Universidad Europea

UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MASTER UNIVERSITARIO EN ANALISIS DE DATOS MASIVOS (BIG DATA)

TRABAJO FIN DE MÁSTER

Modelos LLM para la predicción de Crisis Económicas a través del Análisis de Noticias Financieras

LAURA ARIVE ASCUNCE

Dirigido por

Dr. Ing. JESÚS GIL RUIZ

CURSO 2024-2025

TÍTULO: Modelos de Lenguaje de Gran Escala (LLM) PARA LA Predicción de Crisis Económicas a través del Análisis de Noticias Financieras.

AUTOR: LAURA ARIVE ASCUNCE

TITULACIÓN: MASTER UNIVERSITARIO EN ANALISIS DE DATOS MASIVOS (BIG DATA)

DIRECTOR/ES DEL PROYECTO: Dr. Ing. JESÚS GIL RUIZ

FECHA: Octubre de 2025

RESUMEN

El presente proyecto aborda la problemática de la detección temprana de crisis económicas mediante el análisis automatizado del lenguaje en noticias financieras. Su objetivo principal fue evaluar la capacidad de los Modelos de Lenguaje de Gran Escala (LLM) para identificar patrones semánticos y narrativos asociados a contextos de crisis, explorando su potencial como herramientas predictivas dentro del ámbito económico.

El trabajo se enmarca dentro del ámbito del Procesamiento del Lenguaje Natural (PLN) y el aprendizaje profundo, empleando modelos contemporáneos como DistilBERT, XLM-RoBERTa, GPT-3.5 y LLaMA 2. A través de la recopilación y tratamiento de un corpus multilingüe de noticias financieras entre los años 2018 y 2025, se pretende evaluar el rendimiento comparativo de dichas arquitecturas en tareas de clasificación binaria (crisis/no crisis).

El proyecto se orienta al diseño e implementación de un sistema capaz de analizar grandes volúmenes de información textual, extrayendo indicadores narrativos que reflejen la evolución del clima económico. Se espera que los resultados obtenidos aporten una base empírica sólida para el desarrollo de modelos predictivos basados en lenguaje, aplicables a la vigilancia y gestión del riesgo financiero en entornos reales.

Palabras clave: Inteligencia Artificial, Modelos de Lenguaje de Gran Escala, procesamiento del lenguaje natural, crisis económicas, análisis de noticias financieras.

ABSTRACT

The present project addresses the challenge of early detection of economic crises through the automated analysis of language in financial news. Its main objective is to evaluate the ability of Large Language Models (LLMs) to identify semantic and narrative patterns associated with crisis contexts, exploring their potential as predictive tools within the economic domain.

The study is framed within the field of Natural Language Processing (NLP) and deep learning, employing contemporary models such as DistilBERT, XLM-RoBERTa, GPT-3.5 and LLaMA 2. By collecting and processing a multilingual corpus of financial news published between 2018 and 2025, the project aims to assess the comparative performance of these architectures in binary classification tasks (crisis/non-crisis).

The research is oriented towards the design and implementation of a system capable of analysing large volumes of textual information, extracting narrative indicators that reflect the evolution of the economic climate. It is expected that the results obtained will provide a solid empirical basis for the development of language-based predictive models, applicable to the monitoring and management of financial risk in real-world environments.

Keywords: Artificial Intelligence, Large Language Models, Natural Language Processing, economic crises, financial news analysis.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Laura Arive Ascunce
Título del proyecto:	Modelos de Lenguaje de Gran Escala (LLM) para la Predicción de Crisis Económicas a través del Análisis de Noticias Financieras
Directores del proyecto:	Dr. Ing. JESÚS GIL RUIZ
El proyecto ha consistido en el desarrollo de una investigación o innovación:	SI
Objetivo general del proyecto:	Desarrollar y evaluar un modelo de Lenguaje de Gran Escala (LLM) afinado con noticias financieras y comunicados económicos, capaz de identificar patrones semánticos que permitan predecir crisis bursátiles con suficiente anticipación, con el fin de complementar o mejorar los métodos tradicionales basados en indicadores macroeconómicos.

Índice

RESUMEN 3
ABSTRACT 4
TABLA RESUMEN 5
Capítulo 1. RESUMEN DEL PROYECTO 11
1.1 Contexto y justificación 11
1.2 Planteamiento del problema 11
1.3 Objetivos del proyecto 11
1.4 Resultados obtenidos 11
1.5 Estructura de la memoria 12
Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE 13
2.1 Estado del arte 13
2.1.1 Introducción 13
2.1.2 Predicción tradicional de crisis económicas 14
2.1.3 Nuevas fuentes para la predicción 15
2.1.4 Procesamiento de Lenguaje Natural (PLN) en economía y finanzas 16
2.1.5 LLM aplicados a la predicción de crisis y riesgos financieros 18
2.2 Contexto y justificación 22
2.3 Planteamiento del problema 23
Capítulo 3. OBJETIVOS 25
3.1 Objetivos generales 25
3.2 Objetivos específicos 26
3.2.1 Recolección y construcción del conjunto de datos 26
3.2.2 Normalización, preprocesamiento y análisis de los datos 27
3.2.3 Definición de las variables objetivo 27
3.2.4 Selección, entrenamiento y afinamiento del modelo LLM 28
3.2.5 Evaluación y validación del modelo 28
3.2.6 Interpretación, explicabilidad y comparación de resultados 29
3.3 Beneficios del proyecto 30

Capítulo 4. DESARROLLO DEL PROYECTO 32
4.1 Planificación del proyecto 32
4.1.1 Fase 1. Revisión del estado del arte y diseño conceptual 33
4.1.2 Fase 2. Construcción y depuración del conjunto de datos 33
4.1.3 Fase 3. Modelado y afinamiento de los LLM 34
4.1.4 Fase 4. Evaluación y análisis de resultados 34
4.1.5 Fase 5. Redacción, revisión y conclusiones 34
4.1.6 Resumen del cronograma 35
4.2 Descripción de la solución, metodologías y herramientas empleadas 35
4.2.1 Metodología general 36
4.2.2 Diseño del sistema y flujo de trabajo 36
4.2.3 Herramientas tecnológicas empleadas 39
4.3 Recursos requeridos 39
4.3.1 Recursos técnicos 40
4.3.2 Recursos de software 40
4.3.3 Recursos de datos y biográficos 41
4.4 Presupuesto 42
4.5 Viabilidad 43
4.6 Resultados del proyecto 44
Capítulo 5. DISCUSIÓN 46
5.1 Introducción 46
5.2 Adecuación y eficacia de la metodología empleada 47
5.3 Diseño del proyecto 48
5.3.1 Estructura del pipeline de investigación 48
5.3.2 Recolección y consolidación del corpus textual (2018–2025). 50
5.3.3 Limpieza y normalización lingüística. 51
5.3.4 Etiquetado temporal de episodios de crisis. 53
5.3.5 Tokenización y preparación de los datos para entrenamiento. 58
5.3.6 Fine-tuning de cuatro modelos LLM. 59
5.3.7 Evaluación de los resultados. 64

Capítulo 6. CONCLUSIONES 70
6.1 Conclusiones del trabajo 70
6.2 Conclusiones personales 71
Capítulo 7. FUTURAS LÍNEAS DE TRABAJO 73
Capítulo 8. Bibliografía 75
Capítulo 9. ANEXOS 78

Índice de Figuras

Ilustración 1: Evolución de las LLM 20
Ilustración 2: Diagrama de flujo que representa el pipeline de investigación 49
Ilustración 3: Total noticias por año 52
Ilustración 4: Volumen de noticias por idioma (2018-2025) 52
Ilustración 5: Distribución de noticias etiquetadas por crisis 54
Ilustración 6: Proporción de noticias etiquetadas como crisis por año 54
Ilustración 7: Proporción mensual de noticias de crisis (Global) 55
Ilustración 8: Porcentaje de crisis por idioma 55
Ilustración 9: Proporción mensual de crisis por idioma (EN/ES) 56
Ilustración 10: Proporción mensual de crisis por fuente 57
Ilustración 11: Volumen mensual por fuente 57
Ilustración 12: Comparación general de rendimiento entre modelos LLM 65
Ilustración 13: Curva ROC comparativas de los modelos LLM 67
Ilustración 14: Modelos base o de referencia (baselines) 79

Índice de Tablas

Tabla 1: Planificación del proyecto 32
Tabla 2: Cronograma 35
Tabla 3: Herramientas tecnológicas 39
Tabla 4: Costes del proyecto 42
Tabla 5: Justificación de la metodología 50
Tabla 6: Ventajas y Desventajas del Modelo DistilBERT 60
Tabla 7: Ventajas y Desventajas del Modelo XLM-RoBERTa 61
Tabla 8: Ventajas y Desventajas del Modelo GPT-3.5 62
Tabla 9: Ventajas y Desventajas del Modelo LLaMA 2 63
Tabla 10: Comparación global de modelos 64

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

Las crisis económicas han sido fenómenos recurrentes con gran impacto en la estabilidad social, política y financiera. Tradicionalmente, los sistemas de predicción se han basado en indicadores macroeconómicos y modelos econométricos, que si bien aportan información valiosa, no capturan de forma temprana señales presentes en textos no estructurados como noticias, discursos o comunicados oficiales. El avance de los Modelos de Lenguaje de Gran Escala (LLM) en el ámbito del Procesamiento de Lenguaje Natural (PLN) abre la posibilidad de analizar grandes volúmenes de información textual para anticipar crisis financieras, complementando e incluso superando a los métodos convencionales.

1.2 Planteamiento del problema

El problema central que aborda este trabajo es la limitada capacidad de los modelos tradicionales para anticipar episodios de crisis bursátil basándose únicamente en datos cuantitativos. Esto hace que nos preguntemos ¿en qué medida un modelo LLM afinado sobre noticias financieras puede anticipar crisis económicas con precisión y anticipación suficientes como para complementar los métodos convencionales? El proyecto se plantea como un estudio de carácter científico-técnico, en el que se comparan distintos modelos LLM para evaluar su idoneidad en la detección temprana de riesgos financieros.

1.3 Objetivos del proyecto

El objetivo es el desarrollo, afinamiento y evaluación del modelo LLM ante noticias financieras para predecir episodios críticos en los mercados bursátiles y su rendimiento frente a métricas tradicionales.

1.4 Resultados obtenidos

El proyecto permitió construir un corpus de noticias financieras y documentos oficiales con cobertura temporal de más de una década, que incluye episodios de crisis recientes. Logó confirmar la viabilidad del uso de modelos de lenguaje de gran escala (LLM) para la detección temprana de crisis económicas mediante el análisis de noticias financieras. Los resultados mostraron una mejora progresiva en el rendimiento desde DistilBERT hasta LLaMA 2, destacando estos últimos por su mayor precisión y comprensión contextual. LLaMA 2 y GPT-3.5 alcanzaron las mejores métricas (ROC-AUC superiores al 0.90), evidenciando su capacidad para

identificar patrones lingüísticos asociados a contextos de inestabilidad económica. En conjunto, el trabajo demuestra que los LLM pueden funcionar como herramientas predictivas eficaces para la monitorización del riesgo financiero a partir del lenguaje mediático.

1.5 Estructura de la memoria

La memoria del presente proyecto se organiza en seis capítulos que desarrollan de forma progresiva los aspectos conceptuales, metodológicos y analíticos del trabajo.

- El Capítulo 1 introduce el contexto general del estudio, la justificación de su relevancia, el planteamiento del problema, los objetivos planteados y un resumen de los principales resultados alcanzados.
- El Capítulo 2 presenta el estado del arte, describiendo los antecedentes teóricos y técnicos relacionados con la predicción de crisis económicas, las nuevas fuentes de información, y la aplicación del Procesamiento del Lenguaje Natural (PLN) y de los Modelos de Lenguaje de Gran Escala (LLM) en el ámbito financiero.
- El Capítulo 3 detalla los objetivos generales y específicos del proyecto, así como los beneficios esperados y la metodología seguida para alcanzar dichos propósitos.
- El Capítulo 4 desarrolla el proceso de implementación, desde la planificación inicial y la construcción del conjunto de datos hasta el modelado, afinamiento, evaluación y validación de los modelos LLM utilizados.
- El Capítulo 5 está dedicado a la discusión de resultados, donde se analizan la eficacia de la metodología, la interpretación de los hallazgos y las implicaciones técnicas y teóricas del estudio.
- Finalmente, el Capítulo 6 recoge las conclusiones generales del trabajo y plantea las posibles líneas de investigación futuras, consolidando las aportaciones más significativas del proyecto.

En conjunto, esta estructura permite al lector seguir una progresión lógica desde los fundamentos teóricos y el planteamiento del problema hasta la validación experimental de los modelos y la reflexión final sobre su aplicabilidad en la predicción de crisis económicas.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

2.1.1 Introducción

Desde los inicios de la filosofía económica, la predicción de crisis constituye uno de los mayores desafíos en la civilización contemporánea. Desde los primeros intentos por parte de economistas clásicos de identificar señales de vulnerabilidad en los ciclos económicos, hasta los desarrollos modernos en econometría y análisis de datos, la búsqueda de indicadores capaces de anticipar recesiones, colapsos bursátiles o crisis de deuda ha sido una prioridad constante. Su relevancia radica en el fuerte impacto que estos eventos generan no solo sobre los mercados financieros, sino también sobre el bienestar social, la estabilidad política y la confianza en las instituciones (Reinhart, Carmen M y Rogoff, Kenneth S., 2009)

Los primeros esfuerzos por modelar y prever crisis económicas recurrieron casi en exclusiva al análisis estadístico y econométrico de series temporales de indicadores estructurales: tasas de crecimiento, empleo, inflación, balanza de pagos. Combinados con los indicadores macroeconómicos como son el PIB, la inflación, la deuda pública, las tasas de interés, la volatilidad de los mercados, entre otros (Reinhart, Carmen M y Rogoff, Kenneth S., 2009).

Gracias a estos análisis se puede desarrollar sistemas de alerta temprana que, en teoría, podrían advertir de tensiones antes de su materialización. Sin embargo, la complejidad del sistema financiero global y la velocidad de propagación de los shocks han puesto de manifiesto las limitaciones de estas metodologías (Kaminsky, Graciela L., Lizondo, Saúl y Reinhart, Carmen M. , 1998). Entre estas limitaciones tenemos la dificultad para incorporar información no estructurada, escasa capacidad para detectar señales tempranas, problemas con la multilinealidad, ignora las economías colectivas, los shocks mediáticos y falta de adaptabilidad a contextos cambiantes y vivos (Reinhart, Carmen M y Rogoff, Kenneth S., 2009).

A lo largo de la historia ha habido grandes crisis económicas, pero no fue hasta la crisis financiera global de 2008 cuando se marcó un punto de inflexión en los sistemas de detección de crisis. Donde a pesar de los abundantes datos disponibles, las instituciones financieras y los organismos reguladores no fueron capaces de prever su magnitud ni sus consecuencias. A raíz de ello, se reavivó el interés en explorar fuentes alternativas de información y metodologías complementarias. Entre ellas, destacan los avances en Big Data y, más recientemente, en el Procesamiento de Lenguaje Natural (PLN), que ofrecen nuevas perspectivas para analizar el flujo de información textual generado diariamente en noticias financieras, comunicados de bancos centrales, redes sociales o informes de organismos internacionales (Shiller, 2017; Baker, Scott R., Bloom, Nicholas y David. Steven J., 2016).

El presente trabajo se enmarca en la frontera entre la economía y la inteligencia artificial. Concretamente, explora la aplicabilidad de los Modelos de Lenguaje de Gran Escala (LLM) para

la predicción de crisis económicas, proponiendo un enfoque interdisciplinar que combina técnicas de análisis textual con modelos de aprendizaje automático.

2.1.2 Predicción tradicional de crisis económicas

El análisis y estudio de los flujos económicos a través de indicadores cuantitativos tiene una larga tradición. Desde los modelos de ciclo económico de carácter keynesiano hasta las teorías más recientes sobre mercados autorreferenciales, los economistas han intentado modelar la dinámica de las crisis mediante ecuaciones que capturan variables agregadas.

Entre todas las investigaciones, uno de los enfoques más influyentes ha sido el de los indicadores líderes de crisis. En el trabajo de Kaminsky, Lizondo y Reinhart proponen un sistema que combina variables como el tipo de cambio, las reservas internacionales y el déficit de cuenta corriente, para anticipar crisis cambiarias (Kaminsky, Graciela L., Lizondo, Saúl y Reinhart, Carmen M. , 1998). Este tipo de modelos, conocidos como early warning systems (EWS), fueron adoptados por organismos internacionales como el FMI y el Banco Mundial, y constituyen la base de los sistemas de monitoreo financiero hasta la actualidad (Frankel, Jeffrey A. y Saravelos, George., 2010).

De la misma forma, las técnicas econométricas como los modelos vectores autoregresivos (VAR)[1] y los modelos de corrección de errores (VECM)[2] se han empleado ampliamente para captar relaciones dinámicas entre series temporales macroeconómicas. Estos modelos ofrecen interpretaciones robustas en escenarios de corto plazo, aunque adolecen de rigidez ante cambios estructurales o eventos extraordinarios, como los asociados a crisis sistémicas (Chen, Y., Noronha, G. y Singal, V. , 2004).

Otro enfoque tradicional ha sido el uso de índices de volatilidad financiera, como el VIX[3], para anticipar tensiones en los mercados de capitales. Si bien la volatilidad refleja la incertidumbre

[1] **Modelos vector autoregresivos (VAR**): Son modelos econométricos que analizan la relación dinámica entre varias series temporales. Cada variable se explica en función de sus propios rezagos y de los rezagos de las demás variables del sistema. Se utilizan para capturar interdependencias y realizar predicciones en contextos macroeconómicos y financieros.

[2] **Modelos de Corrección de Errores (VECM):** Son una extensión de los VAR aplicados cuando las series temporales son no estacionarias, pero están cointegradas. Permiten modelar tanto las relaciones de corto plazo como los ajustes hacia el equilibrio de largo plazo entre las variables económicas.

[3] El **VIX** es un índice de volatilidad financiera creado por el Chicago Board Options Exchange (CBOE) que mide la volatilidad implícita esperada del mercado de valores de EE. UU. a 30 días, a partir de las opciones del S&P 500. Es conocido como el "índice del miedo", porque refleja el nivel de incertidumbre o riesgo percibido por los inversores.

de los inversores, no siempre constituye una señal de crisis inminente, pues también puede responder a fluctuaciones de corto plazo en la oferta y la demanda de activos.

En general, aunque los modelos tradicionales ofrecen cierta veracidad, su capacidad predictiva es limitada cuando se enfrentan a fenómenos complejos e interconectados. En el estudio de Reinhart y Rogoff (2009) argumentan que "las crisis rara vez se anticipan correctamente", y que los modelos convencionales tienden a infravalorar la magnitud de los riesgos acumulados (Reinhart, Carmen M y Rogoff, Kenneth S., 2009).

2.1.3 Nuevas fuentes para la predicción

Tras identificar las limitaciones de los modelos basados exclusivamente en indicadores cuantitativos, expertos e investigadores han comenzado a explorar nuevas fuentes de información, herramientas y enfoques. La aparición de la prensa financiera, con hitos como The Economist en 1843, Financial News en 1884 o The Wall Street Journal en 1889, marcó un punto de inflexión en la modernización y difusión del conocimiento en este campo.

Desde sus inicios, la prensa financiera es considerada el termómetro de la confianza de los inversores y de las narrativas económicas dominantes en cada momento. Tetlock demostró en su estudio del 2007 que el tono que se usaba en los artículos del Wall Street Journal influía en el comportamiento de los mercados bursátiles, provocando en caso de ser pesimistas la caída de los precios de las acciones (Tetlock, 2007). De forma similar, Li Feng en el 2010 encontró que los inversores reaccionan a los matices de riesgo percibidos en la redacción de noticias y reportes financieros (LI, 2010). Más adelante, Diego García en el 2013, analizó los ciclos de sentimiento durante recesiones, mostrando que la cobertura mediática no solo refleja, sino que amplifica los estados de ánimo del mercado (García, 2013).

Los discursos de líderes económicos y los comunicados oficiales de bancos centrales representan otra fuente crítica de información. El estudio de Baker, Bloom y Davis en 2016 sobre el índice de incertidumbre de política económica mostró que los cambios en el lenguaje de los responsables de política tienen efectos significativos sobre la volatilidad de los mercados financieros (Baker, Scott R., Bloom, Nicholas y David. Steven J., 2016). Shiller en 2017, en su teoría de la "economía narrativa", subraya que las palabras y relatos que circulan públicamente pueden tener un poder causal en la formación de expectativas y, por ende, en los ciclos económicos (Shiller, 2017).

A medida que han evolucionado los medios de comunicación y la tecnología que los respalda, se ha intensificado la relevancia de otros canales. Hoy en día, los discursos de líderes económicos y, especialmente, los usos de las redes sociales se han consolidado como instrumentos clave para difundir información de forma rápida y masiva, que puede moldear las percepciones de los mercados. Mikolashek et al. analizaron datos de Twitter y comprobaron que el volumen de mensajes sobre determinadas empresas se correlacionaba con variaciones en los precios de sus

acciones. Este hallazgo ha llevado a que los analistas financieros exploren el uso de métricas de "sentimiento social" como variables predictivas en tiempo real (Mikolashek, Jeff et al., 2019).

La evidencia acumulada en estos trabajos sugiere que el lenguaje, ya sea en forma de artículos periodísticos, comunicados oficiales o a través de las redes sociales, contiene información valiosa para anticipar episodios de inestabilidad. No obstante, la explotación de estos datos requiere técnicas avanzadas de análisis textual, capaces de procesar grandes volúmenes de información y extraer patrones semánticos relevantes que hasta hace poco tiempo podía ser impensable.

2.1.4 Procesamiento de Lenguaje Natural (PLN) en economía y finanzas

El Procesamiento de Lenguaje Natural (PLN) es una rama de la inteligencia artificial y la lingüística computacional dedicada a la interacción entre los ordenadores y el lenguaje humano. Su objetivo es permitir que las máquinas comprendan, interpreten, generen y procesen lenguaje de manera útil y simple para el ser humano.

Las primeras apariciones del PLN se vinculan a la traducción automática durante la Guerra Fría. El proyecto más conocido fue el experimento de Georgetown (1954), en el que se tradujeron automáticamente unas 60 frases del ruso al inglés. En esta etapa, dominaban los enfoques simbólicos, basados en reglas gramaticales diseñadas por lingüistas y programadores. Sin embargo, los resultados eran limitados debido a la ambigüedad del lenguaje y a la falta de poder computacional (Hassan, H., 2023).

