preprint arXiv:2405.09770*.
- Zhao, X., Xu, R., y Guo, X. (2023). Post-training quantization or quantization-aware training? that is the question. En *2023 china semiconductor technology international conference (cstic)* (pp. 1–3).
- Zheng, M., Pei, J., Logeswaran, L., Lee, M., y Jurgens, D. (2024). When” a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large

language models. En *Findings of the association for computational linguistics: Emnlp 2024* (pp. 15126–15154).