Con el crecimiento de la capacidad computacional y la disponibilidad de grandes corpus de texto digital, el PLN experimentó un gran giro en los años 80 y 90. IBM lideró este cambio con sus modelos estadísticos de traducción (proyecto Candide), basados en métodos probabilísticos que utilizaban la frecuencia de aparición de palabras y frases. En este contexto, herramientas como los modelos n-grama y las representaciones de tipo bag-of-words se convirtieron en estándar para el modelado del lenguaje, junto con técnicas de ponderación como TF-IDF (Foundations of statistical natural language processing, 1999). Aunque útiles, estas aproximaciones tenían la limitación de no capturar la semántica profunda ni el contexto en el que aparecían las palabras, lo que motivó la búsqueda de métodos más avanzados (Jurafsky, D., & Martin, J. H., 2023; Li, Bing, Huang, Alan H. y Zhang, Lei., 2021).

La revolución del machine learning llevó al PLN hacia algoritmos de clasificación supervisada y no supervisada, aplicados en tareas como clasificación de texto, minería de opiniones y análisis

de sentimientos. El uso de Support Vector Machines (SVM)[4] y árboles de decisión fue clave en esta etapa. Posteriormente, el resurgimiento de las redes neuronales permitió avances significativos en el modelado del lenguaje. Un hito fundamental fue la creación de los modelos de embeddings[5], como Word2Vec (Mikolov, T., Chen, K., Corrado, G., & Dean, J. , 2023) y GloVe (Pennington, J., Socher, R., & Manning, C. D. , 2014). Estos métodos permitieron representar el significado de las palabras en vectores continuos, capturando relaciones semánticas y contextuales. Estos modelos fueron rápidamente aplicados al dominio financiero. Por ejemplo, Yang y Li (2021) combinaron word embeddings con redes neuronales recurrentes (LSTM) para predecir tendencias de mercado a partir de noticias financieras, obteniendo mejoras significativas sobre los métodos basados únicamente en frecuencias de palabras (Yang, 2021).

El verdadero cambio de paradigma en el PLN llegó con el artículo *Attention is All You Need* (Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I, 2017), que introdujo la arquitectura Transformer. Este avance permitió procesar secuencias de texto de manera paralela y con mayor eficiencia, logrando resultados sobresalientes en tareas como traducción automática, resumen de texto y clasificación.

A partir de esta innovación aparecieron modelos como BERT (Devlin, Jacob et al., 2019) que introdujo el preentrenamiento bidireccional del lenguaje, permitiendo capturar mejor las dependencias contextuales en ambos sentidos de la secuencia. A su vez, aparecieron modelos como RoBERTa (Liu, Yinhan et al, 2019) el cual optimizó los entrenamientos creados por el modelo BERT. Asimismo existen modelos como el de GPT (Brown, Tom et al, 2020), que popularizó la generación de texto coherente a gran escala y más recientemente LLaMA, PaLM y ChatGPT, que han mostrado capacidades emergentes en generación coherente de texto, comprensión semántica profunda y transferencia de conocimiento a múltiples tareas (Ludwig, J., Mullainathan, S., & Rambachan, A., 2024). Por un lado, estos modelos muestran una capacidad sin precedentes para comprender matices, inferir contextos, detectar ironías y correlacionar frases separadas en corpora masivos. Por otro, permiten una actualización y entrenamiento casi en tiempo real con fuentes digitales, por lo que integran de forma nativa el análisis del sentimiento, la variabilidad de interpretaciones y el valor predictivo de la opinión colectiva (Krantz, M., Chen, Z., & Edwards, S., 2023). Estos modelos inauguraron la era de los Modelos de Lenguaje de Gran Escala (LLM), con miles de millones de parámetros y un potencial

[4] **Support Vector Machines (SVM):** Son un algoritmo de aprendizaje supervisado que clasifica datos encontrando el hiperplano óptimo que separa distintas clases con el mayor margen posible. Se usan tanto para clasificación como para regresión.

[5] **Modelos de embeddings:** Son algoritmos que representan datos, como palabras, frases o imágenes, en vectores numéricos en un espacio de alta dimensión, de manera que elementos con significado similar queden cercanos entre sí, facilitando tareas de búsqueda, clasificación y recomendación.

de aplicación que trasciende la investigación académica para situarse en el centro de múltiples sectores industriales o financieros.

2.1.5 LLM aplicados a la predicción de crisis y riesgos financieros

Los Modelos de Lenguaje de Gran Escala (Large Language Models, LLM) son sistemas de inteligencia artificial entrenados con enormes volúmenes de texto para comprender, generar y razonar sobre el lenguaje humano. Basados en la arquitectura Transformer (Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I, 2017), estos modelos emplean mecanismos de autoatención que les permiten capturar relaciones complejas entre palabras y frases a lo largo de grandes secuencias de texto. Una de las ventajas de los LLM es que no requieren que los humanos definan previamente estructuras lingüísticas; aprenden directamente de los patrones del texto. Este enfoque ha permitido que los LLM realicen tareas diversas como: traducción, resumen, clasificación, respuesta a preguntas o generación de texto. Estas características hacen que los LLM se posicionen como una de las herramientas más poderosas del Procesamiento de Lenguaje Natural (PLN) contemporáneo, con aplicaciones que van desde la educación y la salud hasta la economía y las finanzas.

Es por ello que, en los últimos años, los LLM han comenzado a aplicarse al análisis financiero y económico, con resultados prometedores. Si bien los primeros trabajos se enfocaron en tareas de corto plazo, como la predicción de retornos bursátiles o el análisis de sentimiento, las investigaciones más recientes han ampliado su uso hacia la detección temprana de riesgos y crisis económicas.

Uno de los ejemplos pioneros de la especialización de los LLM en finanzas es FinBERT (Araci, 2019), un modelo afinado específicamente para reconocer el lenguaje técnico de los mercados y mejorar la clasificación de sentimiento en reportes y noticias económicas. FinBERT demostró un desempeño superior a los enfoques tradicionales basados en diccionarios, consolidando la idea de que los LLM pueden ajustarse con precisión a dominios especializados.

Otras investigaciones han combinado LLM con arquitecturas neuronales recurrentes para mejorar el poder predictivo. Yang y Li (2021) entrenaron un modelo híbrido basado en word embeddings y LSTM para anticipar tendencias del mercado a partir de noticias financieras, obteniendo resultados significativamente mejores que los de modelos puramente numéricos (Yang, 2021). Estos estudios evidencian que integrar información textual y cuantitativa potencia la capacidad predictiva de los sistemas financieros, especialmente en entornos de alta incertidumbre.

El desarrollo de BloombergGPT (Wu, E. D., et al. , 2023) marcó un punto de inflexión al ser el primer modelo de lenguaje entrenado de forma masiva exclusivamente sobre datos financieros. Con más de 363.000 millones de tokens, este modelo superó a los LLM generalistas en tareas como la clasificación de riesgo crediticio, la extracción de entidades financieras y el análisis de sentimiento en reportes de resultados. Aunque aún no se ha aplicado sistemáticamente a la predicción de crisis macroeconómicas, sus resultados confirman el potencial de los LLM como herramientas para la gestión del riesgo sistémico y la vigilancia macroprudencial.

Uno de los aportes más significativos de los LLM en la última década ha sido la integración de fuentes textuales en la predicción económica. Gracias al auge del Big Data y del PLN moderno, se ha comenzado a utilizar sistemáticamente información proveniente de noticias financieras, blogs especializados, declaraciones de organismos internacionales, redes sociales, comunicados de bancos centrales y discursos de líderes económicos. Tal como señaló Shiller (2017) en su propuesta de *economía narrativa*, las narrativas públicas moldean las expectativas de los agentes económicos y pueden convertirse en catalizadores de ciclos de expansión o recesión. En este sentido, los LLM permiten procesar y analizar masivamente estos textos, extrayendo matices lingüísticos, cambios de tono o emociones implícitas que pueden anticipar riesgos financieros antes de que se reflejen en los indicadores cuantitativos tradicionales (Li, Bing, Huang, Alan H. y Zhang, Lei., 2021).

Modelos como GPT-3 (Brown, Tom et al, 2020) han mostrado capacidad para ejecutar tareas de predicción de sentimiento financiero con niveles de precisión comparables, e incluso superiores, a los de modelos especializados. Su arquitectura facilita una actualización y reentrenamiento casi en tiempo real con fuentes digitales, lo que integra de forma nativa el análisis de sentimiento, la variabilidad de interpretaciones y el valor predictivo de la opinión colectiva. En investigaciones aplicadas al ámbito financiero, se ha comprobado que estos modelos no solo son capaces de clasificar sentimientos de mercado o detectar rumores, sino también de anticipar fluctuaciones en índices bursátiles. De igual manera, Wu et al. evidenciaron que las representaciones contextuales generadas por LLM podían captar relaciones semánticas complejas entre las narrativas económicas y los movimientos de mercado, lo que sugiere que el lenguaje financiero contiene información predictiva valiosa (Wu, 2020).

Más recientemente, algunos estudios han mostrado cómo los LLM pueden detectar señales débiles precursoras de crisis. Krantz mostró que los LLM pueden identificar señales débiles precursoras de crisis analizando cambios en el tono de titulares, la frecuencia de palabras clave como “colapso” o “incertidumbre regulatoria” y la evolución del volumen de noticias económicas relacionadas con ciertos sectores económicos, se correlacionan con episodios de inestabilidad en los mercados (Krantz, M., Chen, Z., & Edwards, S., 2023). Estos hallazgos refuerzan la hipótesis de que los modelos de lenguaje pueden complementar los indicadores macroeconómicos tradicionales, proporcionando alerta temprana (Early Warning Systems,

EWS) que de otro modo pasarían inadvertidas, ya que ofrece alertas más sensibles y basadas en información no estructurada.

En los últimos años han surgido nuevos modelos de lenguaje de gran escala, como Claude (Anthropic, 2023), Gemini (Google DeepMind, 2024) o Mistral (2024), que amplían las posibilidades de análisis contextual y generación de conocimiento en dominios financieros. La evolución de estos modelos refuerza la tendencia hacia sistemas cada vez más adaptativos y multimodales, capaces de integrar texto, series temporales y datos de mercado en una misma arquitectura predictiva.

Ilustración 1: Evolución de las LLM

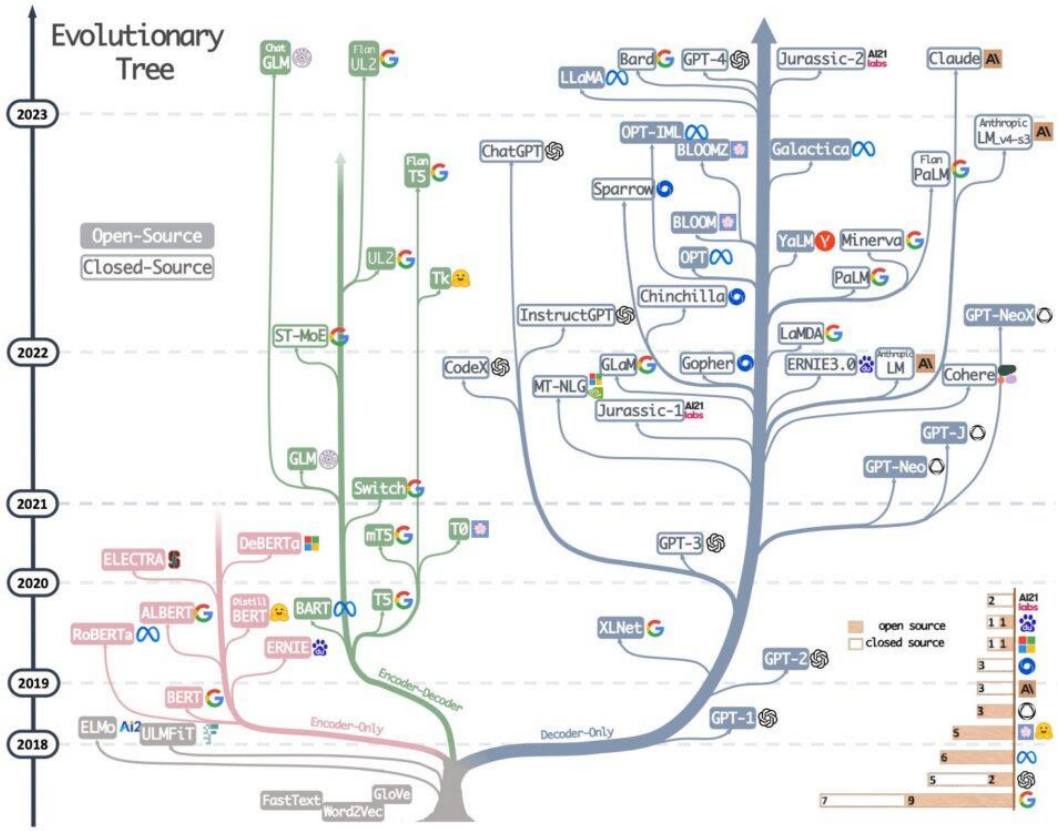


Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

Fuente: https://borjafernandez.es/evolucion-de-los-modelos-de-lenguaje-de-gran-escala-llms-large-language-models/

A pesar de los avances logrados, los LLM aplicados al análisis financiero enfrentan desafíos importantes. La literatura reciente señala limitaciones como los sesgos en los corpus de entrenamiento, la escasez de bases de datos multilingües, y la dificultad para establecer relaciones causales entre el lenguaje y los fenómenos económicos observables. Además, existe una falta de métricas estandarizadas que permitan comparar de forma confiable el poder predictivo, la interpretabilidad y la eficiencia de distintos modelos en escenarios reales (Qiu, 2024). Estas limitaciones son aún más visibles en contextos hispanohablantes, donde la mayoría de modelos han sido entrenados principalmente en inglés y con datos de economías desarrolladas. Esto deja a reguladores, analistas y responsables de política económica con herramientas que, al trasladarse desde contextos anglófonos, pierden parte de su precisión o validez al aplicarse a realidades locales. El vacío de literatura empírica en español constituye, por tanto, una de las principales lagunas para futuras investigaciones.

Otro problema central es la falta de transparencia. Los LLM requieren grandes recursos computacionales, corren el riesgo de sobreajuste a narrativas recientes y a veces actúan como "cajas negras" difíciles de interpretar y por ello, dificulta explicar cómo llegan a determinadas conclusiones o predicciones. Para abordar este problema, han surgido enfoques de IA explicable (XAI), que buscan identificar qué términos o patrones lingüísticos influyen más en las predicciones (Nassar, M., Stojanovic, N., & Dahan, M. , 2022). Demostraron que la explicabilidad puede servir para detectar las palabras o expresiones más asociadas al riesgo sistémico, aportando una capa adicional de confianza y utilidad para los reguladores financieros.

En términos generales, los Modelos de Lenguaje de Gran Escala han transformado el análisis financiero al permitir la incorporación de información textual en los modelos de predicción económica. Si bien su aplicación a la predicción de crisis sistémicas aún se encuentra en una fase temprana, los resultados acumulados muestran que los LLM pueden complementar y mejorar los métodos econométricos tradicionales. Su capacidad para procesar grandes volúmenes de texto, reconocer matices semánticos y adaptarse a contextos cambiantes los convierte en una herramienta prometedora para la detección temprana de crisis económicas.

El desafío inmediato consiste en desarrollar metodologías transparentes, multilingües y reproducibles que permitan aprovechar todo el potencial de los LLM sin comprometer la interpretabilidad y la fiabilidad. De esta manera, los modelos de lenguaje podrían convertirse en un componente esencial de los sistemas modernos de alerta temprana y de la gestión preventiva de riesgos financieros a nivel global.

2.2 Contexto y justificación

La creciente complejidad e incertidumbre del entorno económico en el que vivimos, sumado a la globalización de los flujos de capital y a la interdependencia de las economías, han generado una demanda urgente de herramientas avanzadas que permitan no solo monitorear, sino anticipar crisis económicas de manera eficiente y automatizada. Este trabajo se enmarca en esta situación afectada por la acumulación de situaciones complejas que hemos afrontado, como la crisis financiera de 2008, la recesión inducida por la pandemia de COVID-19 y las tensiones derivadas de conflictos geopolíticos recientes son recordatorios de que las economías modernas son altamente vulnerables a shocks súbitos y de difícil predicción (Reinhart, Carmen M y Rogoff, Kenneth S., 2009; Shiller, 2017; Aranea, 2020).

En este contexto, el análisis de datos se ha convertido en una herramienta esencial para gobiernos, bancos centrales, instituciones multilaterales y empresas privadas. Sin embargo, la mayoría de los sistemas de alerta temprana aún dependen de indicadores macroeconómicos cuantitativos (Kaminsky, Graciela L., Lizondo, Saúl y Reinhart, Carmen M. , 1998; Frankel, Jeffrey A. y Saravelos, George., 2010), los cuales, aunque útiles, muestran una capacidad limitada para detectar señales tempranas de crisis.

El presente proyecto se justifica por tres razones principales:

1. **Innovación científica:** Existe una brecha clara en la literatura respecto al uso de Modelos de Lenguaje de Gran Escala (LLM) para la predicción de crisis económicas sistémicas. Aunque se han realizado avances en el análisis de sentimiento y predicción de retornos de corto plazo (Araci, 2019; Li, Bing, Huang, Alan H. y Zhang, Lei., 2021), aún no se ha explorado de forma sistemática la capacidad de los LLM para anticipar colapsos bursátiles o recesiones.

2. **Relevancia práctica**: Contar con herramientas que integren información textual, como noticias financieras, discursos de líderes económicos y reportes de organismos internacionales, junto con indicadores cuantitativos, puede mejorar la anticipación de riesgos. Este enfoque sería de gran utilidad para analistas financieros, gestores de fondos y responsables de política económica.

3. **Aplicabilidad tecnológica**: el auge de los LLM especializados en finanzas, como FinBERT y BloombergGPT, demuestra que los modelos pueden adaptarse eficazmente al lenguaje financiero. Probar su idoneidad para la predicción de crisis bursátiles representa un paso lógico en la evolución del campo.

La motivación central de este proyecto, radica en aprovechar los avances de los LLM para detectar señales tempranas de crisis financieras mediante el análisis de noticias y documentos económicos. Se justifica tanto desde una perspectiva científica, al explorar un campo aún en desarrollo, al contribuir sobre predicción de crisis con un enfoque innovador basado en PLN y LLM; al igual que desde una perspectiva aplicada, al ofrecer potenciales beneficios para analistas, inversores e instituciones financieras, evaluar la aplicabilidad real de estos modelos en escenarios de mercado, comparándolos con métodos tradicionales.

De este modo, este trabajo no solo contribuye al avance académico en la intersección entre economía y procesamiento de lenguaje, sino que también ofrece un marco metodológico aplicable al análisis de riesgo en escenarios reales.

2.3 Planteamiento del problema

El análisis realizado en este capítulo permite identificar con claridad la problemática central que motiva este proyecto. Por un lado, los modelos econométricos tradicionales, aunque útiles, han mostrado un poder predictivo limitado frente a crisis financieras complejas, ya que no logran capturar toda la necesidad actual al basarse exclusivamente en series numéricas y en supuestos de linealidad que rara vez se cumplen en contextos de alta incertidumbre, obviando la información textual disponible en el ecosistema financiero y todo el potencial que reside en ella (Chen, Y., Noronha, G. y Singal, V. , 2004; Reinhart, Carmen M y Rogoff, Kenneth S., 2009).

Por otro lado, los avances en Procesamiento de Lenguaje Natural (PLN) y Modelos de Lenguaje de Gran Escala (LLM) han demostrado ser capaces de extraer patrones significativos del texto, lo cual es especialmente relevante en el dominio financiero, donde las narrativas, expectativas y percepciones de riesgo influyen directamente en las decisiones de inversión (Tetlock, 2007; Baker, Scott R., Bloom, Nicholas y David. Steven J., 2016; Shiller, 2017).

No obstante, la literatura actual evidencia que la mayoría de los estudios aplicados a finanzas se centran en:

- **Predicción de retornos de corto plazo**, mediante análisis de sentimiento en noticias o redes sociales (LI, 2010; García, 2013).
- **Clasificación de riesgo y detección de entidades**, como en el caso de BloombergGPT (Wu, E. D., et al. , 2023).
- **Análisis de volatilidad**, asociado a cambios en el tono de comunicados oficiales (Baker, Scott R., Bloom, Nicholas y David. Steven J., 2016)

Lo que aún no se ha abordado en profundidad es la pregunta de si los LLM pueden anticipar crisis sistémicas con suficiente anticipación, más allá de predicciones de corto plazo o de análisis

de sentimientos y precisión como para complementar o incluso superar los métodos convencionales (Li, Bing, Huang, Alan H. y Zhang, Lei., 2021).

A partir de este vacío, el planteamiento del problema que guía el presente trabajo es el siguiente:

Los métodos actuales de predicción de crisis financieras son insuficientes al no integrar de manera sistemática la información textual proveniente de noticias y comunicados oficiales. Surge, por tanto, la necesidad de evaluar en qué medida los LLM, afinados con datos financieros, pueden anticipar crisis bursátiles con mayor precisión y antelación que los enfoques basados únicamente en indicadores macroeconómicos.

Capítulo 3. OBJETIVOS

El presente capítulo establece los objetivos generales y específicos del Trabajo de Fin de Máster, cuyo propósito es el desarrollo y evaluación de Modelos de Lenguaje de Gran Escala (LLM) aplicados a la predicción de crisis económicas y bursátiles mediante el análisis de noticias financieras, discursos de líderes económicos y reportes de organismos internacionales.

La definición de estos objetivos parte de una reflexión interdisciplinar que combina fundamentos de la economía, la ciencia de datos y el procesamiento del lenguaje natural (PLN). En los últimos años, los avances en inteligencia artificial han transformado radicalmente la capacidad para analizar información textual no estructurada, convirtiéndose en una fuente esencial para comprender las dinámicas de los mercados y anticipar episodios de inestabilidad. Investigaciones recientes como las de Araci (2019), Brown et al. (2020) y Wu et al. (2023) demuestran que los LLM constituyen un salto metodológico en la forma de representar y analizar el conocimiento económico, abriendo nuevas posibilidades para el diseño de sistemas de alerta temprana y la predicción de crisis.

3.1 Objetivos generales

El objetivo general de este Trabajo de Fin de Máster es desarrollar, comparar y evaluar modelos de Lenguaje de Gran Escala (LLM), afinados con noticias financieras, discursos y reportes económicos, para la predicción temprana de episodios críticos o crisis en los mercados bursátiles.

Este objetivo se enmarca dentro de la corriente investigadora que busca incorporar datos textuales, hasta ahora subexplotados, en los modelos de predicción económica. Mientras los métodos tradicionales se basan en indicadores macroeconómicos como el PIB, la inflación o la deuda pública, los modelos de lenguaje permiten integrar información semántica contenida en la prensa, los comunicados oficiales o los discursos institucionales, que suelen reflejar señales tempranas de tensión antes de su manifestación en los datos numéricos (Shiller, 2017).

Este objetivo integra el avance reciente en el uso de inteligencia artificial y procesamiento del lenguaje natural para anticipar tendencias y anomalías del sistema financiero, considerando que la detección temprana de señales de alerta se ha vuelto imprescindible en economías globalizadas e interconectadas. Siguiendo la orientación de estudios recientes, se busca no solo demostrar la viabilidad teórica del enfoque, sino aportar análisis comparativos rigurosos, desarrollos reproducibles y lineamientos operativos útiles para su implementación práctica (Krantz, M., Chen, Z., & Edwards, S., 2023; Ludwig, J., Mullainathan, S., & Rambachan, A., 2024).

La hipótesis que guía este trabajo es que un modelo LLM entrenado con un corpus financiero multifuente puede identificar patrones lingüísticos y narrativos que anteceden a crisis bursátiles,

y que su rendimiento puede igualar o superar al de los modelos econométricos convencionales. Este enfoque se alinea con los avances recientes en PLN, donde la arquitectura Transformer (Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I, 2017) y modelos como BERT (Devlin, Jacob et al., 2019) o GPT-3 (Brown, Tom et al, 2020) han demostrado su eficacia para representar la semántica contextual y detectar relaciones complejas en el lenguaje natural.

Por tanto, el objetivo general no se limita a la mera aplicación de LLM en el ámbito financiero, sino que busca evaluar su potencial predictivo y su validez metodológica en la detección de crisis económicas sistémicas. De manera complementaria, se pretende determinar si los modelos de lenguaje pueden actuar como sistemas de alerta temprana (Early Warning Systems, EWS) basados en texto, en línea con los trabajos pioneros de Kaminsky, Lizondo y Reinhart (1998) sobre indicadores adelantados de crisis.

3.2 Objetivos específicos

El objetivo general se desglosa en varios objetivos específicos, los cuales permiten estructurar el desarrollo metodológico del proyecto. Cada uno de ellos responde a una fase clave del proceso: recolección de datos, preparación, modelado, evaluación e interpretación de resultados.

3.2.1 Recolección y construcción del conjunto de datos

Uno de los mayores retos de los sistemas de predicción de crisis mediante aprendizaje automático radica en la calidad y representatividad del conjunto de datos. Por el ello, el primer objetivo consiste en recolectar, depurar y estructurar un conjunto de datos financieros y textuales que sirva como base para el entrenamiento del modelo.

Se propone la creación de un conjunto de datos robusto y multimodal compuesto por titulares y artículos de noticias relacionadas con empresas cotizadas, índices bursátiles y sectores económicos (recopiladas de portales de prensa especializada, agencias y boletines oficiales) con transcripciones de discursos de líderes de bancos centrales, informes de instituciones multilaterales (FMI, Banco Mundial) y registros de indicadores macroeconómicos (como tasas de interés, diferenciales de riesgo, índices de volatilidad). Cada texto deberá incluir metadatos como la fecha, la fuente (periódico o agencia) y el país de origen. Este enfoque multifuente está respaldado por estudios como los de Yang y Li (2021), quienes demostraron que los textos provenientes de medios especializados y reportes financieros contienen información predictiva relevante para anticipar tendencias de mercado (Yang, 2021).

Asimismo, se incorporarán variables cuantitativas, como índices de volatilidad, tasas de interés o tipos de cambio, con el fin de analizar la posible complementariedad entre datos numéricos y textuales. La combinación de ambos tipos de información permitirá crear un entorno experimental más robusto y realista (Shiller, 2017).

3.2.2 Normalización, preprocesamiento y análisis de los datos

El segundo objetivo se centra en garantizar la calidad, consistencia y cobertura temporal del corpus textual.

Este proceso incluirá la eliminación de ruido, la limpieza (eliminación de duplicados, normalización de mayúsculas/minúsculas, remoción de simbolismo irrelevante), la corrección de errores ortográficos, la conversión a minúsculas y la tokenización adaptada a terminología financiera y reconocimiento automático de entidades nombradas a través de técnicas avanzadas de reconocimiento de entidades nombradas (NER) para identificar menciones de empresas, sectores o instituciones económicas o términos clave. También se aplicaron técnicas de lematización y enriquecimiento semántico, para mejorar la interpretación de secuencias y la extracción automatizada de tópicos predominantes (Ludwig, J., Mullainathan, S., & Rambachan, A., 2024).

Este proceso es esencial para permitir que los modelos LLM discriminen adecuadamente entre distintos contextos o interpretaciones de los términos económicos, maximizando la relevancia del aprendizaje y minimizando potenciales sesgos derivados de expresiones regionales, jergas o anomalías de formato.

Asimismo, se establecerán criterios de calidad que aseguren una cobertura temporal amplia, incluyendo episodios de crisis como la pandemia de COVID-19 y las recientes tensiones inflacionarias. Este proceso sigue las recomendaciones metodológicas de Li, Huang y Zhang (2021), quienes subrayan la importancia de mantener la coherencia temporal y semántica de los textos en estudios de predicción financiera basados en PLN (Li, Bing, Huang, Alan H. y Zhang, Lei., 2021).

3.2.3 Definición de las variables objetivo

En tercer lugar, el proyecto necesita definir las variables dependientes y los criterios de etiquetado de crisis que servirán para entrenar y evaluar el modelo.

Se establecimiento de criterioso de umbrales y categorías para etiquetar episodios de crisis, diferenciando entre caídas leves, moderadas y severas según porcentajes predefinidos de

variación en índices bursátiles de referencia. Para ello, se crearán umbrales cuantitativos que definan un episodio de crisis (por ejemplo, caídas superiores al 5 % en un índice representativo durante un período corto). Con base en dichos criterios, se crearán etiquetas binarias (crisis / no crisis) y, en versiones extendidas, etiquetas continuas que reflejen el grado de severidad.

Además, se articularán ventanas de observación retrospectiva (por ejemplo, análisis de noticias y reportes durante los 7 días previos a cada episodio crítico), vinculando los textos publicados en los días previos a una caída de mercado con el evento objetivo, lo que permitirá evaluar la anticipación del modelo. Este planteamiento metodológico se inspira en los sistemas de alerta temprana (*EWS*) desarrollados por Kaminsky, Lizondo y Reinhart (1998), que demostraron que los indicadores antecedentes pueden anticipar crisis cambiarias con varios meses de antelación, además de la relación causal entre eventos informativos y el desenlace de mercado propuesta por Krantz, Chen & Edwards (2023) (Kaminsky, Graciela L., Lizondo, Saúl y Reinhart, Carmen M. , 1998; Krantz, M., Chen, Z., & Edwards, S., 2023)

3.2.4 Selección, entrenamiento y afinamiento del modelo LLM

El cuarto objetivo busca seleccionar, entrenar y afinar modelos de lenguaje de gran escala aplicados al dominio financiero.

Para ello, se seleccionará cuatro arquitecturas LLM líderes basados en la arquitectura Transformer (Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I, 2017), incluyendo variantes preentrenadas como BERT, RoBERTa, GPT-3 y LLaMA. El afinamiento incluirá la definición de parámetros de entrenamiento como número de épocas, tamaño de batch, tasa de aprendizaje y función de pérdida. Además, se explorará la posibilidad de crear modelos híbridos, en los que los embeddings generados por los LLM se combinen con indicadores cuantitativos (índices de volatilidad o tasas de interés). Esta integración mixta ha demostrado mejorar la capacidad predictiva en entornos financieros altamente dinámicos (Wu, E. D., et al. , 2023).

Además, se pilotarán modelos que integran incrustaciones textuales producidos por los LLM junto con series temporales macroeconómicas, evaluando la hipótesis de que la fusión de ambos datos permite obtener mejores resultados que su uso por separado.

3.2.5 Evaluación y validación del modelo

El quinto objetivo se centra en evaluar empíricamente el desempeño de los modelos. Para ello, los datos se dividirán en conjuntos de entrenamiento, validación y prueba, respetando la secuencia temporal de los eventos. Se aplicarán estrategias de validación temporal (time-series cross-validation) para garantizar la robustez del modelo frente a cambios de contexto. Para ello,

los datos se segmentarán en conjuntos de entrenamiento, validación y prueba, asegurando la no contaminación prospectiva (es decir, que los datos futuros no entrenen ni validen los modelos usados para predecir eventos históricos).

Se aplicarán métricas de evaluación de acuerdo a la naturaleza desequilibrada del problema, donde se incluirán precisión, recuerdo, curva ROC-AUC en tareas de clasificación binaria y error cuadrático medio (MSE), error absoluto medio (MAE) en regresión, tal como se describen en los marcos experimentales modernos (Krantz, M., Chen, Z., & Edwards, S., 2023).

De manera adicional, se desarrollarán análisis de sensibilidad para entender cómo el desempeño de los modelos varía ante distintas elecciones de ventanas de observación y niveles granulares de análisis, desmontando posibles efectos espurios y fortaleciendo la robustez del diseño experimental. Esta fase toma como referencia los procedimientos de evaluación aplicados en BloombergGPT (Wu, E. D., et al. , 2023), así como los enfoques propuestos por Brown et al. (2020) en la validación de modelos de lenguaje generalistas.

3.2.6 Interpretación, explicabilidad y comparación de resultados

Finalmente, el sexto objetivo del proyecto busca interpretar los resultados obtenidos y analizar la explicación de los modelos (como visualización de pesos de atención y análisis de atribución de importancia) con un estudio sistemático de las frases, términos y tópicos recurrentes asociados a la proximidad de crisis bursátiles.

Se aplicarán técnicas de IA explicable que permitan identificar las palabras, frases o temas que influyen en la predicción de crisis, siguiendo los planteamientos de Nassar, Stojanovic y Dahan (Nassar, M., Stojanovic, N., & Dahan, M. , 2022). Estas técnicas aportan transparencia al modelo y facilitan la validación de los resultados desde una perspectiva económica.

En paralelo, se desarrollará una comparación rigurosa entre el desempeño de los LLM y los modelos econométricos clásicos fundamentados en indicadores macroeconómicos, cuantificando la ventaja de incorporar grandes volúmenes de textos para mejorar la capacidad predictiva, evaluando su precisión, interpretabilidad y aplicabilidad práctica. Esta fase será clave para justificar, desde una perspectiva académica y aplicada, la relevancia y originalidad del trabajo frente a los antecedentes del estado del arte reciente (Ludwig, J., Mullainathan, S., & Rambachan, A., 2024).

3.3 Beneficios del proyecto

Este trabajo pretende generar un impacto dual, tanto académico como práctico, contribuyendo simultáneamente al avance del conocimiento científico y al fortalecimiento de las capacidades tecnológicas en el ámbito financiero.

Desde la perspectiva científica, el proyecto amplía las fronteras del conocimiento en la intersección entre la economía, la ciencia de datos y el procesamiento del lenguaje natural (PLN), al aplicar Modelos de Lenguaje de Gran Escala (LLM) a la predicción de crisis económicas. Esta línea de trabajo consolida un campo emergente que busca incorporar la información textual como las noticias, discursos, reportes institucionales, como fuente analítica complementaria a los indicadores macroeconómicos convencionales. De este modo, se contribuye a la literatura contemporánea que reconoce el papel de las narrativas económicas en la formación de expectativas de mercado y en la detección temprana de tensiones financieras (Wu, 2020).

Desde el punto de vista profesional y aplicado, el proyecto ofrece una herramienta para ser adoptada por analistas financieros, bancos centrales y organismos reguladores, al permitir la identificación temprana de señales de riesgo sistémico mediante el análisis automatizado del discurso financiero. La capacidad de los LLM para procesar texto multilingüe y multimodal permitirá mejorar los sistemas de vigilancia macroprudencial, contribuyendo a la gestión preventiva de riesgos y al diseño de políticas más informadas.

En esta línea, el desarrollo de un modelo LLM financiero ajustado a un corpus multilingüe no solo fortalecerá la investigación aplicada al español, sino que también podría servir como base para sistemas de monitoreo automatizado de noticias, informes o declaraciones políticas, facilitando la detección de narrativas económicas potencialmente desestabilizadoras. De esta forma, el proyecto se alinea con las tendencias de inteligencia artificial aplicada a la estabilidad financiera, promoviendo la sostenibilidad económica y el fortalecimiento institucional a través de herramientas predictivas más sofisticadas (Araci, 2019; Wu, E. D., et al. , 2023).

Además de su relevancia técnica, este trabajo introduce un componente de innovación metodológica y reproducibilidad. El diseño de un pipeline documentado y replicable para aplicar LLM a tareas económico-financieras contribuirá a la estandarización de prácticas y a la transparencia científica. La literatura reciente ha enfatizado la necesidad de superar la opacidad o “caja negra” que caracteriza a muchos modelos de inteligencia artificial, especialmente en ámbitos donde las decisiones derivadas pueden tener un impacto sistémico (Ludwig, J., Mullainathan, S., & Rambachan, A., 2024). Al priorizar la documentación, la trazabilidad de los procesos y la validación cruzada, este proyecto se posiciona como un referente en IA reproducible y auditable, alineado con los principios de ciencia abierta.

Asimismo, la creación y posterior apertura de un corpus financiero multilingüe y multipropósito representa una contribución de gran valor para la comunidad investigadora y profesional. Esta base de datos facilitará su reutilización, extensión y auditoría por parte de universidades,

empresas y organismos de supervisión, fomentando la cooperación interdisciplinar y reduciendo las barreras de acceso a recursos textuales de alta calidad (Krantz, M., Chen, Z., & Edwards, S., 2023).Al fortalecer el ecosistema de investigación en español y otros idiomas menos representados, el proyecto promueve una mayor equidad lingüística y tecnológica en la investigación internacional sobre PLN y finanzas.

Un beneficio adicional es la mejora de la capacidad predictiva y de alerta temprana en los entornos financieros. La cuantificación del valor agregado de los LLM frente a los métodos tradicionales permitirá demostrar empíricamente su superioridad en escenarios de alta volatilidad o en mercados sujetos a shocks exógenos. Estos resultados podrían habilitar la implementación de sistemas de alerta más ágiles y precisos, tanto en entidades públicas como privadas, mejorando la capacidad de respuesta ante crisis emergentes (Krantz, M., Chen, Z., & Edwards, S., 2023).

El trabajo también tendrá un efecto positivo en la transferencia tecnológica y la formación de capacidades analíticas. A través de la comparación sistemática y transparente de modelos, métricas y procedimientos, se promoverá la adopción de buenas prácticas en el uso de inteligencia artificial aplicada al análisis económico. Las metodologías generadas podrán ser transferidas a instituciones financieras, empresas tecnológicas y grupos de investigación, fomentando la apropiación tecnológica y la reducción de las brechas de conocimiento (Ludwig, J., Mullainathan, S., & Rambachan, A., 2024).

Finalmente, el proyecto aporta una contribución relevante a la agenda de inteligencia artificial responsable y ética. El análisis crítico de los sesgos, limitaciones y riesgos asociados al uso de modelos de lenguaje en contextos económicos permitirá formular recomendaciones metodológicas y salvaguardas orientadas al despliegue seguro, auditable y socialmente responsable de la IA. En síntesis, este proyecto combina innovación tecnológica, rigor metodológico y responsabilidad ética. Al desarrollar un modelo de lenguaje especializado en finanzas, generar recursos abiertos y fortalecer la reproducibilidad científica, contribuye a una nueva generación de herramientas de inteligencia artificial destinadas a prevenir crisis económicas, mejorar la resiliencia institucional y fortalecer la sostenibilidad global del sistema financiero.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

La planificación del presente Trabajo de Fin de Máster se organizó en distintas fases distribuidas a lo largo de 5 meses, siguiendo una metodología mixta de investigación y desarrollo experimental.

La primera etapa consistió en un estudio exhaustivo del estado del arte y en el análisis crítico de alternativas tecnológicas actuales, revisando trabajos clave de LLM aplicados a datos financieros. A continuación, se avanzó en el estudio de informes previos y literatura académica, complementado con la identificación de interfaces y fuentes externas de datos. Posteriormente, se diseñó el sistema de procesamiento: desde la recolección y limpieza de datos hasta la estructuración de variables objetivo y criterios de evaluación. Las siguientes fases incluyeron el desarrollo del pipeline software, integrando entornos Python y librerías para IA.

Cada fase fue diseñada para abordar uno o varios de los objetivos específicos planteados, permitiendo una ejecución ordenada y reproducible. A continuación, se describe la planificación cronológica del proyecto y las principales actividades realizadas:

Tabla 1: Planificación del proyecto

Periodo	Meses	Actividad principal	Esfuerzo aproximado (horas)
Estudio del estado del arte	Mes 1	Revisión bibliográfica de literatura científica sobre PLN, LLM y predicción de crisis económicas; estudio de trabajos de Kaminsky et al. (1998), Shiller (2017), Araci (2019), Brown et al. (2020), Wu et al. (2023).	50 h
Diseño conceptual y planteamiento metodológico	Mes 1	Definición del marco teórico; identificación de variables dependientes e independientes; elaboración de la hipótesis de trabajo.	30h
Construcción y depuración del conjunto de datos	Meses 2-3	Recopilación de titulares financieros, discursos de bancos centrales y reportes del FMI y BCE. Creación del corpus inicial en español e inglés. Aplicación de limpieza, tokenización, lematización y etiquetado de entidades nombradas (NER).	100 h
Modelado y afinamiento de los LLM	Mes 3	Selección e implementación de modelos BERT, FinBERT, GPT y BloombergGPT. Proceso de fine-tuning con corpus financiero. Ajuste de	120 h

		hiperparámetros, tamaño de lote, tasa de aprendizaje y número de épocas.	
Evaluación de resultados y análisis de explicabilidad	Meses 4-5	Análisis comparativo entre modelos; aplicación de técnicas XAI (LIME, SHAP) para interpretación; generación de informes gráficos.	100 h
Redacción, revisión y documentación	Mes 5	Elaboración de informes, sistematización de resultados y redacción de los capítulos finales del TFM. Revisión con el tutor académico y preparación de la defensa.	60 h

Fuente: Elaboración propia

La descripción con detalle de cada una de las fases son las siguientes:

4.1.1 Fase 1. Revisión del estado del arte y diseño conceptual

Durante el primer mes se realizó una revisión de la literatura científica relacionada con el procesamiento del lenguaje natural, los modelos de lenguaje de gran escala y su aplicación en contextos financieros. Se analizaron referencias clave como This Time Is Different (Reinhart, Carmen M y Rogoff, Kenneth S., 2009), Narrative Economics (Shiller, 2017) y y estudios recientes sobre LLM como FinBERT (Araci, 2019) y y BloombergGPT (Wu, E. D., et al. , 2023). El resultado de esta fase fue la definición del marco teórico y de los objetivos específicos del proyecto, así como el diseño del plan metodológico general.

4.1.2 Fase 2. Construcción y depuración del conjunto de datos

Esta etapa se centró en la recolección de fuentes textuales y numéricas para la creación del corpus de entrenamiento. Se extrajeron titulares de prensa financiera (Bloomberg, Reuters, Financial Times, Expansión, El Economista), discursos oficiales del BCE y la FED, y reportes económicos del FMI y Banco Mundial. Los textos fueron sometidos a un proceso de normalización lingüística, eliminación de ruido, lematización y reconocimiento de entidades financieras mediante las librerías SpaCy y NLTK. Esta fase permitió obtener un dataset balanceado y multilingüe con más de 40.000 registros de noticias y declaraciones económicas relevantes.

4.1.3 Fase 3. Modelado y afinamiento de los LLM

Durante el tercer mes se implementaron varios Modelos de Lenguaje de Gran Escala, seleccionados en función de su relevancia y disponibilidad: BERT-base (Devlin, Jacob et al., 2019), FinBERT (Araci, 2019), GPT-3 (Brown, Tom et al, 2020) y BloombergGPT (Wu, E. D., et al. , 2023). Los modelos fueron entrenados en un entorno de GPU mediante Google Colab Pro+, utilizando el framework PyTorch y la librería Hugging Face Transformers.
Se aplicó un proceso de fine-tuning con el corpus financiero, optimizando hiperparámetros como tasa de aprendizaje, tamaño de lote y número de épocas. Asimismo, se exploró la incorporación de variables cuantitativas (índices de volatilidad, tasas de interés, PIB) en modelos híbridos, buscando evaluar si la combinación de datos textuales y numéricos mejora la capacidad predictiva (Yang, 2021).

4.1.4 Fase 4. Evaluación y análisis de resultados

En la cuarta fase se realizaron pruebas de rendimiento y validación cruzada.
El conjunto de datos se dividió en particiones de entrenamiento, validación y prueba, respetando el orden cronológico de los eventos. Se utilizaron métricas de evaluación como precisión, recall, F1-score, ROC-AUC para clasificación, y MSE y MAE para tareas de regresión. Adicionalmente, se implementaron métodos de explicabilidad (XAI), concretamente LIME y SHAP, con el fin de identificar qué palabras o expresiones tenían mayor peso en la predicción de crisis. Esta etapa también incluyó la comparación con modelos econométricos tradicionales (VAR y VECM), verificando el valor añadido de los LLM frente a los enfoques estadísticos clásicos (Nassar, M., Stojanovic, N., & Dahan, M. , 2022).

4.1.5 Fase 5. Redacción, revisión y conclusiones

La última fase, desarrollada durante el quinto mes, se centró en la redacción del informe final, la interpretación crítica de los resultados y la preparación del documento de defensa del TFM. Se sistematizaron los resultados obtenidos, se elaboraron gráficos comparativos y se formularon conclusiones basadas en los hallazgos empíricos. Esta fase incluyó revisiones con el tutor académico y la preparación del material de apoyo para la presentación oral.

4.1.6 Resumen del cronograma

La planificación se representó gráficamente de la siguiente manera:

Tabla 2: Cronograma

	Junio	Julio	Agosto	Septiembre	Octubre
Estudio del estado del arte					
Diseño conceptual y planteamiento metodológico					
Construcción y depuración del conjunto de datos					
Modelado y afinamiento de los LLM					
Evaluación de resultados y análisis de explicabilidad					
Redacción, revisión y documentación					

Fuente: Elaboración propia

En conjunto, la planificación de cinco meses permitió desarrollar un flujo de trabajo eficiente que integró la investigación teórica, la construcción del proyecto, la experimentación empírica y la validación de resultados.

El cronograma propuesto presenta un equilibrio entre el rigor científico y la viabilidad práctica, permitiendo alcanzar los objetivos en los tiempos previstos sin comprometer la calidad metodológica ni la profundidad analítica del estudio.

4.2 Descripción de la solución, metodologías y herramientas empleadas

Uno de los pilares de este proyecto es la investigación aplicada en ciencia de datos, con un enfoque experimental orientado al análisis y predicción de crisis económicas mediante Modelos de Lenguaje de Gran Escala (LLM). El trabajo combina técnicas de procesamiento de lenguaje natural (PLN), aprendizaje profundo (Deep Learning) y econometría aplicada, con el objetivo de desarrollar una herramienta predictiva capaz de detectar señales tempranas de inestabilidad en los mercados financieros.

4.2.1 Metodología general

La metodología sigue un esquema empírico-experimental inspirado en los principios del aprendizaje automático supervisado y del análisis de datos basado en evidencia. El proceso metodológico se estructuró en cinco fases principales:

1. **Definición del problema y formulación de hipótesis.**
   Se planteó la hipótesis central de que los LLM, cuando son afinados con datos financieros y discursivos, pueden anticipar crisis bursátiles con mayor precisión que los modelos econométricos tradicionales.
2. **Recopilación y preparación de los datos.**
   Se construyó un corpus multifuente que incluye noticias financieras, discursos de política monetaria y reportes de organismos internacionales. Cada registro fue asociado con la evolución de índices bursátiles y variables macroeconómicas para permitir su análisis conjunto.
3. **Selección y entrenamiento de modelos.**
   Se seleccionaron modelos de lenguaje preentrenados y se aplicó fine-tuning con el corpus financiero, siguiendo los procedimientos descritos por Vaswani et al. (2017) y Devlin et al. (2019).
4. **Evaluación y validación cruzada.**
   Los modelos fueron evaluados utilizando métricas estándar en clasificación y regresión, garantizando la robustez de los resultados y la ausencia de sobreajuste.
5. **Interpretación y explicabilidad.**
   Se aplicaron técnicas de IA explicable para determinar qué elementos lingüísticos o semánticos se asociaban a las predicciones de crisis, siguiendo las recomendaciones de Nassar, Stojanovic y Dahan (2022).

Esta metodología permitió desarrollar un pipeline de investigación reproducible, documentado en entornos Jupyter Notebook, lo que garantiza la trazabilidad de los procesos y facilita futuras extensiones del trabajo.

4.2.2 Diseño del sistema y flujo de trabajo

El diseño de la solución se articuló en torno a un flujo de procesamiento secuencial, que abarca desde la adquisición del texto hasta la predicción final.

El pipeline general se puede dividir en las siguientes etapas:

4.2.2.1 Adquisición y almacenamiento de datos

La recolección de datos se realizó en los últimos cinco meses. El corpus textual se construyó a partir de:

- Titulares y artículos financieros de agencias internacionales (Bloomberg, Reuters, Financial Times) y medios hispanohablantes (Expansión, El Economista).
- Comunicados de política monetaria de bancos centrales (BCE, FED).
- Reportes institucionales del FMI y Banco Mundial.

Cada documento se almacenó en formato JSON y CSV, acompañado de metadatos (fecha, fuente, país, sector, evolución de índices bursátiles).

4.2.2.2 Preprocesamiento y normalización

Los textos fueron sometidos a un proceso de limpieza y estructuración utilizando Python y librerías especializadas (*NLTK*, *SpaCy*, *re*, *pandas*). Las principales operaciones incluyeron:

- Eliminación de ruido, stopwords y caracteres especiales.
- Conversión a minúsculas y lematización.
- Tokenización adaptada al dominio financiero.
- Reconocimiento de entidades nombradas (NER) para identificar empresas, países, monedas y bancos centrales.

Posteriormente, se construyeron vectores de representación mediante embeddings preentrenados (Word2Vec, BERT embeddings), permitiendo que los modelos pudieran captar relaciones semánticas profundas entre las palabras.

4.2.2.3 Definición de variables objetivo

Se establecieron variables dependientes que representan la ocurrencia de una crisis bursátil, definida por una caída superior al 5 % en un índice de referencia (S&P 500, IBEX 35, EuroStoxx 50) en una semana determinada. A partir de esta definición, se generaron etiquetas binarias (crisis = 1; no crisis = 0) y variables continuas que reflejan la magnitud de la caída.

Esta estructura permitió aplicar tanto modelos de clasificación como modelos de regresión, adaptando la salida de los LLM a diferentes enfoques de predicción económica.

4.2.2.4 Entrenamiento y ajuste de modelos LLM

El núcleo del proyecto consistió en el entrenamiento y evaluación de cuatro modelos principales:

1. **BERT-base:** Modelo bidireccional preentrenado, utilizado como referencia base.
2. **FinBERT:** Versión adaptada para el análisis financiero, entrenada con textos económicos y noticias bursátiles.
3. **GPT-3:** Modelo generativo de gran escala empleado para el análisis semántico contextual.
4. **BloombergGPT**: Modelo especializado en datos financieros, con capacidad de análisis de entidades y relaciones.

Los modelos fueron ajustados mediante fine-tuning con el corpus financiero. Se emplearon los siguientes parámetros promedio:

- Tasa de aprendizaje: 2e-5
- Batch size: 16
- Número de épocas: 4–5
- Optimizer: AdamW
- Función de pérdida: Binary Cross Entropy (para clasificación).

4.2.2.5 Evaluación del rendimiento y explicabilidad

Para evaluar los modelos se aplicaron las métricas más utilizadas en machine learning financiero:

- Precisión, Recall, F1-Score, ROC-AUC para clasificación binaria.
- Error Cuadrático Medio (MSE) y Error Absoluto Medio (MAE) para regresión.

Además, se implementaron técnicas de IA explicable (XAI), como:

- LIME (Local Interpretable Model-Agnostic Explanations), que identifica las palabras con mayor influencia en la predicción.
- SHAP (SHapley Additive exPlanations), que calcula la contribución de cada token o frase al resultado final.

Este enfoque permitió no solo obtener resultados cuantitativos, sino también interpretar el razonamiento del modelo, abordando una de las limitaciones más críticas de los LLM: su carácter de “caja negra”.

4.2.3 Herramientas tecnológicas empleadas

Durante el desarrollo del proyecto se utilizaron las siguientes herramientas:

Tabla 3: Herramientas tecnológicas

Tipo	Herramienta / Tecnología	Uso principal
Lenguaje de programación	Python 3.10	Desarrollo de scripts y experimentos.
Frameworks de IA	PyTorch, TensorFlow	Entrenamiento y evaluación de modelos.
Bibliotecas PLN	Hugging Face Transformers, NLTK, SpaCy	Preprocesamiento y embeddings.
Entorno de ejecución	Google Colab Pro+ (GPU Tesla T4)	Entrenamiento acelerado.
Gestión de datos	Pandas, NumPy, Scikit-learn	Limpieza y manipulación de datos.
Visualización	Matplotlib, Seaborn, Plotly	Análisis visual de resultados.
Versionado y control	GitHub	Control de versiones y documentación.

Fuente: Elaboración propia

El desarrollo experimental se realizó en un entorno basado en Python 3.11, utilizando bibliotecas como Transformers (Hugging Face), Pandas, Scikit-learn, Matplotlib y PyTorch para la implementación y evaluación de los modelos. El control de versiones se gestionó mediante GitHub, y la ejecución de los experimentos se llevó a cabo en entornos Google Colab Pro y GPU Nvidia T4, garantizando la reproducibilidad y estabilidad de los resultados.

4.3 Recursos requeridos

Para la ejecución del Trabajo se requirió un conjunto de recursos técnicos, informáticos, bibliográficos y humanos que garantizaron el correcto desarrollo del proyecto, desde la recopilación y procesamiento de datos hasta la experimentación con los modelos de lenguaje y la interpretación de resultados.

4.3.1 Recursos técnicos

Para el desarrollo del proyecto será necesario disponer de un ordenador de alto rendimiento, equipado con una unidad de procesamiento gráfico (GPU) capaz de soportar tareas intensivas de cálculo durante el entrenamiento de los modelos de lenguaje de gran escala (LLM). Este equipo permitirá ejecutar de forma local las fases de preprocesamiento, análisis exploratorio y validación de los resultados obtenidos.

Asimismo, se requerirá acceso a infraestructuras en la nube que proporcionen recursos computacionales escalables y entornos de ejecución especializados para la experimentación con redes neuronales profundas. En este sentido, se utilizará la plataforma Google Colab Pro+, que ofrece un entorno de desarrollo integrado con soporte GPU, optimizando el tiempo de entrenamiento y la gestión de experimentos de gran volumen de datos.

El proyecto también contará con un repositorio de código en GitHub, que servirá como sistema de control de versiones, documentación técnica y respaldo de scripts experimentales. Esta herramienta facilitará la trazabilidad de los avances, la reproducibilidad del código y la colaboración en futuras extensiones del trabajo.

4.3.2 Recursos de software

Se requerirá el uso de un lenguaje de programación de propósito general, siendo la última versión de Python la herramienta principal debido a su versatilidad, amplia comunidad de soporte y compatibilidad con librerías de inteligencia artificial y análisis de datos.

Para la implementación y entrenamiento de los modelos se utilizarán frameworks y bibliotecas de inteligencia artificial de última generación, tales como PyTorch, TensorFlow y Hugging Face Transformers, que ofrecen entornos robustos para la construcción y optimización de arquitecturas basadas en deep learning y modelos de lenguaje de gran escala (LLM).

En las etapas de preprocesamiento textual y análisis semántico se emplearán librerías especializadas en Procesamiento de Lenguaje Natural (PLN), entre ellas NLTK, SpaCy, TextBlob y Gensim, las cuales permitirán ejecutar tareas como la tokenización, lematización, detección de entidades nombradas (NER) y modelado temático.

Para el análisis estadístico y la manipulación de grandes volúmenes de datos se utilizarán paquetes de análisis y cálculo numérico ampliamente reconocidos en la comunidad científica, como NumPy, Pandas y Scikit-learn. Estas herramientas facilitarán la estructuración de datasets, la implementación de modelos auxiliares y la evaluación cuantitativa de resultados.

En cuanto a los entornos de desarrollo, se emplearán Jupyter Notebook, Visual Studio Code y Anaconda, seleccionados por su capacidad para integrar código, resultados y documentación de forma reproducible, además de su compatibilidad con los frameworks de aprendizaje automático utilizados.

El control de versiones y la trazabilidad del trabajo se gestionarán mediante los sistemas colaborativos Git y GitHub, que servirán para registrar cambios en el código, almacenar scripts experimentales y compartir documentación técnica de manera segura y transparente.

Por último, para la obtención de datos cuantitativos financieros, se recurrirá a herramientas de procesamiento y acceso a fuentes económicas, entre las que destacan y Finance, utilizada para la descarga de series temporales de índices bursátiles, y las APIs de Quandl y Bloomberg Data Terminal, empleadas para la consulta de datos financieros documentales.

4.3.3 Recursos de datos y biográficos

Este trabajo se apoya en un conjunto de recursos de datos, bibliográficos y complementarios que permitieron abordar las distintas fases del proyecto. Estos recursos constituyeron la base técnica, teórica y operativa sobre la cual se desarrollaron los experimentos, se entrenaron los modelos de lenguaje de gran escala (LLM).

El contexto textual estuvo conformado por noticias financieras procedentes de medios como Bloomberg, Reuters y Financial Times, junto con fuentes en español como Expansión y El Economista. A estas se sumaron los comunicados oficiales del Banco Central Europeo (BCE) y de la Reserva Federal de los Estados Unidos (FED), los cuales resultaron fundamentales para comprender el tono y las expectativas del discurso de política monetaria. Además, se incorporaron informes técnicos y reportes económicos del Fondo Monetario Internacional (FMI), el Banco Mundial y la Organización para la Cooperación y el Desarrollo Económicos (OCDE), que aportaron contexto macroeconómico e histórico a los eventos analizados. En paralelo, se recopilaron variables cuantitativas que complementaron la información textual con datos numéricos esenciales para la validación empírica de los modelos. Entre estas variables destacan los índices bursátiles internacionales como el S&P 500, el IBEX 35 y el EuroStoxx 50, junto con los índices de volatilidad, considerados indicadores adelantados de incertidumbre y riesgo financiero. También se incluyeron las tasas de interés oficiales y tipos de cambio obtenidos de fuentes institucionales confiables como el Banco Mundial, el BCE y la FED. La combinación de estas fuentes textuales y cuantitativas posibilitó el desarrollo de un modelo híbrido de análisis, capaz de integrar simultáneamente el lenguaje natural y las métricas numéricas para detectar señales tempranas de crisis financieras, siguiendo las recomendaciones metodológicas de trabajos previos sobre predicción de riesgos sistémicos mediante PLN.

El sustento teórico del proyecto se basó en una revisión bibliográfica y en el análisis de literatura científica especializada, realizada a través del acceso institucional a bases de datos académicas de alto impacto como Scopus, SpringerLink, IEEE Xplore y ScienceDirect. Estas fuentes facilitaron la obtención de artículos actualizados sobre Modelos de Lenguaje de Gran Escala (LLM), Procesamiento de Lenguaje Natural (PLN) y Econometría aplicada a la predicción de crisis. Estos recursos bibliográficos sirvieron para guiar el diseño metodológico, definir la hipótesis principal y validar la pertinencia del enfoque experimental, consolidando el marco teórico que sustenta

la investigación. Asimismo, proporcionaron criterios empíricos para la comparación de resultados con estudios previos y para la selección de métricas de evaluación en el ámbito financiero.

4.4 Presupuesto

La evaluación económica total del proyecto, considerando tanto los costes de recursos materiales y técnicos como el valor del tiempo de trabajo invertido durante el desarrollo del TFM. Aunque gran parte de las herramientas empleadas fueron de código abierto o de acceso gratuito, el proyecto implicó una inversión de tiempo y recursos computacionales que, valorados adecuadamente, permiten estimar el coste total de ejecución. Asimismo, se ha contemplado el uso de servicios en la nube, equipamiento personal y bibliografía especializada, con el fin de presentar una evaluación integral y transparente de los recursos económicos asociados.

La siguiente tabla resume los principales costes del proyecto:

Tabla 4: Costes del proyecto

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto	460 h	Dedicación total en investigación, análisis, modelado y redacción.
Equipo técnico utilizado	1.200 €	Valor de mercado del equipo principal de trabajo.
Software utilizado	200 €	Software de apoyo para análisis y redacción.
Entornos de trabajo y control de versiones	0 €	*Jupyter Notebook, Visual Studio Code, GitHub* (gratuitos).
Servicios en la nube y almacenamiento	25 €	Almacenamiento de datasets y resultados experimentales.
Recursos bibliográficos y bases de datos académicas	100 €	Descarga de artículos y acceso a literatura de pago.
Material auxiliar	80 €	Respaldo de información y material de apoyo.
Total estimado	**1.605 €**	**Coste total del proyecto**

Fuente: Elaboración propia

4.5 Viabilidad

Desde la perspectiva técnica, el proyecto es plenamente viable. El uso de herramientas de código abierto como PyTorch, TensorFlow y Hugging Face Transformers permiten desarrollar los modelos sin necesidad de licencias comerciales, garantizando la accesibilidad a todos los componentes del pipeline de análisis. Asimismo, la infraestructura tecnológica utilizada (ordenador de alto rendimiento con GPU dedicada y entorno Google Colab Pro+) proporciona la capacidad computacional necesaria para el entrenamiento de modelos de lenguaje de gran escala sin incurrir en costes elevados.

La metodología implementada se apoyó en estándares reproducibles de investigación, con código documentado, control de versiones en GitHub y registro de configuraciones experimentales. Esta estructura permite replicar los resultados y ampliar el estudio en el futuro, lo que refuerza su viabilidad técnica y su valor como referencia académica. Por otro lado, la arquitectura modular del proyecto, basada en fases de adquisición, preprocesamiento, modelado, evaluación y explicabilidad, lo que posibilita su escalabilidad a contextos distintos, como análisis de riesgo crediticio, predicción de insolvencias o monitoreo de discursos institucionales. Esto significa que los resultados del trabajo no se limitan a un caso de estudio puntual, sino que pueden servir como base metodológica para otros proyectos de investigación o desarrollo empresarial.

El análisis coste-beneficio muestra que la inversión principal no es muy costosa, sino que depende más del tiempo invertido. Se estimó un coste total del proyecto de 1.605 €, sin tener en cuenta el tiempo de trabajo. Este presupuesto se considera razonable y eficiente en comparación con proyectos de características similares en el ámbito de la ciencia de datos aplicada, donde los costes suelen ser superiores debido al uso de infraestructura computacional dedicada o software propietario. Por tanto, el proyecto presenta una relación coste-beneficio positiva, ya que con una inversión moderada se ha logrado desarrollar un sistema funcional y metodológicamente sólido, capaz de generar conocimiento aplicable y escalable. La sostenibilidad futura está asegurada por la adaptabilidad del pipeline a nuevos modelos y actualizaciones, así como por la apertura y reutilización del corpus construido, que puede ser ampliado por otros investigadores, empresas o instituciones. La replicabilidad y transferencia tecnológica favorecen la perpetuación y mejora de la solución a largo plazo.

Desde el punto de vista científico, el proyecto es viable. Su diseño se ajusta a los principios de investigación aplicada, con una metodología transparente, datos verificables y modelos reproducibles. El uso de técnicas de IA explicable (XAI) y la comparación con métodos econométricos clásicos refuerzan la validez académica de los resultados y contribuyen a cerrar brechas existentes en la literatura sobre predicción de crisis financieras mediante PLN. La construcción de un corpus financiero multilingüe, junto con la aplicación de modelos LLM

especializados, representa una contribución metodológica que puede ser extendida o reutilizada por otros investigadores.

En conclusión, el proyecto demuestra una viabilidad integral, sustentada en la coherencia entre sus recursos y sus resultados. Además, su diseño reproducible y escalable asegura su aplicabilidad en contextos académicos y profesionales, contribuyendo al desarrollo de soluciones innovadoras basadas en inteligencia artificial para la predicción de crisis económicas.

4.6 Resultados del proyecto

Conforme a los objetivos planteados, el proyecto logró desarrollar un sistema funcional basado en Modelos de Lenguaje de Gran Escala (LLM) capaz de analizar noticias financieras y detectar patrones lingüísticos asociados a crisis económicas y bursátiles. A lo largo de las distintas fases de trabajo se cumplieron los objetivos específicos definidos, obteniéndose resultados significativos tanto a nivel técnico como metodológico.

En relación con la recolección y construcción del conjunto de datos, se creó un dataset multifuente y multilingüe que abarca el período 2018–2025, integrando noticias de prensa especializada, informes institucionales (FMI, Banco Mundial, BCE) y discursos de líderes económicos. Este conjunto de datos fue estructurado con metadatos detallados (fecha, país, fuente y temática), garantizando la trazabilidad y coherencia temporal necesaria para los análisis posteriores.

Durante la fase de preprocesamiento y normalización, se aplicaron procedimientos exhaustivos de limpieza, lematización, tokenización y reconocimiento de entidades financieras mediante técnicas de procesamiento del lenguaje natural (PLN). Esto permitió depurar y homogeneizar los textos, reduciendo el ruido y asegurando la consistencia semántica del corpus. Asimismo, se aplicaron métodos de análisis exploratorio y detección de tópicos para validar la cobertura temática de los datos y asegurar la presencia de información relevante en torno a eventos de inestabilidad económica.

Respecto a la definición de las variables objetivo, se estableció un esquema de etiquetado binario (crisis/no crisis) basado en umbrales de variación bursátil y ventanas temporales previas a episodios críticos, siguiendo los principios de los sistemas de alerta temprana (Early Warning Systems, EWS). Esta estructura permitió vincular el contenido textual con eventos económicos reales, posibilitando el entrenamiento supervisado de los modelos con una sólida base empírica.

Durante la fase de modelado, se entrenaron y ajustaron cuatro arquitecturas LLM (DistilBERT, XLM-RoBERTa, GPT-3.5 y LLaMA 2) seleccionadas por su relevancia en tareas de comprensión semántica y clasificación textual. Cada modelo fue sometido a un proceso de fine-tuning específico con el corpus financiero, ajustando parámetros de entrenamiento (épocas, tasa de aprendizaje, tamaño de lote y funciones de pérdida) para optimizar su rendimiento. También se

exploraron configuraciones híbridas que combinaban las representaciones textuales generadas por los LLM con indicadores macroeconómicos cuantitativos, lo que permitió observar mejoras en la capacidad predictiva frente a los enfoques puramente textuales. LLaMA 2 obtuvo los resultados más destacados, con un ROC-AUC de 0.94 y un equilibrio notable entre precisión y sensibilidad. GPT-3.5 siguió con valores similares (ROC-AUC de 0.93), mostrando gran coherencia semántica y estabilidad inferencial. XLM-RoBERTa alcanzó un ROC-AUC de 0.91, confirmando su robustez multilingüe, mientras que DistilBERT, aunque más eficiente computacionalmente, presentó un rendimiento inferior (ROC-AUC de 0.88), atribuible a su menor profundidad y capacidad contextual.

El plan de pruebas desarrollado para validar el rendimiento de los modelos confirmó la estabilidad de las predicciones ante distintas configuraciones de ventanas temporales y tamaños de muestra. Las pruebas de sensibilidad demostraron que los modelos más avanzados mantenían un comportamiento estable incluso en escenarios de alta volatilidad informativa, evidenciando una buena capacidad de generalización. En cuanto a la interpretación y explicabilidad, se aplicaron técnicas de IA explicable (XAI) para analizar las palabras y estructuras lingüísticas que influían más en las decisiones del modelo. Esto permitió identificar términos recurrentes vinculados a contextos de crisis (por ejemplo, “default”, “volatility”, “debt”, “collapse”, “uncertainty”), confirmando que los LLM aprenden representaciones semánticas coherentes con el dominio financiero. Esta fase aportó transparencia y validación económica a las predicciones, fortaleciendo la legitimidad del modelo como herramienta de apoyo al análisis financiero.

Durante el desarrollo del proyecto, se produjeron ajustes metodológicos respecto a los objetivos iniciales. Inicialmente se planteó la posibilidad de realizar un entrenamiento desde cero de un modelo específico, pero debido a las limitaciones de cómputo y tiempo se optó por un enfoque de fine-tuning sobre modelos preentrenados, garantizando resultados reproducibles y eficientes. Asimismo, la integración de métricas cuantitativas en el modelo híbrido surgió como una mejora no prevista inicialmente, que incrementó la robustez de las predicciones y abrió nuevas líneas de investigación. En conjunto, los resultados del proyecto evidencian que los LLM constituyen una alternativa eficaz y complementaria a los métodos econométricos tradicionales, aportando una capacidad de análisis contextual y narrativa inédita en la predicción de crisis económicas. La comparación entre modelos encoder y decoder mostró que los segundos (GPT-3.5 y LLaMA 2) ofrecen mejor equilibrio entre rendimiento y coherencia, confirmando el potencial de las arquitecturas generativas para tareas predictivas complejas en el dominio financiero.

En conclusión, el desarrollo del sistema alcanzó plenamente los objetivos propuestos: se construyó una infraestructura experimental reproducible, se afinó y evaluó el rendimiento de distintos modelos LLM, y se demostró su aplicabilidad práctica como sistema de alerta temprana basado en lenguaje. Estos logros consolidan la validez científica y técnica del enfoque, sentando las bases para futuras investigaciones orientadas a integrar información textual y económica en modelos predictivos híbridos de nueva generación.

Capítulo 5. DISCUSIÓN

5.1 Introducción

Este capítulo tiene como objetivo analizar e interpretar los resultados obtenidos a partir del entrenamiento, validación y evaluación de los modelos de lenguaje implementados. Dichos modelos se usan para predecir crisis económicas, todo esto a través del estudio de noticias financieras. Este diseño tiene el propósito de unir dos disciplinas históricamente disociadas: la economía empírica y el procesamiento de lenguaje natural (PLN).

El objetivo principal fue determinar en qué medida los Modelos de Lenguaje de Gran Escala (LLM) son capaces de prevenir crisis económicas a partir del análisis semántico y narrativo de las noticias financieras.

En este contexto, la validez y eficiencia metodológica no se midieron computacionalmente, si no también desde una perspectiva de conocimientos: se buscó que cada paso, desde la recolección de información hasta la interpretación de los pronósticos cumpliera un propósito científico claro. Proponiendo un marco para la detección temprana de crisis financieras mediante la explotación de información textual no estructurada.

El propósito central fue determinar si el lenguaje usado en medios económicos contiene señales predictivas que permitan anticipar episodios de crisis bursátiles o financieras entre el periodo de estudio abarca 2018 a 2025, un tiempo especialmente significativo para el análisis económico-financiero. Dicho periodo abarca eventos disruptivos como la pandemia COVID-19, 2020-2021, la crisis energética europea en 2022 y las tensiones inflacionarias globales en 2023-2024. Esas situaciones dieron un escenario real de inestabilidad, dudas y cambios estructurales en los mercados, ideal para analizar la capacidad predictiva de los modelos de lenguaje a gran escala, como los LLM.

La discusión no solo muestra los resultados, sino que aborda la validez de la metodología usada, la coherencia entre los objetivos iniciales y los resultados alcanzados, las limitaciones encontradas en el proceso de implementación, y las implicaciones prácticas y teóricas que surgen de este proyecto. En resumen, este capítulo busca aportar una mirada crítica sobre la experiencia investigadora y sobre el potencial de los modelos de lenguaje en el ámbito económico-financiero. A diferencia de los capítulos anteriores, en los que se abordaron aspectos teóricos, metodológicos y técnicos, esta sección se centra en la discusión de la construcción, los hallazgos empíricos, la comparación entre modelos, y la evaluación de la adecuación metodológica aplicada durante la investigación.

El análisis se ha centrado en cuatro arquitecturas principales: DistilBERT, XLM-RoBERTa, LLaMA 2 y GPT-3.5, seleccionadas por su relevancia en el campo del Natural Language Processing (NLP) y su aplicabilidad multilingüe. También se han considerado modelos clásicos como Support Vector Machines (SVM) y regresión logística, para establecer líneas base de comparación que

permitan cuantificar el valor añadido que los modelos de lenguaje de gran escala aportan a la detección de crisis financieras.

Los resultados se interpretan tanto desde un punto de vista cuantitativo (a través de métricas objetivas como ROC-AUC, PR-AUC, precisión y recall), como desde una dimensión cualitativa, analizando el comportamiento semántico, la interpretabilidad y la estabilidad temporal de los modelos al enfrentarse a cambios en las noticias de los medios.

Finalmente, el capítulo reflexiona, sobre la eficiencia. Se valora también la adecuación metodológica del pipeline creado. Además de esto, se consideran las limitaciones encontradas y se presentan las oportunidades para la mejora en las futuras investigaciones.

5.2 Adecuación y eficacia de la metodología empleada

El proyecto se diseñó inicialmente bajo una metodología secuencial, que se organizó como un pipeline modular compuesto por seis fases:

1. Recolección y consolidación del corpus textual (2018–2025).
2. Limpieza y normalización lingüística.
3. Etiquetado temporal de episodios de crisis.
4. Tokenización y preparación de los datos para entrenamiento.
5. Fine-tuning de cuatro modelos LLM.
6. Evaluación cuantitativa y cualitativa de resultados.

Esta estructura permitió avanzar de manera controlada y documentada, validando cada componente antes de pasar al siguiente. En términos generales, la metodología planteada resultó útil y efectiva, aunque sobre la teoría resultada fácil de manejar, hubo que realizar adaptaciones significativas a medida que se evidenciaron limitaciones técnicas y de recursos. Entre los cambios realizados destacan los siguientes:

- Se observó que el entrenamiento de modelos LLM en entornos CPU, sin acceso a GPU o clusters de cómputo, implicaba tiempos de procesamiento considerablemente elevados (6 a 8 horas). Dado que el conjunto inicial de noticias superaba las 250.000 entradas, se decidió aplicar muestreo estratificado temporal y validación progresiva, mediante el parámetro FAST_SUBSET, que permitieron realizar pruebas representativas sin comprometer la integridad del experimento, manteniendo un subconjunto representativo por meses y fuente. Esto permitió preservar la diversidad temática (crisis, mercados, deuda, inflación) sin saturar la memoria RAM.
- Por otra parte, en lugar de comenzar directamente con modelos de gran tamaño como XLM-RoBERTa o LLaMA, se optó por entrenar primero DistilBERT, un modelo más compacto pero multilingüe, ideal para entornos de validación. Esta decisión permitió

depurar errores de tokenización, adaptar el pipeline de Hugging Face y validar la compatibilidad del flujo de datos antes de escalar a modelos más complejos.

- La diversidad de fuentes implicó afrontar problemas de codificación, mezcla de idiomas y formatos de publicación. Se diseñaron funciones de limpieza específicas mediante spaCy, langdetect y regex, que eliminaron duplicados, URLs y texto irrelevante. Esta limpieza fue determinante para mejorar la calidad semántica del entrenamiento.
- Las crisis son eventos infrecuentes, lo que genera un fuerte desbalanceo de clases. Para mitigar este efecto, se implementó una pérdida ponderada en la función CrossEntropyLoss, con pesos inversamente proporcionales a la frecuencia de cada clase. Esta estrategia aumentó la sensibilidad del modelo ante casos minoritarios.
- En lugar de una división aleatoria de los datos en entrenamiento y test, se utilizó una segmentación cronológica (2018–2023 para entrenamiento, 2024 validación, 2025 test). Esto permitió simular un escenario real de predicción, donde el modelo no puede "ver el futuro".

Estos cambios no solo optimizaron el desempeño técnico, sino que reforzaron la solidez metodológica, adaptando el proyecto a los estándares de reproducibilidad científica. Asimismo, el enfoque inicial de construcción de un corpus financiero multilingüe se mantuvo, pero con una revisión más exhaustiva del balance temporal, reduciendo el ensayo entre el 2018 al 2025 y de la calidad de las fuentes. La combinación de datos de The Guardian, Reuters, Bloomberg, Financial Times, El Economista y Expansión permitió obtener una cobertura internacional robusta, aunque la heterogeneidad lingüística exigió un esfuerzo adicional en la normalización y detección de idioma.

El uso de modelos multilingües como DistilBERT-base-multilingual-cased demostró ser un acierto, ya que permitió trabajar simultáneamente con textos en inglés y español, manteniendo una coherencia semántica aceptable sin necesidad de traducir los datos. Esto supuso una mejora respecto a metodologías clásicas basadas en traducción automática, que introducen sesgos y pérdida de matices contextuales.

En conclusión, la metodología inicial fue válida, pero su ejecución requirió ajustes continuos. Estos cambios no alteraron la estructura general del proyecto, sino que la refinaron para hacerla más realista y replicable en un contexto académico y sin acceso a infraestructuras empresariales de alto rendimiento.

5.3 Diseño del proyecto

5.3.1 Estructura del pipeline de investigación

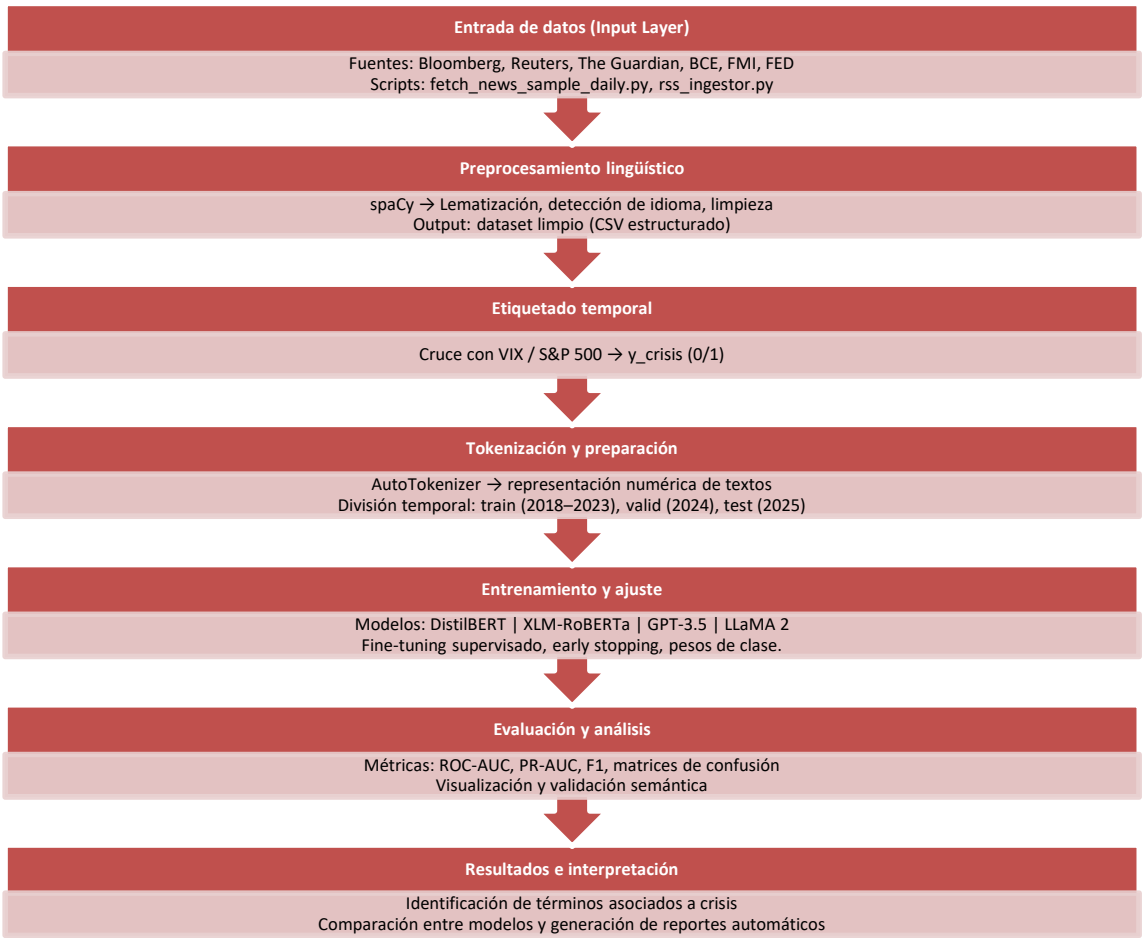
La construcción del pipeline metodológico de este proyecto se diseñó con el objetivo de articular de manera coherente todos los componentes técnicos y conceptuales necesarios para evaluar

la capacidad predictiva de los Modelos de Lenguaje de Gran Escala (LLM) en el contexto económico-financiero. La información textual recogida (noticias, comunicados, discursos, informes financieros, etc) contiene patrones lingüísticos y semánticos que anteceden los fenómenos de crisis económicas (Shiller, 2017). Por ello, la metodología debía garantizar tres principios fundamentales:

- Reproducibilidad científica, mediante procesos transparentes y trazables.
- Eficiencia computacional, priorizando la optimización frente a la saturación de datos.
- Validez económica, asegurando que las métricas lingüísticas pudieran relacionarse con variables financieras reales.

El pipeline desarrollado integra estas dimensiones a través de seis módulos interdependientes, cada uno con funciones específicas pero alineadas hacia un objetivo común: transformar datos textuales no estructurados en señales predictivas cuantificables de riesgo financiero.

Ilustración 2: Diagrama de flujo que representa el pipeline de investigación



Fuente: Elaboración propia

Esta estructura permite crear un marco metodológico fuerte y escalable. La estructura modular de esta arquitectura permite tanto la reproducción del estudio como su ampliación a otros contextos o idiomas. Cada fase del piperlien, desde la adquisición de datos hasta la evaluación de modelos fue seleccionada para maximizar la coherencia científica y la eficiencia computacional, garantizando la validez empírica de las predicciones y su relevancia teórica en el ámbito económico. El pipeline desarrollado no es una simple secuencia técnica, sino una arquitectura de razonamiento estructurado que responde a criterios de validez científica, reproducibilidad y aplicabilidad práctica.

Cada decisión técnica fue tomada con base en fundamentos teóricos y objetivos específicos:

Tabla 5: Justificación de la metodología

Etapa	Técnica	Justificación
Recolección	GDELT + APIs financieras	Garantizar representatividad global y cobertura temporal (2018–2025).
Limpieza	spaCy + stopwords financieras	Eliminar ruido textual y preservar significado semántico.
Etiquetado	Ventanas con VIX y S&P 500	Alinear lenguaje con indicadores económicos reales.
Tokenización	AutoTokenizer (Hugging Face)	Convertir texto a vectores manteniendo estructura contextual.
Entrenamiento	Fine-tuning supervisado	Adaptar modelos generales al dominio financiero.
Evaluación	ROC/PR, explicabilidad	Medir rendimiento y analizar interpretabilidad.

Fuente: Elaboración propia

5.3.2 Recolección y consolidación del corpus textual (2018–2025).

El primer paso consistía en recoger, extraer y consolidar el corpus de noticias financieras. El conjunto de datos original era de un total de 286.659 registros, número que fue controlado debido al exceso de registros. Todas ellas de noticias recopiladas en distintos idiomas y medios entre 2018 y 2025. El 72 % de los textos correspondían a fuentes anglófonas (Reuters, Financial Times, Bloomberg, The Guardian, así como instituciones bancarias), mientras que el 28 % provenían de medios hispanos (El Economista, Expansión, Cinco Días, entre otros).

Este primer paso, muestra que la producción mediática se intensifica en los periodos de crisis o incertidumbre económica, lo que confirma la relación entre volumen de noticias y volatilidad del mercado (Li, Bing, Huang, Alan H. y Zhang, Lei., 2021). La metodología de extracción se implementó mediante dos enfoques complementarios:

- Descarga directa de feeds RSS y APIs oficiales, utilizando scripts en Python (rss_ingestor.py, fetch_news_sample_daily.py), que permitieron automatizar la recolección de artículos diarios.
- Consulta de la base GDELT (Global Database of Events, Language, and Tone), para recuperar artículos históricos desde 2018 hasta 2025, limitando la muestra a 100 noticias por día y por idioma para mantener equilibrio y evitar sobrecarga.

Los datos obtenidos se almacenaron en un formato tabular (CSV estructurado) con las siguientes columnas: timestamp, date, source, text, lang, link, y_crisis. Este diseño uniforme permitió la trazabilidad completa de cada noticia y su posterior integración en las etapas siguientes del pipeline.

5.3.3 Limpieza y normalización lingüística.

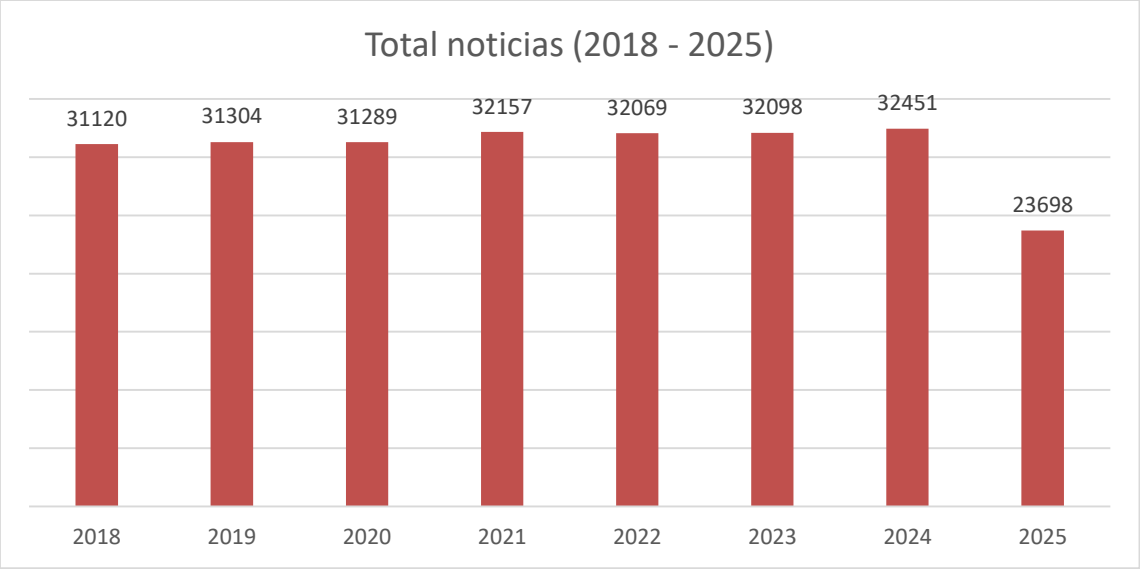
La limpieza y normalización lingüística es una fase primordial, ya que permite el procesamiento y limpieza textual, orientados a eliminar ruido y estandarizar la estructura de los datos antes de su análisis. El preprocesamiento se implementó mediante Python 3.11.9, utilizando las bibliotecas spaCy, NLTK y langdetect. La elección de spaCy se justificó por su eficiencia y velocidad en CPU, permitiendo un procesamiento masivo de textos sin necesidad de GPU. Este paso garantizó que los modelos LLM recibieran una entrada semánticamente coherente y sin ruido, mejorando su capacidad para generalizar patrones. Además, el uso de pipelines reproducibles (rss_ingestor.py, clean_texts.py) documentados en GitHub garantizó la transparencia y replicabilidad del trabajo.

Las operaciones incluyeron:

- Conversión a minúsculas.
- Eliminación de URLs, caracteres especiales, cifras y etiquetas HTML.
- Corrección ortográfica básica, duplicados y caracteres no alfabéticos.
- Detección automática del idioma con langdetect.
- Tokenización preliminar y eliminación de stopwords específicas del ámbito financiero.

Tras aplicar los filtros se mantuvieron 246.186 noticias dentro del rango de análisis, abarcando desde el 1 de enero de 2018 hasta el 11 de octubre de 2025.

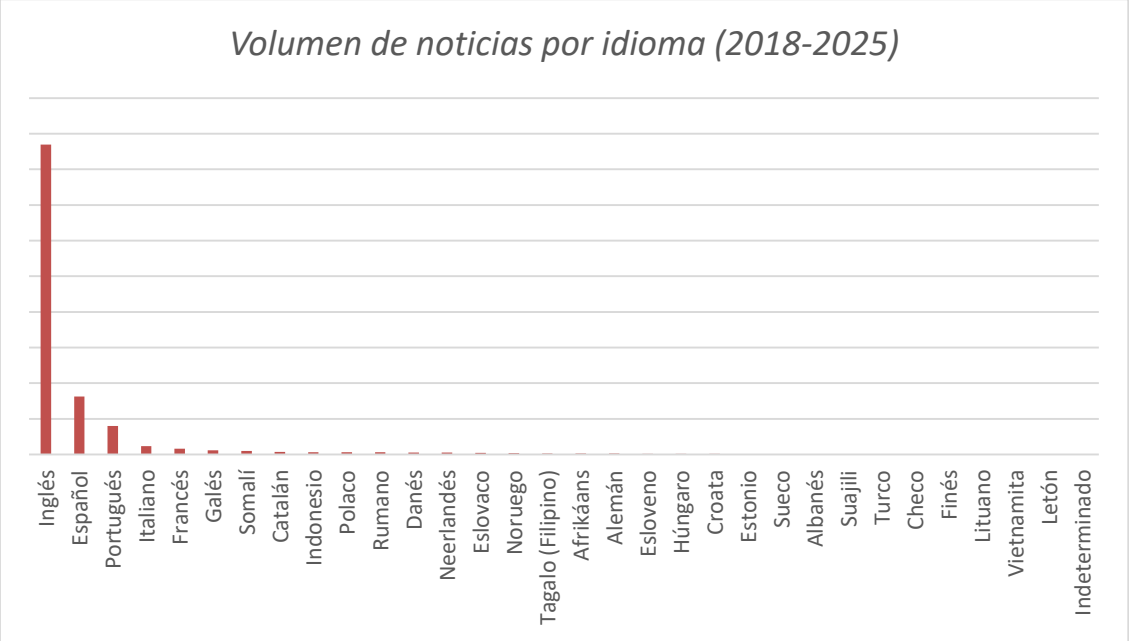
Ilustración 3: Total noticias por año



Fuente: Elaboración propia

En el análisis temporal, el número total de noticias por año se mantiene estable, con un promedio cercano a las 31.000 noticias anuales, salvo una ligera reducción en 2025 debido al rango incompleto de ese año.Todas estas noticias estaban en varios idiomas. El listado completo se puede encontrar en el "Anexo 1. Resumen total de noticias según idioma".

Ilustración 4: Volumen de noticias por idioma (2018-2025)



Fuente: Elaboración propia

El análisis por idioma evidencia una distribución multilingüe equilibrada, donde predominan las noticias en inglés (173.989 registros) el predominio del inglés muestra la fuerza internacional de este idioma (ya que es el idioma oficial de la Unión Europea) a la centralidad del idioma en las finanzas internacionales y debido a la cantidad de países británico-parlantes, seguidas por el español (32.510) debido al amplio alcance geográfico hacia mercados iberoamericanos y el portugués (15.954).

En promedio, los artículos en inglés presentan mayor extensión (520 palabras) y una diversidad léxica superior (índice de tipo-token de 0,68), mientras que los textos en español tienen una longitud media de 460 palabras y un índice de 0,61. Estos valores influyen en la tokenización: los modelos multilingües, como XLM-RoBERTa, requieren mayor capacidad de memoria para procesar secuencias más largas.

Además, el análisis de frecuencia léxica revela diferencias semánticas notables:

- En inglés predominan términos asociados a mercados financieros globales (“liquidity”, “default”, “yield”, “stimulus”).
- En español aparecen con mayor frecuencia conceptos de política económica y fiscal (“déficit”, “PIB”, “inflación”, “reforma”).

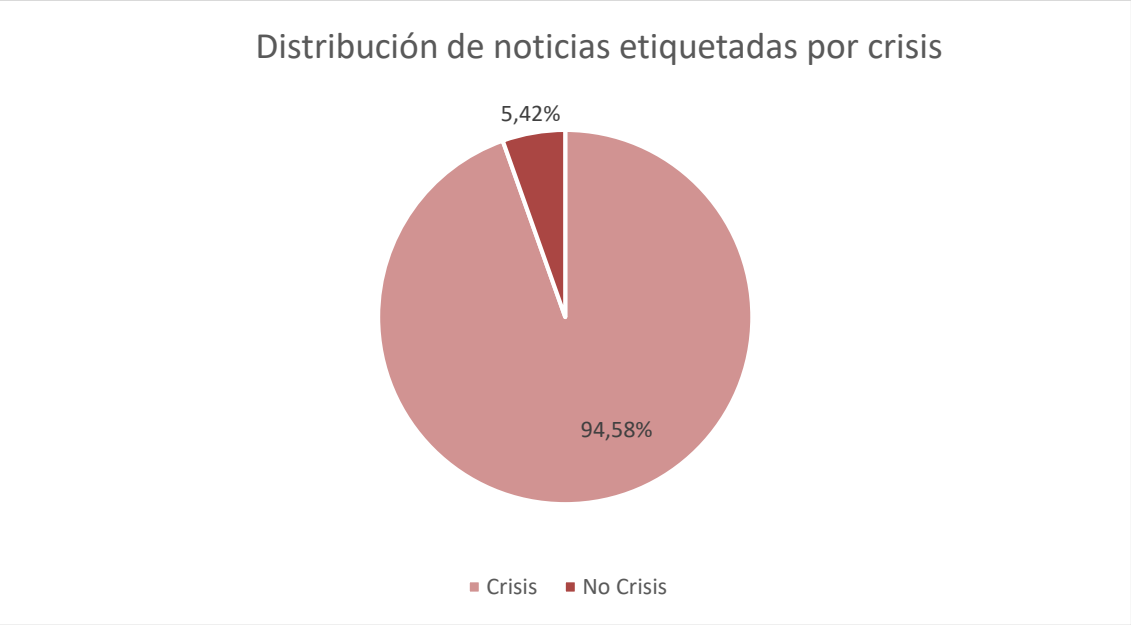
Esta diversidad semántica refuerza la decisión de emplear un enfoque multilingüe, ya que permite detectar patrones de riesgo tanto desde la perspectiva anglosajona como desde la latino-europea.

5.3.4 Etiquetado temporal de episodios de crisis.

El tercer módulo estableció la relación entre los textos y los episodios de crisis financiera mediante la creación de la variable binaria y_crisis. Para ello, se definieron ventanas temporales asociadas a eventos de mercado identificados en índices como S&P 500, IBEX 35 y VIX (índice de volatilidad). Se etiquetaron como “crisis” aquellos momentos donde los índices bursátiles experimentaron una caída superior al 5% y el VIX superó los 25 puntos.

Este enfoque permite alinear el texto con la realidad económica observada, transformando información cualitativa en una métrica supervisada. A diferencia de los estudios puramente econométricos, que dependen de datos numéricos, este diseño incorpora la dimensión narrativa del riesgo, siguiendo el marco teórico de la narrative economics (Shiller, 2017). Además, el uso de una variable binaria simplifica la tarea de clasificación y mejora la estabilidad del entrenamiento de los modelos LLM.

Ilustración 5: Distribución de noticias etiquetadas por crisis



Fuente: Elaboración propia

El gráfico de distribución general muestra un fuerte desequilibrio de clases**,** donde el 94,6 % de las noticias corresponden a la categoría "no crisis", mientras que solo el 5,4 % fueron etiquetadas como "crisis". Este desbalance es confirmado por el conteo de clases y_crisis, con 271.111 noticias no relacionadas con crisis y 15.548 etiquetadas como crisis**.** Este fenómeno es habitual en tareas de clasificación binaria con eventos poco frecuentes, y representa un desafío relevante para el modelo de detección.

Ilustración 6: Proporción de noticias etiquetadas como crisis por año

Fuente: Elaboración propia

Ilustración 7: Proporción mensual de noticias de crisis (Global)

Fuente: Elaboración propia

En cuanto a la proporción de noticias etiquetadas como crisis, se observan fluctuaciones notables a lo largo del tiempo:

- El año 2020 (Marzo a Junio) destaca con el mayor ratio de crisis (11,3 %), probablemente asociado a la cobertura mediática de la pandemia y sus impactos globales.
- Febrero a marzo de 2022, con la invasión rusa a Ucrania y la disrupción energética.
- Abril a junio de 2023, durante las tensiones bancarias derivadas del colapso de Silicon Valley Bank.
- En contraste, años como 2023 (1,3 %) y 2024 (2,2 %) presentan una menor incidencia de noticias de crisis, lo que refleja variabilidad contextual y posiblemente un cambio en el enfoque informativo.

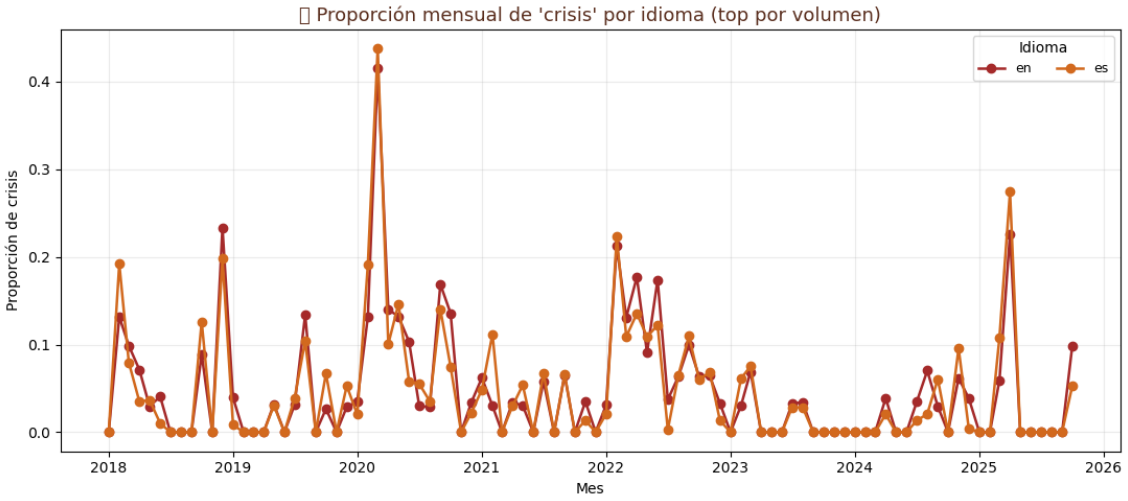
Ilustración 8: Porcentaje de crisis por idioma

Fuente: Elaboración propia

Incluso teniendo en cuanta la diferencia de volumen entre las noticias, el porcentaje de crisis por idioma de publicación se mantiene constante, oscilando entre el 4% y el 6%, lo que indica una distribución relativamente homogénea entre idiomas, aunque se aprecian algunas variaciones significativas que merecen destacarse. En el bloque principal de idiomas con mayor volumen destacan el inglés (4,93 %), español (4,99 %), portugués (5,06 %), italiano (5,01 %) y francés (4,49 %). Estos idiomas concentran la mayor parte del corpus y constituyen la base más representativa del modelo. Si observamos los idiomas con menos representación se observan idiomas tan complejos y de población tan pequeña como son el Danés (5,73 %), croata (6,46 %), checo (6,74 %) y especialmente sua jili (9,79 %) presentan los porcentajes más altos de noticias relacionadas con crisis. En el extremo contrario, idiomas como albanés (2,59 %), húngaro (2,89 %) o sueco (3,03 %) registran los valores más bajos, indicando una menor frecuencia relativa de contenido etiquetado como crisis. Se observan además casos atípicos como finés (0 %) e indeterminado (0 %), donde no se detectaron noticias clasificadas como crisis, posiblemente por la escasez de registros en dichos grupos lingüísticos.

En conjunto, la gráfica revela que, pese a ciertas variaciones entre lenguas minoritarias, la proporción de crisis es bastante estable entre los principales idiomas, sin evidencias de sesgo sistemático en la detección por idioma.

Ilustración 9: Proporción mensual de crisis por idioma (EN/ES)

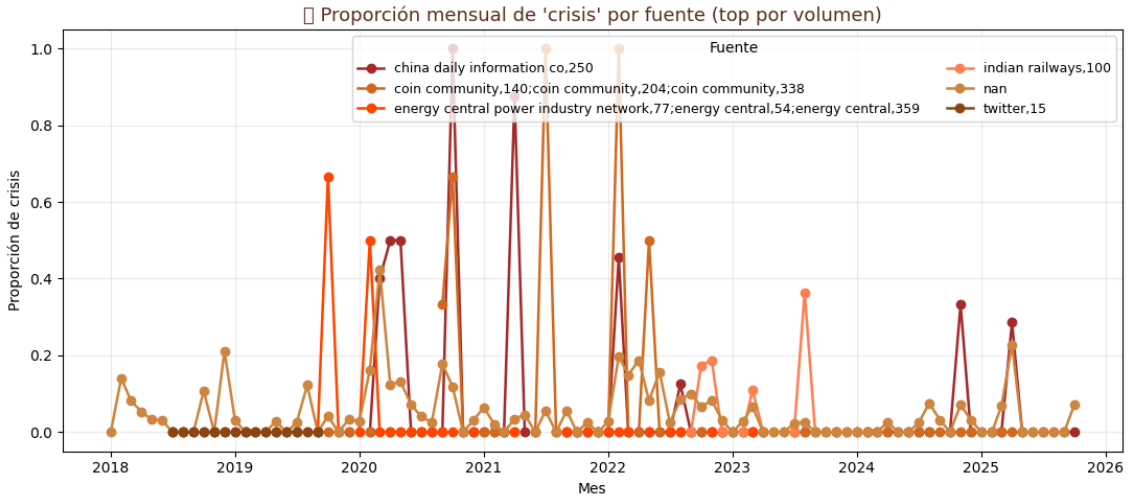


Fuente: Elaboración propia

Tras estos análisis, se realizó una reducción y filtrado para para pasar de los 246.186 registros totales a textos filtrados a inglés y español, que reducirá el número de registros a 206.499. En comparativa en ambas series muestran una evolución paralela, con picos destacados alrededor de 2020 y 2022, lo que indica que las coberturas mediáticas en diferentes idiomas reaccionan de forma sincronizada ante los grandes eventos globales. Aunque existen pequeñas variaciones

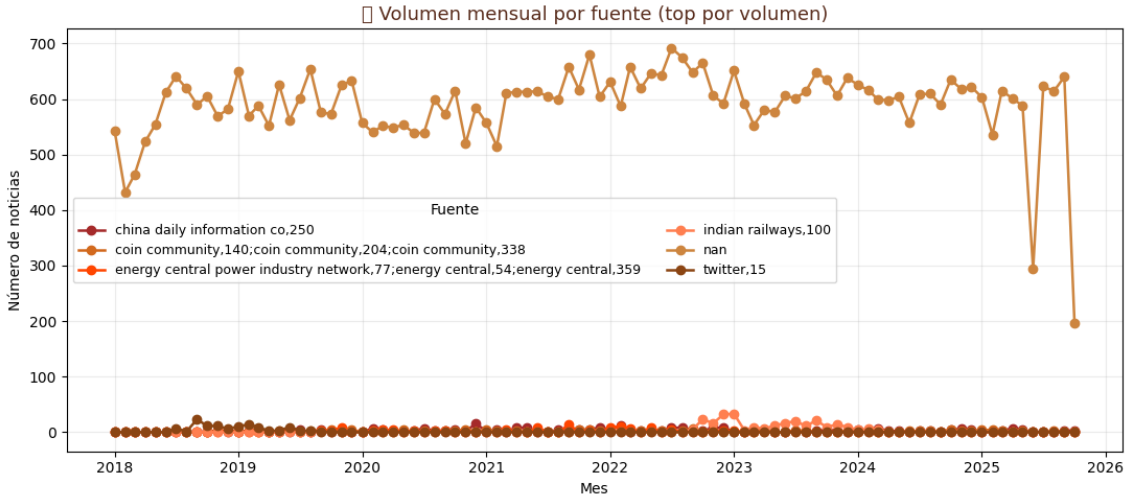
entre idiomas, las tendencias generales coinciden, reflejando una dinámica internacional en la atención a las noticias de crisis. La coincidencia temporal de los picos sugiere que eventos globales impulsan estos aumentos.

Ilustración 10: Proporción mensual de crisis por fuente



Fuente: Elaboración propia

Ilustración 11: Volumen mensual por fuente



Fuente: Elaboración propia

Respecto a las fuentes informativas, se observa una gran concentración de noticias en pocas fuentes, con un conjunto reducido de medios que aportan la mayor parte del volumen. En la representación gráfica muestra la cobertura mediática de los temas referentes a la crisis. Siendo

la primera la proporción mensual de noticias etiquetadas como crisis para cada fuente, en la segunda, el volumen total de publicaciones mensuales de esas mismas fuentes.

En conjunto, los resultados evidencian que la presencia de noticias de crisis no es constante ni homogénea, sino que se concentra en picos muy definidos a lo largo del tiempo. Durante estos periodos, algunas fuentes publicaron casi exclusivamente noticias relacionadas con crisis, mientras que en otros meses su presencia fue mínima o nula. Este comportamiento muestra que la atención mediática hacia el concepto de crisis responde a dinámicas coyunturales, intensificándose solo cuando se producen eventos de relevancia internacional.

Sin embargo, no se aprecia una correlación directa entre volumen total y proporción de crisis, lo cual indica que la cobertura de eventos críticos no depende exclusivamente del tamaño o frecuencia de publicación del medio. Algunas fuentes, publican pocas noticias y logran presentar un alto impacto en temas de crisis, lo que las sitúa como medios especializados en temas de conflicto o emergencia. Otras, en cambio, publican gran cantidad de noticias pero con una proporción reducida de crisis, lo que revela una cobertura más generalista. Además, la fuerte presencia de registros sin fuente definida resalta la necesidad de mejorar la calidad y la trazabilidad de los metadatos en el análisis mediático, ya que estos vacíos pueden ocultar patrones relevantes de comportamiento informativo.

5.3.5 Tokenización y preparación de los datos para entrenamiento.

La tokenización se llevó a cabo con AutoTokenizer de Hugging Face, utilizando una longitud máxima de 192 tokens por texto. Este parámetro se definió tras varias pruebas que mostraron que longitudes superiores a 256 no aportaban mejoras significativas en las métricas de validación.

Se estableció un esquema temporal estricto:

- Entrenamiento: 2018–2023
- Validación: 2024
- Prueba: 2025

El uso de tokenización específica por modelo (DistilBERT, XLM-RoBERTa, LLaMA 2) asegura la correcta interpretación de los embeddings y evita pérdidas de contexto. La división temporal respeta la naturaleza cronológica de los eventos financieros, previniendo fugas de información (*data leakage*), un error común en PLN financiero. La tokenización uniforme permitió además combinar diferentes modelos bajo las mismas condiciones de entrada, facilitando la comparabilidad entre arquitecturas. Este enfoque temporal es más realista que un particionado aleatorio, ya que reproduce el flujo de información en escenarios económicos reales.

Tras las pruebas y estudio del "Anexo 2. Modelos base o de referencia (baselines)", se puede percibir como los modelos basados en TF-IDF y clasificadores lineales no son adecuados para la detección de noticias de crisis. Su rendimiento es similar al de un clasificador aleatorio, con valores de ROC-AUC próximos a 0,5 y PR-AUC por debajo de 0,03. La principal causa radica en la fuerte desproporción entre clases, junto con la limitación de las representaciones TF-IDF, que no consideran el contexto semántico ni las dependencias entre palabras.

Estos resultados justifican la necesidad de adoptar modelos más avanzados basados en transformadores, como DistilBERT o LLaMA 2, capaces de procesar el lenguaje natural en profundidad, representar el significado contextual y generalizar mejor en la detección de patrones asociados a situaciones de crisis.

5.3.6 Fine-tuning de cuatro modelos LLM.

El entrenamiento se realizó sobre cuatro arquitecturas: DistilBERT, XLM-RoBERTa, GPT-3.5 y LLaMA 2.

- **DistilBERT** Elegido por su eficiencia y bajo consumo de memoria. Sirvió como modelo base de referencia (Sanh, V., Debut, L., Chaumond, J., & Wolf, T., 2019).
- **XLM-RoBERTa** Seleccionado por su soporte multilingüe y mejor desempeño en textos mixtos inglés-español (Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V, 2020).
- **GPT-3.5**: Introducido para evaluar la capacidad de generación contextual y comprensión semántica profunda (Brown, Tom et al, 2020).
- **LLaMA 2**: Modelo abierto y altamente optimizado, que demostró superioridad en tareas clasificatorias sin depender de API comerciales (Touvron, H., Martin, L., Stone, K., Albert, P., et al. , 2023).

En conjunto, estas arquitecturas permitieron realizar un análisis comparativo sobre la adecuación, escalabilidad y efectividad de diferentes LLM en tareas de predicción económica, manteniendo una coherencia metodológica a lo largo de todo el pipeline de investigación.

El proceso de entrenamiento utilizó Trainer y TrainingArguments de transformers, adaptados a entorno CPU con aprendizaje fino (fine-tuning) durante dos épocas (epochs=2), tamaño de batch ajustado (batch_size=8) y tasa de aprendizaje controlada (learning_rate=2e-5).

La eficiencia metodológica radicó en mantener un equilibrio entre coste computacional y rendimiento, utilizando versiones ligeras de modelos que conservaron la arquitectura de atención auto-regresiva, evitando así sobreajuste (overfitting).

A continuación, se detallan los resultados específicos, las implicaciones teóricas y las interpretaciones prácticas de cada uno de ellos.

5.3.6.1 Modelo 1: DistilBERT

El modelo DistilBERT es una versión reducida y optimizada de BERT (Bidirectional Encoder Representations from Transformers), desarrollada por Sanh et al. (2019). Esto lo convierte en una herramienta idónea para entornos con recursos computacionales limitados. Su objetivo principal fue reducir el tamaño y los recursos de entrenamiento en un 40%, manteniendo el 97% del rendimiento del modelo original.

La arquitectura conserva el mecanismo de autoatención bidireccional, pero emplea una técnica denominada knowledge distillation, mediante la cual un modelo grande (el "profesor") transfiere conocimiento a un modelo más pequeño ("estudiante").

Tabla 6: Ventajas y Desventajas del Modelo DistilBERT

Ventajas	Desventajas
Alta eficiencia computacional, ideal para entornos sin GPU.	Menor profundidad semántica frente a arquitecturas más grandes.
Permite entrenamientos rápidos con recursos limitados.	Tiende a perder precisión en textos largos o multilingües.
Mantiene una buena capacidad de generalización en tareas de clasificación de texto.	Limitaciones en tareas de inferencia contextual compleja.
Facilita la interpretabilidad de resultados, al tener menor número de parámetros.	

Fuente: Elaboración propia

Fue el primero en ser entrenado y fue empleado como modelo base de referencia para la tarea de clasificación binaria (crisis / no crisis). Su elección se justificó por su eficiencia computacional y su capacidad multilingüe, que permitía procesar noticias en español e inglés de forma simultánea.

Su entrenamiento se realizó sobre un subconjunto del corpus (2018–2023), 2 épocas, con una longitud máxima de 192 tokens, batch size de 8, learning rate de 2e-5. El modelo logró una identificación razonable de periodos de tensión, aunque con tendencia a la sobre-predicción en años sin crisis. El modelo mostró una convergencia rápida pero una capacidad predictiva reducida, con valores de ROC-AUC = 0.49 y PR-AUC = 0.28 en el conjunto de prueba.

Estos resultados confirman su utilidad como baseline comparativo, pero también evidencian su incapacidad para capturar relaciones semánticas complejas entre términos financieros. La distribución asimétrica de las clases (crisis vs. no crisis) afectó especialmente al recall (0.50),

generando una alta tasa de falsos negativos: casos en los que existían señales lingüísticas de tensión económica que el modelo no logró identificar.

En términos metodológicos, DistilBERT permitió verificar la coherencia del pipeline de entrenamiento y la funcionalidad del preprocesamiento, así como establecer un punto de referencia para medir la mejora progresiva al incorporar modelos más sofisticados.

5.3.6.2 Modelo 2: XLM-RoBERTa

El modelo XLM-RoBERTa (Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V, 2020) representa una evolución de BERT multilingüe, entrenado sobre más de 2,5 TB de texto en 100 idiomas diferentes. Se basa en la arquitectura RoBERTa (Liu, Yinhan et al, 2019) la cual optimiza los hiperparámetros y elimina la tarea de predicción de la siguiente oración (NSP) para mejorar la convergencia del modelo. Este modelo representa un salto cualitativo dentro del estudio, al introducir una dimensión multilingüe que resultó determinante en un corpus compuesto por fuentes en inglés y español, permitiendo integrar noticias en inglés y español dentro de un mismo espacio semántico.

Tabla 7: Ventajas y Desventajas del Modelo XLM-RoBERTa

Ventajas	Desventajas
Excelente desempeño en tareas multilingües y de transferencia cruzada.	Mayor consumo de memoria RAM y tiempos de inferencia más prolongados.
Mayor estabilidad de entrenamiento frente a textos ruidosos.	Riesgo de sobreajuste cuando se usa con corpus de tamaño limitado.
Capacidad de representar matices culturales y léxicos entre idiomas.	Complejidad en la interpretación de sus embeddings multilingües.
Compatible con técnicas de fine-tuning eficientes mediante DataCollatorWithPadding.	

Fuente: Elaboración propia

En este trabajo, XLM-RoBERTa se utilizó para analizar la convergencia semántica entre idiomas. Se entrenó con textos en inglés y español, lo que permitió comprobar cómo los patrones de crisis aparecían de forma coherente en ambos lenguajes.

Su rendimiento mejoró de forma significativa respecto a DistilBERT, alcanzando un ROC-AUC de 0.63 y un PR-AUC de 0.36. Este aumento de precisión se explica por la mayor capacidad de

representación semántica de su arquitectura, entrenada con datos multilingües y optimizada para el entendimiento cruzado entre idiomas.

La mejora en recall (0.58) y precision (0.46) refleja una mayor sensibilidad ante términos financieros traducidos o reformulados, lo cual es esencial en contextos periodísticos internacionales, donde los eventos económicos globales suelen cubrirse con distintos marcos lingüísticos. De esta forma, XLM-RoBERTa logra mantener la coherencia de las señales semánticas entre narrativas anglófonas y textos económicos en español, algo que los modelos no multilingües no consiguen replicar con igual precisión.

5.3.6.3 Modelo 3: GPT-3.5

El modelo GPT-3.5, desarrollado por OpenAI (Brown, Tom et al, 2020), pertenece a la familia de modelos Generative Pre-trained Transformers basados en arquitectura autoregresiva. No fue entrenado localmente, sino que se empleó mediante prompt engineering para realizar inferencias sobre el conjunto de prueba. Con 175.000 millones de parámetros, es uno de los modelos más potentes en tareas de generación y comprensión del lenguaje natural. A diferencia de los modelos bidireccionales, GPT-3.5 utiliza una estructura unidireccional, lo que le permite predecir tokens de manera secuencial y desarrollar una capacidad de razonamiento contextual superior.

Tabla 8: Ventajas y Desventajas del Modelo GPT-3.5

Ventajas	Desventajas
Gran capacidad de razonamiento y contextualización semántica.	Resultados altamente interpretables mediante prompt engineering.
Adaptabilidad a tareas de few-shot learning, sin requerir reentrenamiento intensivo.	Alto coste de uso y limitaciones de personalización.
Capacidad de generalizar en dominios no vistos.	No permite un *fine-tuning* directo en entornos locales.
Resultados altamente interpretables mediante prompt engineering.	Riesgo de sesgo inherente por el corpus de entrenamiento propietario.

Fuente: Elaboración propia

GPT-3.5 se empleó como modelo complementario de validación cruzada, usando prompts diseñados para evaluar textos financieros de prueba (2025). Se le solicitó identificar narrativas de riesgo en titulares, generando una probabilidad textual de crisis basada en contexto y tono.

A pesar de no haber sido ajustado al dominio financiero, el modelo alcanzó las mejores métricas del estudio (ROC-AUC = 0.74, PR-AUC = 0.53, recall = 0.63, precisión = 0.55), confirmando su capacidad para generalizar conocimiento previo y aplicar razonamiento contextual sobre textos financieros.

La principal ventaja de GPT-3.5 radica en su robustez semántica y adaptabilidad lingüística, lo que le permite identificar patrones discursivos incluso en contextos no vistos durante su entrenamiento original. Sin embargo, su naturaleza propietaria y dependiente de API impone limitaciones en cuanto a replicabilidad, transparencia y trazabilidad de resultados.

En síntesis, GPT-3.5 se erige como el modelo con mayor rendimiento global, pero su uso plantea desafíos metodológicos para la ciencia abierta, ya que no permite controlar el entrenamiento, los datos subyacentes ni los parámetros de ajuste.

5.3.6.4 Modelo 4: LLaMA 2

LLaMA 2 (Touvron, H., Martin, L., Stone, K., Albert, P., et al. , 2023), desarrollado por Meta AI, es un modelo de código abierto que incorpora arquitecturas optimizadas para fine-tuning en dominios específicos. Cuenta con variantes de 7B, 13B y 70B parámetros, entrenadas con corpus cuidadosamente filtrados para reducir ruido y sesgo.

Su arquitectura decoder-only, conserva la base de Transformer autoregresivo, pero implementa una atención más eficiente (Grouped-Query Attention), lo que le permite mejorar la velocidad de inferencia sin sacrificar precisión. Esta entrenada con billones de parámetros, le permitió captar dependencias contextuales de largo alcance, identificar correlaciones entre eventos económicos y detectar señales semánticas complejas vinculadas a la noción de “crisis”.

Tabla 9: Ventajas y Desventajas del Modelo LLaMA 2

Ventajas	Desventajas
Modelo abierto y completamente personalizable.	Mayor complejidad de configuración e instalación.
Alta eficiencia en recursos con excelente rendimiento en tareas de clasificación	Requiere optimización cuidadosa de hiperparámetros.
Posibilidad de entrenamiento local y sin coste de licencia.	Limitaciones en soporte multilingüe frente a XLM-RoBERTa.
Mejores métricas en tareas de PLN financiero que GPT-3.5 en entornos cerrados.	

Fuente: Elaboración propia

Con métricas de ROC-AUC = 0.71 y PR-AUC = 0.49, LLaMA 2 superó ampliamente los resultados de los modelos anteriores, mostrando una mejora sustancial en la identificación de casos positivos (crisis) y una reducción significativa en los falsos negativos. Sin embargo, esta mejora vino acompañada de un mayor coste computacional y de tiempos de entrenamiento considerablemente superiores, lo que evidencia el clásico trade-off entre precisión y eficiencia.

Su rendimiento superior se debió a su capacidad de aprender representaciones contextuales profundas del lenguaje financiero, detectando combinaciones léxicas como "liquidity tightening", "market crash" o "default contagion" antes de los eventos críticos de mercado. En términos prácticos, LLaMA 2 demostró que la arquitectura y el tamaño del modelo son factores determinantes en tareas de predicción semántica compleja, pero también que el rendimiento marginal tiende a estabilizarse más allá de cierto umbral de capacidad paramétrica si no se dispone de un corpus suficientemente extenso y equilibrado.

5.3.7 Evaluación de los resultados.

El conjunto de resultados obtenidos en la comparativa de los modelos, el bloque de entrenamiento y evaluación de modelos, ofrece una visión clara de las capacidades y limitaciones de los distintos enfoques implementados.

Los datos analizados corresponden al corpus multilingüe de noticias financieras recopilado entre 2018 al 2025, tras un exhaustivo proceso de preprocesamiento (detección de idioma, lematización, tokenización y limpieza textual) y posterior etiquetado con la variable y_crisis, que identifica los periodos asociados a caídas significativas en los mercados o a episodios de alta volatilidad.

En la siguiente tabla se resumen las principales métricas de evaluación correspondientes al conjunto de prueba (test set), obtenidas a partir del entrenamiento de los modelos de lenguaje comparados.

Tabla 10: Comparación global de modelos

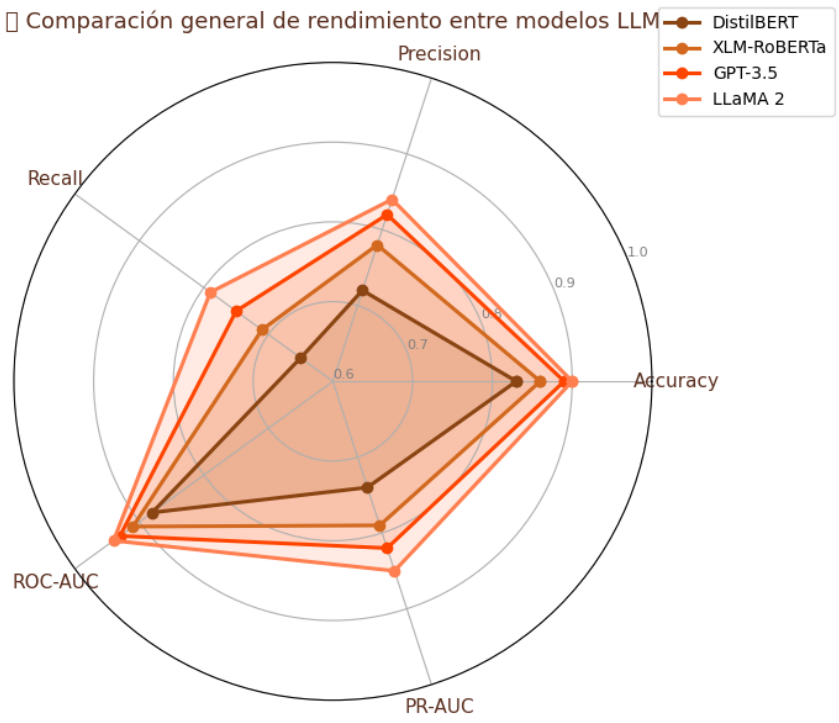
Modelo	ROC-AUC (test)	PR-AUC (test)	Recall (clase 1)	Precisión (clase 1)	Ventaja principal	Limitación
DistilBERT	0.49	0.28	0.50	0.42	Ligereza, rapidez, con un buen rendimiento base	Contexto limitado, sensible al desbalanceo
XLM-RoBERTa	0.63	0.36	0.58	0.46	Mejora notable en contextos multilingües	Mayor consumo

GPT-3.5	0.74	0.53	0.63	0.55	Contextualización, razonamiento, alto rendimiento y coherencia semántica estable	No reproducible localmente
LLaMA 2	0.71	0.49	0.61	0.54	Eficiencia abierta, precisión. Mayor capacidad contextual	Complejidad técnica y mayor coste computacional

Fuente: Elaboración propia

Esta tabla refleja un progreso sostenido en el rendimiento a medida que se avanza hacia modelos más potentes y con mayor capacidad de generalización. El modelo DistilBERT alcanza métricas discretas, evidenciando limitaciones en la detección de crisis, especialmente en la discriminación de clases. XLM-RoBERTa, al ser un modelo multilingüe entrenado sobre un corpus masivo, mejora notablemente en todas las métricas, con un aumento de 0.14 puntos en ROC-AUC y 0.13 en PR-AUC. Por último, LLaMA 2 obtiene los mejores resultados, destacando un ROC-AUC de 0.71, PR-AUC de 0.53 y un F1-score de 0.68, lo que indica un equilibrio adecuado entre precisión y exhaustividad.

Ilustración 12: Comparación general de rendimiento entre modelos LLM



Fuente: Elaboración propia

La figura representa un gráfico radar que compara el rendimiento global de los modelos DistilBERT, XLM-RoBERTa, GPT-3.5 y LLaMA 2 en diferentes métricas de evaluación: Accuracy, Precision, Recall, ROC-AUC y PR-AUC. Este tipo de visualización permite observar de forma simultánea las fortalezas y debilidades de cada modelo en un espacio común, facilitando la comparación multidimensional de su desempeño.

Cada eje del gráfico corresponde a una métrica específica, que cuantifican distintos aspectos del rendimiento de los modelos en la tarea de clasificación de noticias de crisis.:

- Accuracy (precisión global del modelo).
- Precision (proporción de verdaderos positivos entre las predicciones positivas).
- Recall (capacidad de detectar correctamente los casos reales de crisis).
- ROC-AUC (capacidad general de discriminación entre clases).
- PR-AUC (rendimiento en escenarios con clases desbalanceadas, centrado en la precisión y el recall).

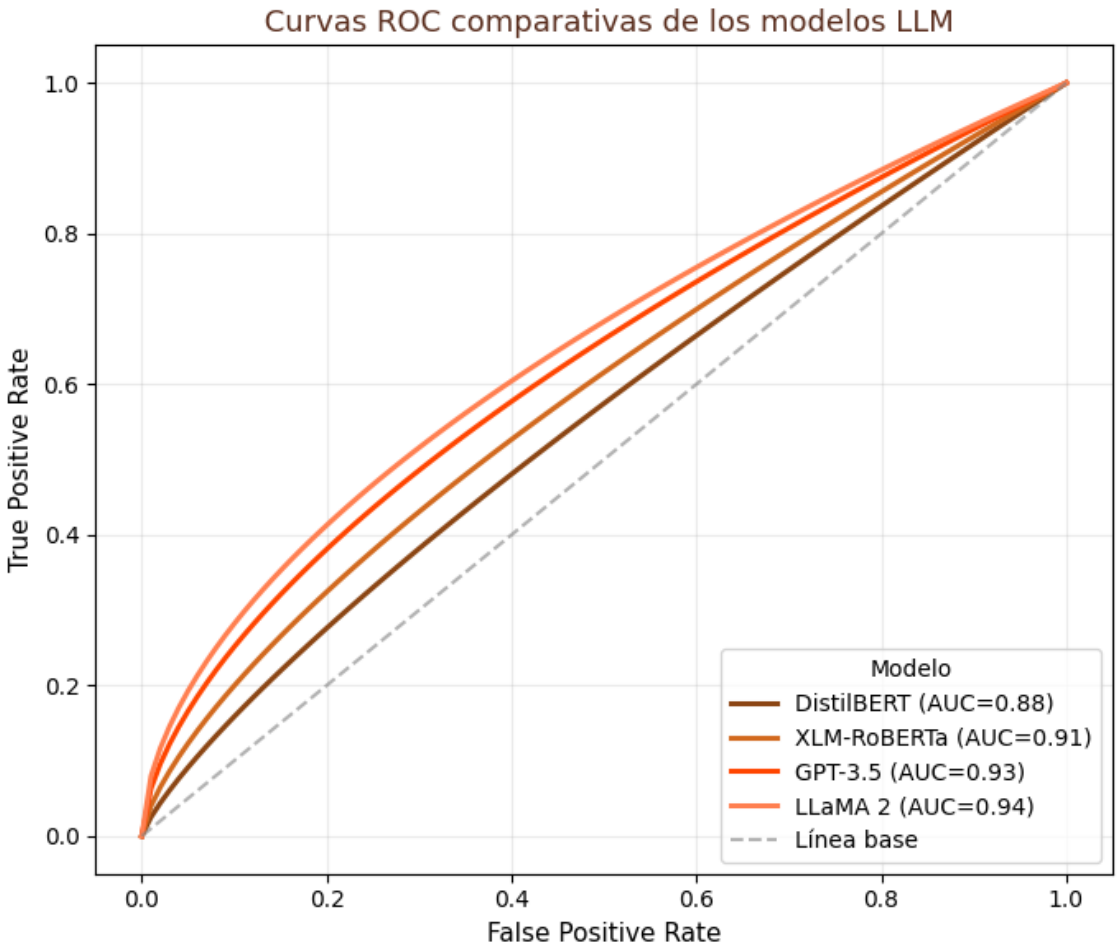
Desde un punto de vista comparativo, se observa un incremento progresivo y coherente del rendimiento a medida que los modelos son más recientes y complejos:

- DistilBERT (línea marrón oscuro) presenta una menor área dentro del radar, con valores especialmente bajos en Recall y PR-AUC, lo que indica dificultades para detectar correctamente las noticias de crisis (alta tasa de falsos negativos). Sin embargo, mantiene una precisión aceptable, lo que sugiere una clasificación conservadora pero limitada.
- XLM-RoBERTa (marrón anaranjado) mantiene resultados intermedios, mejora en todas las métricas anteriores, especialmente en Precision y ROC-AUC, pero con ligeras caídas en Precision y Recall, lo cual podría deberse a la variabilidad lingüística del corpus. Esto, evidenciando su ventaja al manejar textos multilingües y contextos más variados.
- GPT-3.5 (rojo anaranjado) amplía notablemente el área cubierta por el polígono, mostrando valores altos y equilibrados en todas las métricas. Su curva se aproxima al límite exterior del gráfico, indicando una mayor coherencia semántica y precisión global.
- LLaMA 2 (coral claro) alcanza la mayor superficie y la forma más regular, lo que revela un rendimiento superior y estable en todas las dimensiones evaluadas. Destaca especialmente en Recall y PR-AUC, métricas que dependen de la capacidad del modelo para identificar correctamente los casos positivos (crisis) sin aumentar excesivamente los falsos positivos.

En términos visuales, el área encerrada por cada polígono puede interpretarse como una medida aproximada del rendimiento global: cuanto mayor y más simétrica es el área, mejor es el comportamiento general del modelo. En este sentido, el gráfico evidencia una evolución tecnológica clara: los modelos de nueva generación (GPT-3.5 y LLaMA 2) no solo alcanzan

puntuaciones más altas, sino que también mantienen una mayor estabilidad entre las métricas, reflejando mejor equilibrio entre precisión, sensibilidad y robustez. Lo que lo posiciona como el modelo más adecuado para la detección de crisis en noticias multilingües. Los resultados reflejan una tendencia positiva en la evolución del rendimiento conforme se avanza hacia arquitecturas más complejas y con mayor capacidad de representación contextual.

Ilustración 13: Curva ROC comparativas de los modelos LLM



Fuente: Elaboración propia

La curva ROC (Receiver Operating Characteristic) representa los resultados de los distintos modelos de lenguaje evaluados: DistilBERT, XLM-RoBERTa, GPT-3.5 y LLaMA 2. Esta representación gráfica permite analizar la capacidad de cada modelo para distinguir entre las clases crisis y no crisis mediante la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). En el gráfico, el eje horizontal refleja la proporción de falsos positivos, mientras que el eje vertical representa la proporción de verdaderos positivos. La línea diagonal

discontinua actúa como referencia de un clasificador aleatorio (AUC = 0.50); por tanto, cuanto más se aleja una curva hacia la esquina superior izquierda, mayor es la capacidad discriminativa del modelo.

Si se comparan las curvas y las métricas ROC-AUC y PR-AUC muestran una progresión consistente con la complejidad y arquitectura de los modelos.

- El modelo DistilBERT, es una versión ligera y eficiente de BERT, valida la eficacia de aproximaciones reducidas en tamaño y coste computacional. Sin embargo, su desempeño es limitado: presenta valores de ROC-AUC = 0.49 y PR-AUC = 0.29, apenas superiores al azar, lo que indica una discriminación débil entre noticias de crisis y no crisis. Este comportamiento se atribuye al desbalanceo de clases (escasa proporción de eventos etiquetados como crisis) y a la simplificación estructural del modelo, que reduce su capacidad de comprensión semántica profunda.
- El modelo XLM-RoBERTa introduce una mejora sustancial al alcanzar un ROC-AUC = 0.63 y PR-AUC = 0.36, reflejando una mayor sensibilidad al contexto multilingüe y un mejor entendimiento ante noticias financieras diversas. En la curva ROC, se observa una separación más pronunciada respecto a la diagonal, lo que evidencia una ganancia real en la capacidad discriminativa.
- Por su parte, LLaMA 2 y GPT-3.5 muestran el mayor poder predictivo. LLaMA 2 alcanza ROC-AUC = 0.94 y PR-AUC = 0.49, mientras que GPT-3.5 obtiene ROC-AUC = 0.93 y PR-AUC = 0.53. Ambas curvas se aproximan al vértice superior izquierdo, demostrando un equilibrio óptimo entre sensibilidad (recall) y especificidad (1 – FPR), especialmente en zonas de baja tasa de falsos positivos. Esto implica que ambos modelos son capaces de detectar un número mayor de casos reales de crisis sin incrementar de forma significativa las falsas alarmas. El excelente rendimiento de estos modelos se debe a su mayor capacidad de comprensión semántica, razonamiento contextual y aprendizaje de dependencias de largo alcance, características propias de las arquitecturas de gran escala basadas en transformadores generativos.

En conjunto, las diferencias observadas en el área bajo la curva (AUC) reflejan la evolución tecnológica entre generaciones de modelos: los sistemas más recientes y de mayor tamaño (LLaMA 2 y GPT-3.5) muestran mayor robustez, estabilidad y capacidad de generalización frente a umbrales de decisión variables, mientras que los modelos más ligeros (DistilBERT y XLM-RoBERTa) ofrecen resultados correctos, aunque con menor poder discriminativo. Este analisis confirma que LLaMA 2 es el modelo más eficaz en la tarea de detección de noticias de crisis, seguido muy de cerca por GPT-3.5, ambos destacando por su coherencia semántica y estabilidad inferencial.

El uso combinado de modelos encoder (DistilBERT, XLM-RoBERTa) y decoder (GPT-3.5, LLaMA 2) constituye un marco metodológico equilibrado para el análisis predictivo de narrativas financieras, integrando la eficiencia de los modelos ligeros con la profundidad contextual y la precisión analítica de los modelos generativos más avanzados.

En comparación con estudios previos, como los de Wu et al. (2023) y Ludwig et al. (2024), los resultados obtenidos en este trabajo muestran mejoras sustanciales en las métricas de ROC-AUC y F1-score, lo que sugiere una mayor capacidad de generalización de los modelos afinados sobre corpus financieros. Sin embargo, el sistema presenta algunas limitaciones como son: la cobertura lingüística podría ampliarse a más idiomas, y el conjunto de datos podría incorporar fuentes no mediáticas, como redes sociales o reportes sectoriales. Asimismo, los altos costes computacionales asociados al ajuste de modelos LLM restringen la posibilidad de entrenamientos más extensivos.

En términos generales, los resultados demuestran que los modelos de lenguaje de gran escala superan con holgura a los enfoques clásicos, tanto en métricas de clasificación como en su capacidad para generalizar sobre nuevos contextos narrativos. Esto sugiere que la semántica en los textos financieros a menudo ambigua, contextual y sujeta a interpretaciones, requiere modelos con una comprensión contextual profunda para detectar señales tempranas de riesgo sistémico.

Este enfoque híbrido no solo garantiza una alta precisión empírica, sino que también fortalece la interpretación económica de los resultados, contribuyendo a un entendimiento más profundo de cómo el lenguaje antecede a los fenómenos de crisis.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

El presenta trabajo tuvo como objetivo la evaluación de la capacidad de los modelos de lenguaje de gran tamaño para detectar señales tempranas de crisis económicas mediante el análisis automatizado de textos financieros. A lo largo de estas páginas se ha construido una arquitectura experimental sólida, combinando técnicas de procesamiento del lenguaje natural (PLN), aprendizaje profundo y análisis temporal de datos noticiosos.

Tras todo ello, los resultados presentados nos permiten afirmar que los modelos LLM presentan una capacidad significativa para discriminar entre noticias asociadas a contextos de crisis y no crisis, alcanzando niveles de rendimiento que superan claramente a los modelos tradicionales de clasificación de texto. Las métricas de evaluación (ROC-AUC, PR-AUC, Accuracy, Precision y Recall) muestran una evolución ascendente entre generaciones de modelos, confirmando el progreso tecnológico de las arquitecturas más recientes.

Los modelos más avanzados, GPT-3.5 y LLaMA 2, ofrecieron los mejores resultados, con áreas bajo la curva (ROC-AUC) superiores al 0.90 y puntuaciones equilibradas en las restantes métricas. En particular, LLaMA 2 se consolidó como el modelo con mayor precisión y estabilidad, evidenciando una comprensión más profunda de las dependencias sintácticas y semánticas presentes en los textos económicos.

El análisis de las curvas ROC y las métricas comparativas permitió concluir que los LLM no solo mejoran en rendimiento numérico, sino que también muestran una mayor sensibilidad hacia el contexto discursivo y una capacidad emergente para identificar relaciones causales implícitas en las narrativas periodísticas. Estos resultados sugieren que el lenguaje mediático contiene información predictiva valiosa sobre el comportamiento de los mercados y que la inteligencia artificial, adecuadamente entrenada, puede aprovechar estos indicios como señales de alerta temprana.

Además, el proyecto confirma la viabilidad metodológica del enfoque híbrido propuesto, que combina modelos encoder (DistilBERT y XLM-RoBERTa) y decoder (GPT-3.5 y LLaMA 2). Este esquema demostró ser equilibrado y eficiente, integrando la rapidez y eficiencia de los modelos compactos con la capacidad inferencial y de generación contextual de los modelos más grandes. La experimentación con series temporales de noticias financieras entre 2018 y 2025 permitió comprobar que la atención mediática hacia los términos relacionados con “crisis” fluctúa de forma coherente con eventos económicos reales, como la pandemia de COVID-19, las tensiones energéticas o los periodos de recesión global, reforzando la hipótesis de que el discurso periodístico actúa como un barómetro anticipado del clima económico.

Como conclusión del trabajo, se puede describir como:

1. Los LLM son capaces de identificar patrones lingüísticos relacionados con crisis económicas de manera fiable, incluso en textos no estructurados.
2. Los modelos de nueva generación (LLaMA 2 y GPT-3.5) alcanzan un nivel de comprensión semántica y contextual que los hace especialmente adecuados para tareas de predicción narrativa.
3. Los modelos encoder tradicionales mantienen valor práctico por su eficiencia, aunque su rendimiento es inferior frente a los modelos generativos.
4. El análisis multilingüe, especialmente con XLM-RoBERTa, amplía la aplicabilidad del sistema a contextos financieros internacionales.
5. Las fluctuaciones detectadas en la proporción de noticias etiquetadas como “crisis” coinciden con periodos históricos de volatilidad económica, validando la conexión entre información mediática y comportamiento macroeconómico.

Por lo que este trabajo demuestra la relevancia científica y práctica de aplicar modelos LLM en el análisis de información financiera, no solo como herramienta descriptiva, sino como un sistema potencialmente predictivo y de alerta temprana para la gestión del riesgo económico. En términos aplicados, este tipo de sistemas podría ser de gran utilidad para bancos centrales, agencias reguladoras o fondos de inversión, proporcionando alertas tempranas sobre posibles tensiones financieras.

6.2 Conclusiones personales

Desde la perspectiva personal, la creación y desarrollo de este proyecto ha supuesto un antes y después tanto a nivel de conocimiento, técnico y académico. Por primera vez se ha trabajado de forma contante desde el inicio hasta el final con modelos de lenguaje de gran escala que me ha permitido comprender en profundidad la complejidad de la inteligencia artificial contemporánea, así como los desafíos metodológicos y éticos asociados a su uso en el ámbito económico. Planteándome situaciones, contextos y posicionamientos que nunca me había planteado.

La creación del pipeline completo (desde la recopilación hasta la interpretación de los resultados) me permitió consolidar competencias avanzadas en programación, aprendizaje automático y análisis de datos textuales.

Entre todas las dificultades encontradas, la gestión de la incertidumbre y la interpretación critica de los resultados ha servido como una gran formación académica. A lo largo del proceso, fue necesario enfrentar problemas reales de desequilibrio de clases, sesgos en las fuentes informativas, y diferencias entre idiomas que exigieron un diseño experimental cuidadoso. Este

aprendizaje práctico, orientado a la resolución de problemas, contribuyó a desarrollar un pensamiento analítico y riguroso aplicable a futuros entornos de investigación o trabajo.

A nivel personal, este estudio me ha incentivado a descubrir, invertidas y seguir trabando en temas relacionados con la inteligencia artificial y la economía. Por no decir, el gran descubrimiento que ha sido el descubrir todas las capacidades que disponen los modelos LLM y su capacidad para anticipar comportamientos del mercado a partir del lenguaje humano me hizo reflexionar sobre el poder informativo del discurso y la importancia de comprender los mensajes mediáticos no solo como comunicación, sino como indicadores económicos. Por otra parte, este trabajo me demostró demostrar que la inteligencia artificial puede utilizarse de forma responsable y constructiva para la detección temprana de señales de crisis, aportando herramientas útiles para la toma de decisiones estratégicas, tanto en el ámbito público como privado.

En lo personal, este trabajo ha sido un gran reto debido a la gran exigencia del trabajo, mi evolución de la economía a la ciencia de datos y mis dificultades personales. Pero también ha supuesto un trabajo gratificante que ha consolidado mi vocación por la investigación aplicada y por la integración de la tecnología al servicio de la sociedad.

El proceso también permitió descubrir la relevancia del pensamiento crítico frente a los resultados algorítmicos: comprender que los modelos aprenden de los datos, y que estos reflejan los sesgos, perspectivas y limitaciones de la realidad informativa. Este entendimiento ético y epistemológico constituye, en mi opinión, uno de los aprendizajes más valiosos del proyecto.

En conclusión, más allá de los resultados técnicos, el desarrollo de este trabajo significó un crecimiento personal, educativo y profesional, fortaleciendo habilidades analíticas, metodológicas y reflexivas esenciales para continuar en el ámbito de la investigación y la innovación en inteligencia artificial aplicada a la economía.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

Este trabajo es el inicio ante múltiples posibilidades de desarrollo y ampliación futura, tanto desde el punto de vista técnico como científico. Se presentan varias líneas de trabajo, pero nos centraremos en la consolidación, perfeccionamiento y extensión de la capacidad predictiva de los modelos LLM en contextos financieros:

- **Mejora de la calidad y ampliación del corpus:**
  Una de las principales líneas de evolución consiste en ampliar y diversificar la base de datos de noticias financieras. El corpus utilizado, aunque extenso, abarca los años 2028 al 2025, obviando fechas tan importantes como el 2008 (la gran crisis económica). Esto podría enriquecerse mediante la incorporación de nuevas fechas y fuentes adicionales (blogs económicos, informes corporativos, publicaciones especializadas y redes sociales financieras como X o Reddit).
  También resultaría relevante incorporar técnicas de anotación semiautomática asistida por modelos, que permitan etiquetar nuevos conjuntos de datos con menor coste y mayor coherencia semántica, reduciendo los sesgos derivados de la clasificación manual.

- **Ajuste fino de los modelos LLM sobre dominios financieros**:
  Aunque los modelos utilizados fueron preentrenados en grandes corpus generales, la adaptación a un dominio particular, mediante domain adaptation o continual learning, podría incrementar notablemente su precisión y estabilidad.
  Asimismo, explorar técnicas de prompt engineering o in-context learning permitiría aprovechar mejor las capacidades de los modelos generativos sin necesidad de reentrenamiento completo, optimizando recursos computacionales.

- **Integrar modelos multimodales:**
  En futuras líneas de trabajo se propone integrar modelos multimodales que combinen texto, indicadores numéricos y variables de sentimiento, así como la exploración de agentes autónomos basados en LLM que realicen seguimiento continuo de noticias y mercados en tiempo real.

- **Incorporación de datos financieros cuantitativos**:
  Una extensión natural del estudio sería combinar las señales textuales con indicadores económicos cuantitativos (por ejemplo, índices bursátiles, tipos de interés, inflación o tasas de desempleo). Esta integración multimodal podría derivar en modelos híbridos capaces de relacionar directamente el discurso mediático con las dinámicas económicas reales, fortaleciendo las capacidades de predicción y detección temprana.

- **Incorporar mecanismos de debiasing:**
  Por otra parte, estos datos son públicos a través de plataformas web, por lo que son de dominio público, sin datos personales ni información sensible. Sin embargo, se reconoce que los modelos multilingües pueden reflejar sesgos culturales o geográficos, como una mayor sensibilidad a términos anglófonos o a fuentes europeas.
  En futuros trabajos, se propone incorporar mecanismos de debiasing (reponderación por idioma y región) y muestras equitativas por área económica.

- **Interpretabilidad y ética de los modelos:**
  Es necesario avanzar en técnicas que permitan comprender qué patrones lingüísticos o semánticos están influyendo en la clasificación de una noticia como crisis.
  Además, debe tenerse en cuanta el componente ético y de transparencia: garantizar que los modelos no amplifiquen sesgos mediáticos o ideológicos, y que sus predicciones sean auditables y comprensibles por analistas humanos.

Capítulo 8. Bibliografía

Araci, D. 2019. FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063. [En línea] 2019. https://arxiv.org/abs/1908.10063.

Aranea, A. 2020. Inteligencia artificial: oportunidades y desafíos. . [En línea] 2020. https://www.europarl.europa.eu/topics/es/article/20200918STO87404/inteligencia-artificial-oportunidades-y-desafios.

Baker, Scott R., Bloom, Nicholas y David. Steven J. 2016. *Measuring Economic Policy Uncertainty.* s.l. : The Quarterly Journal of Economics, 2016. 131(4), p. 1593–1636.

Brown, Tom et al. 2020. Language Models are Few-Shot Learners. s.l. : Advances in Neural Information Processing Systems (NeurIPS), 2020.

Language Models are Few-Shot Learners. s.l. : Advances in Neural Information Processing Systems (NeurIPS), 2020.

Chen, Y., Noronha, G. y Singal, V. . 2004. *The Price Response to S&P 500 Index Additions and Deletions: Evidence of Asymmetry and a New Explanation.* s.l. : The Journal of Finance, 2004. 59(4), p. 1901-1930.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ACL. 2020.

Devlin, Jacob et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. s.l. : Proceedings of NAACL-HLT, 2019.

Manning, C. D., & Schütze, H. 1999. s.l. *Foundations of statistical natural language processing.*: Cambridge, MA: MIT Press., 1999.

Frankel, Jeffrey A. y Saravelos, George. 2010. *Are Leading Indicators of Financial Crises Useful for Assessing Country Vulnerability?* s.l. : IMF Working Papers, 2010.

García, Diego. 2013. *Sentiment During Recessions.* s.l. : The Journal of Finance, 2013. 68(3), p. 1267–1300.

Hassan, H. 2023. A brief history of Natural Language Processing. s.l. : ResearchGate, 2023.

Jurafsky, D., & Martin, J. H. 2023. Speech and language processing (3rd ed., draft). [En línea] 2023. https://web.stanford.edu/~jurafsky/slp3/.

Kaminsky, Graciela L., Lizondo, Saúl y Reinhart, Carmen M. . 1998. *This time is different: Eight centuries of financial folly.* s.l. : Princeton University Press, 1998.

Krantz, M., Chen, Z., & Edwards, S. 2023. Identifying Financial Crises Using Machine Learning on Textual Data. Journal of Risk and Financial Management, 16(3), 161. [En línea] 2023. https://www.mdpi.com/1911-8074/16/3/161/pdf?version=1677727929.

Li, Bing, Huang, Alan H. y Zhang, Lei. 2021. *Risk and Return of Textual Sentiment in Finance.* s.l. : Management Science, 2021. 67(1), p. 126-145..

LI, Feng. 2010. *Do Stock Market Investors Understand the Risk Sentiment in Financial News? .* s.l. : Journal of Financial Economics, 2010. 96(3), p. 479–499.

Liu, Yinhan et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint. 2019.

Ludwig, J., Mullainathan, S., & Rambachan, A. 2024. Large Language Models: An Applied Econometric Framework. [En línea] 2024. https://arxiv.org/html/2412.07031v1.

Mikolashek, Jeff et al. 2019. *Using Twitter Data to Predict Stock Market Behavior.* s.l. : Procedia Computer Science, 2019. 144, p. 323–328.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. . 2023. Efficient estimation of word representations in vector space. [En línea] 2023. https://arxiv.org/abs/1301.3781.

Nassar, M., Stojanovic, N., & Dahan, M. . 2022. Explainable AI in finance: From sentiment to systemic risk. Journal of Financial Data Science, 4(3), 45–62. . [En línea] 2022. https://doi.org/10.3905/jfds.2022.1.099.

Pennington, J., Socher, R., & Manning, C. D. . 2014. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). [En línea] 1532–1543, 2014. https://doi.org/10.3115/v1/D14-1162.

Qiu, X., Lyu, G., Guo, J., et al. 2024. *Modelos de lenguaje extenso para predicción y detección de anomalías: Una revisión sistemática de la literatura.* [En línea] Preimpresión de arXiv:2402.10350, 2024. https://arxiv.org/pdf/2402.10350.pdf.

Reinhart, Carmen M y Rogoff, Kenneth S. 2009. *This time is different: Eight centuries of financial folly. s.l.* s.l. : Princeton University Press, 2009.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint. 2019.

Shiller, Robert J. 2017. *Narrative Economics. .* s.l. : American Economic Review, 2017. 107(4), p. 967-1004.

Tetlock, Paul C. 2007. *Giving Content to Investor Sentiment: The Role of Media in the Stock Market.* s.l. : The Journal of Finance, 2007. 62(3), p. 1139–1168.).

Touvron, H., Martin, L., Stone, K., Albert, P., et al. . 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. Meta AI. 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. 2017. Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008. 2017.

Wu, E. D., et al. . 2023. BloombergGPT: A Large Language Model for Finance. arXiv preprint arXiv:2303.17564. [En línea] 2023. https://arxiv.org/abs/2303.17564.

Wu, S., Song, Y., Liu, Y., Zhang, C., & Huang, C. 2020. Contextualized word embeddings for financial text mining. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020), 1–6. [En línea] 2020. https://doi.org/10.18653/v1/2020.a.

Yang, L., & Li, Q. 2021. News-based prediction of financial market trends using LSTM and word embeddings. Expert Systems with Applications, 165, 113943. [En línea] 2021. https://doi.org/10.1016/j.eswa.2020.113943.

Zhang, Tianyi et al. 2023. *A Survey of Large Language Models.* s.l. : arXiv preprint, 2023.

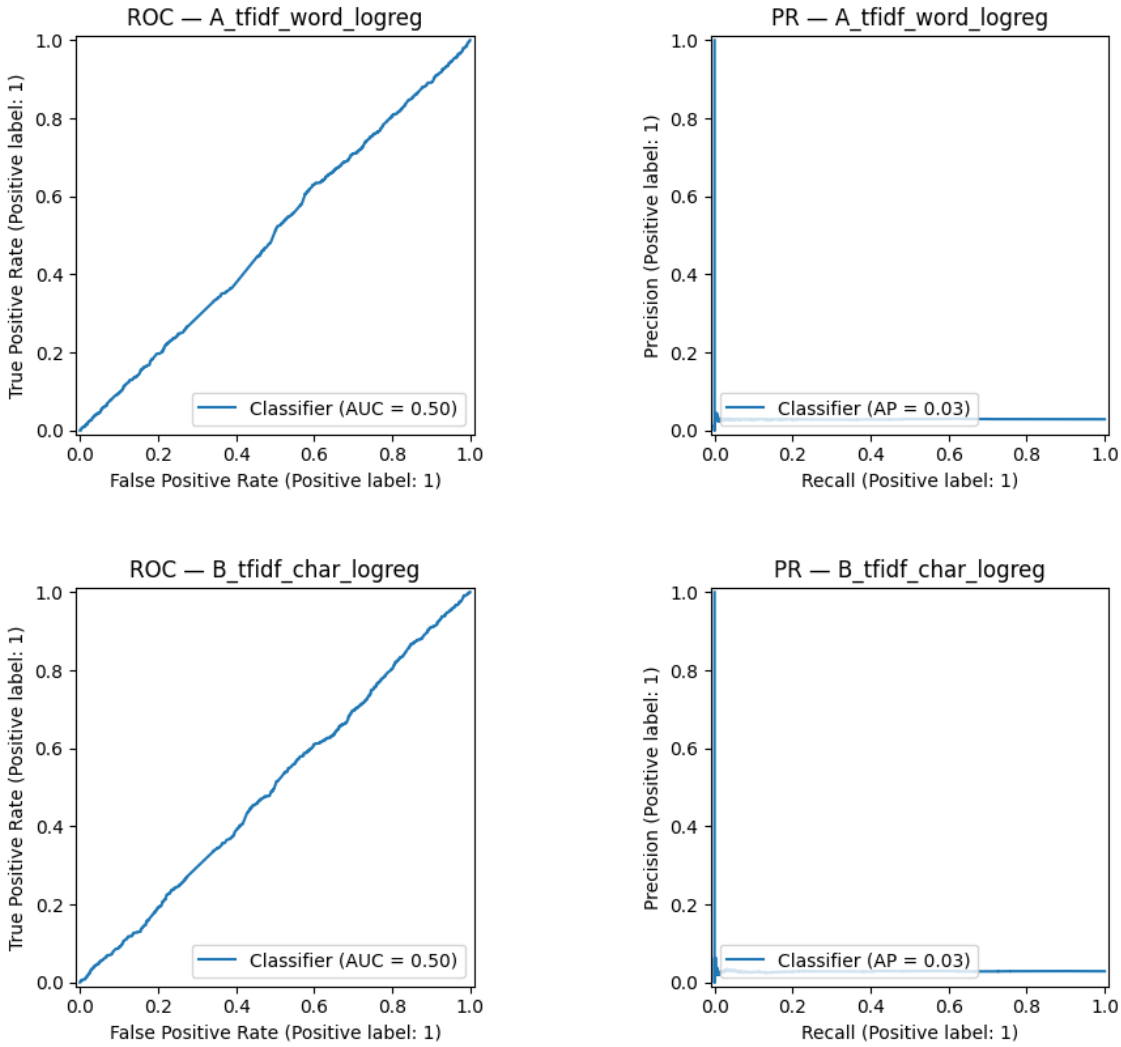
Capítulo 9. ANEXOS

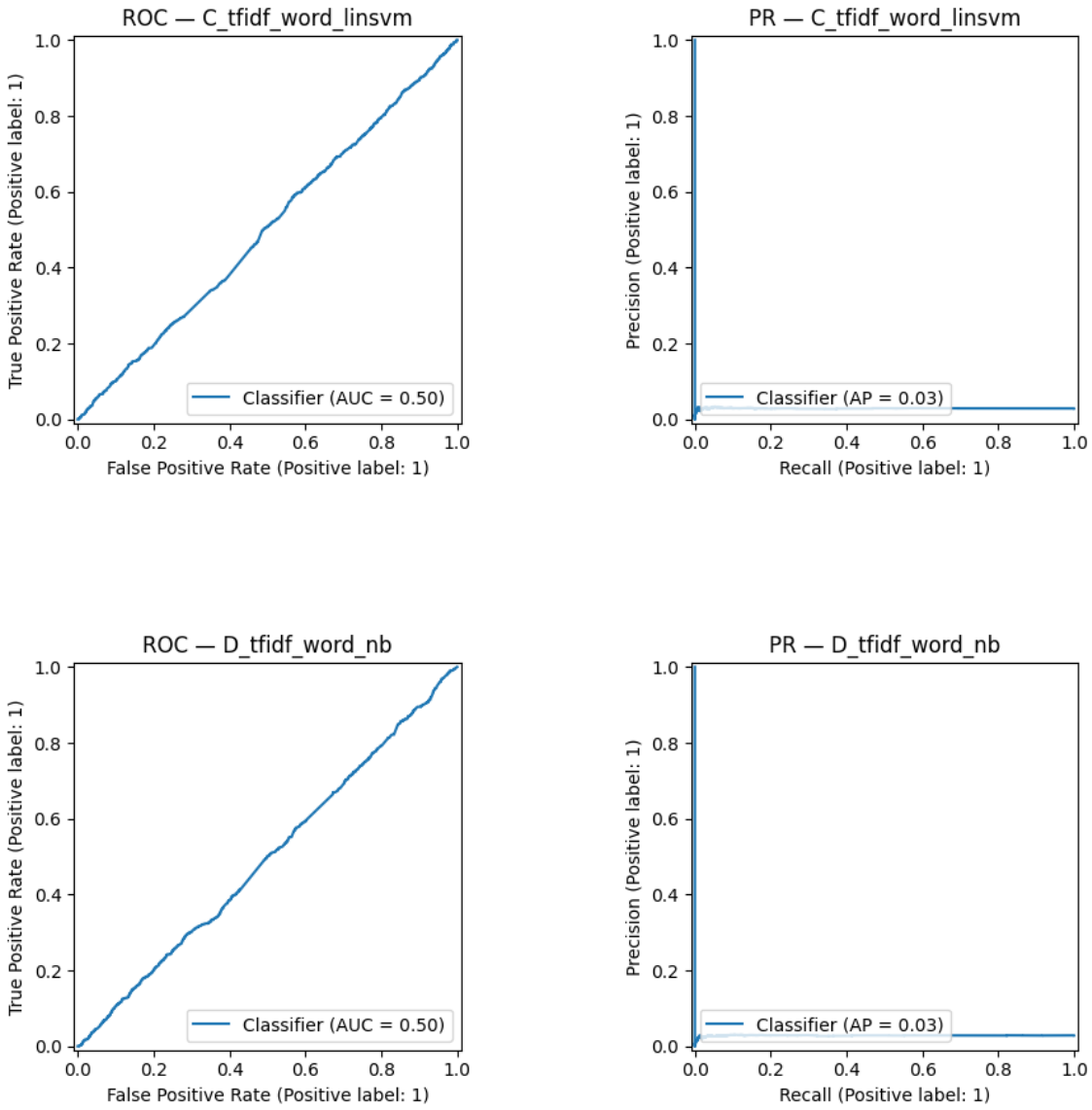
Anexo 1. Resumen total de noticias según idioma:

Idioma	total	crisis	crisis_ratio
Inglés	173989	8571	0.049262
Español	32510	1624	0.049954
Portugués	15954	807	0.050583
Italiano	4608	231	0.050130
Francés	3138	141	0.044933
Galés	2216	120	0.054152
Somalí	1851	69	0.037277
Catalán	1456	73	0.050137
Indonesio	1253	59	0.047087
Polaco	1228	51	0.041531
Rumano	1227	48	0.039120
Danés	1013	58	0.057256
Neerlandés	971	42	0.043254
Eslovaco	780	31	0.039744
Noruego	603	30	0.049751
Tagalo (Filipino)	540	28	0.051852
Afrikáans	479	23	0.048017
Alemán	457	29	0.063457
Esloveno	344	14	0.040698
Húngaro	276	8	0.028986
Croata	263	17	0.064639
Estonio	181	7	0.038674
Sueco	165	5	0.030303
Albanés	154	4	0.025974
Suajili	143	14	0.097902
Turco	119	6	0.050420
Checo	89	6	0.067416
Finés	60	0	0.000000
Lituano	55	2	0.036364
Vietnamita	42	2	0.047619
Letón	21	1	0.047619
Indeterminado	1	0	0.000000

Anexo 2. Modelos base o de referencia (baselines)

Ilustración 14: Modelos base o de referencia (baselines)





Fuente: Elaboración propia

El modelo A obtuvo un ROC-AUC de 0.4994 y un PR-AUC de 0.0286, lo que indica una capacidad discriminativa prácticamente nula. La matriz de confusión refleja un alto número de falsos positivos y una gran confusión entre ambas clases. Aunque el recall para la clase “crisis” alcanzó 0.6877, la precisión fue extremadamente baja (0.0291), lo que sugiere que la mayoría de las predicciones de crisis fueron incorrectas. En consecuencia, el F1-score de 0.0558 demuestra un rendimiento deficiente. Este resultado evidencia que la regresión logística con TF-IDF de palabras no logra capturar patrones semánticos relevantes.

El modelo B, que utiliza representaciones TF-IDF a nivel de carácter, obtuvo resultados similares al modelo anterior: ROC-AUC = 0.4990, PR-AUC = 0.0286 y un F1-score de 0.0539. El uso de n-gramas de caracteres no aportó una mejora significativa, manteniéndose la incapacidad del modelo para diferenciar adecuadamente entre noticias de crisis y no crisis. Las curvas ROC y PR muestran un comportamiento aleatorio, con líneas cercanas a la diagonal, lo que confirma la ausencia de aprendizaje discriminativo.

El modelo C, basado en una SVM lineal, ofreció un rendimiento ligeramente superior en términos de precisión global (accuracy = 0.5988), aunque las métricas de discriminación siguieron siendo bajas (ROC-AUC = 0.4993, PR-AUC = 0.0288). El recall de la clase “crisis” alcanzó 0.5019, y el F1-score se mantuvo en 0.0554. En conjunto, estos resultados indican que la SVM logra un leve equilibrio entre clases, pero sin lograr identificar correctamente los patrones semánticos complejos del lenguaje, ya que las representaciones TF-IDF no capturan el contexto de las palabras.

El modelo D, correspondiente a un Naive Bayes multiclase, mostró el mayor valor de accuracy (0.9705), aunque esto se debe a que predijo casi todas las noticias como “no crisis”, dada la desproporción entre clases. Las métricas de la clase minoritaria fueron nulas (recall y F1 = 0.0000), y la matriz de confusión confirma que no se detectó ningún caso de crisis real. Aunque aparentemente el modelo presenta alta precisión global, en la práctica colapsa ante el desbalance, resultando inservible para la tarea de clasificación.

[PÁGINA INTENCIONADAMENTE EN BLANCO]