GRADO EN BUSINESS ANALYTICS

TRABAJO FIN DE GRADO

ANÁLISIS DEL VALOR PREDICTIVO DE LAS NOTICIAS FINANCIERAS SOBRE EL NASDAQ MEDIANTE PROCESAMIENTO DE LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO

Autor: María Castilla Montes

Director: Dr. Antonio Muñoz San Roque

Subdirector: Dr. Guillermo Mestre Marcos

Madrid

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

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

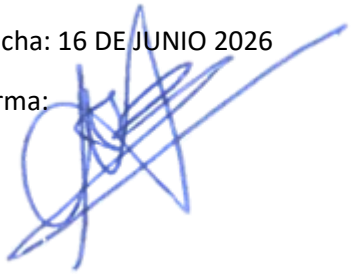
Por la presente, yo, MARIA CASTILLA MONTES, estudiante de BUSINESS ANALYTICS de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado ANÁLISIS DEL VALOR PREDICTIVO DE LAS NOTICIAS FINANCIERAS SOBRE EL NASDAQ MEDIANTE PROCESAMIENTO DE LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación [el alumno debe mantener solo aquellas en las que se ha usado ChatGPT o similares y borrar el resto. Si no se ha usado ninguna, borrar todas y escribir “no he usado ninguna”]:

1. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
2. **Interpretador de código:** Para realizar análisis de datos preliminares.
3. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
4. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
5. **Generador de encuestas:** Para diseñar cuestionarios preliminares.
6. **Traductor:** Para traducir textos de un lenguaje a otro.

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

Fecha: 16 DE JUNIO 2026

Firma:



ANÁLISIS DEL VALOR PREDICTIVO DE LAS NOTICIAS FINANCIERAS SOBRE EL NASDAQ MEDIANTE PROCESAMIENTO DE LENGUAJE NATURAL Y APRENDIZAJE AUTOMÁTICO

Autor: Castilla Montes, María.
Director: Muñoz San Roque, Antonio.

Subdirector: Mestre Marcos, Guillermo.

Entidad colaboradora: ICAI.

RESUMEN DEL PROYECTO

Este Trabajo Fin de Grado analiza si las noticias financieras, transformadas en variables cuantitativas mediante técnicas de procesamiento de lenguaje natural, pueden mejorar la predicción del comportamiento del NASDAQ frente a un modelo basado únicamente en datos históricos del índice. Para ello, se construye un dataset diario que combina variables de mercado con información extraída de noticias financieras, incluyendo sentimiento, volumen informativo, dispersión del tono, polarización y atención por compañía. Los resultados muestran mejoras moderadas en algunas métricas y configuraciones, especialmente en F1-score y en la detección de movimientos significativos del índice.

Palabras clave: NASDAQ, noticias financieras, procesamiento de lenguaje natural, FinBERT, aprendizaje automático, predicción financiera.

1. Introducción

Predecir los mercados financieros es una tarea difícil, especialmente cuando se trabaja con horizontes cortos y datos muy ruidosos. Los precios no dependen solo de su evolución histórica, sino también de expectativas, cambios de narrativa, resultados empresariales, decisiones regulatorias y noticias que modifican la percepción de los inversores. Por eso, los modelos cuantitativos tradicionales, que operan con variables como retornos pasados, volatilidad, volumen o medias móviles, dejan fuera una parte esencial: la información no estructurada, como noticias, informes o discursos, que también influyen en la dinámica del mercado.

A partir de esta idea, el trabajo analiza si las noticias financieras pueden transformarse en variables que aporten señal para explicar o anticipar el comportamiento del NASDAQ. El objetivo no es demostrar que las noticias predicen el mercado por sí solas, sino comprobar si añaden algo de información frente a un modelo basado únicamente en la evolución histórica del índice.

El NASDAQ resulta especialmente interesante para este análisis por su composición sectorial, dominada por compañías tecnológicas. En este tipo de empresas, la valoración depende en gran medida de las expectativas sobre beneficios futuros, innovación, regulación o evolución del ciclo tecnológico. Por ello, los cambios en la cobertura

informativa, el tono de las noticias o la atención sobre determinadas compañías pueden ser relevantes para entender el comportamiento agregado del índice.

2. Definición del proyecto

El objetivo del trabajo es comparar dos enfoques predictivos. El primero, denominado NASDAQ-only, utiliza exclusivamente variables de mercado: retornos pasados, volatilidad, volumen, medias móviles, *momentum* y variables de régimen. El segundo, denominado NASDAQ + News, parte de esa misma información de mercado, pero añade variables construidas a partir de noticias financieras.

La base de datos textual utilizada procede del repositorio Financial News Dataset from Bloomberg and Reuters, publicado por Philippe Remy y Xiao Ding, que recopila noticias financieras procedentes de Bloomberg y Reuters. A partir de esta base de datos se realiza un proceso de limpieza, filtrado y agregación diaria para convertir texto no estructurado en variables que puedan utilizarse en modelos de aprendizaje automático.

El proyecto se organiza en torno a tres tareas: predicción direccional del índice, detección de movimientos significativos y estimación del retorno futuro. De esta forma, el análisis no se limita a estudiar si el NASDAQ sube o baja, sino que también considera la intensidad del movimiento y la rentabilidad esperada en distintos horizontes temporales.

3. Descripción de la metodología

La metodología se desarrolla en varias fases. Primero se procesan los datos históricos del NASDAQ y se generan indicadores técnicos de mercado. Después se limpia la base de datos de noticias y se conservan los textos adecuados para el análisis de sentimiento financiero.

Para medir el sentimiento de las noticias se utiliza FinBERT, un modelo basado en BERT y adaptado al lenguaje financiero. A diferencia de los métodos basados solo en diccionarios de palabras positivas o negativas, FinBERT tiene en cuenta el contexto lingüístico de cada texto y asigna probabilidades de sentimiento positivo, negativo y neutral. A partir de esas probabilidades se calcula una puntuación de sentimiento para cada noticia como la diferencia entre la probabilidad positiva y la negativa.

Una vez calculado el sentimiento individual de cada noticia, las variables se agregan por fecha. Así se construyen medidas como sentimiento medio, dispersión del sentimiento, fuerza del tono, polarización, ratio positivo/negativo, volumen de noticias, número de compañías mencionadas y cambios recientes en la cobertura informativa. Además, se crean variables específicas por compañía para capturar la atención sobre valores relevantes del índice, como AAPL, AMZN, MSFT, NVDA o TSLA.

La comparación entre modelos se realiza respetando el orden temporal de los datos. Esto es importante porque, en un problema financiero, el modelo debe entrenarse con información pasada y evaluarse en fechas posteriores. Se prueban modelos lineales, modelos basados en árboles, métodos de *ensemble*, *boosting* y una arquitectura híbrida de combinación. La evaluación se realiza mediante métricas de clasificación, regresión y detección de movimientos significativos.

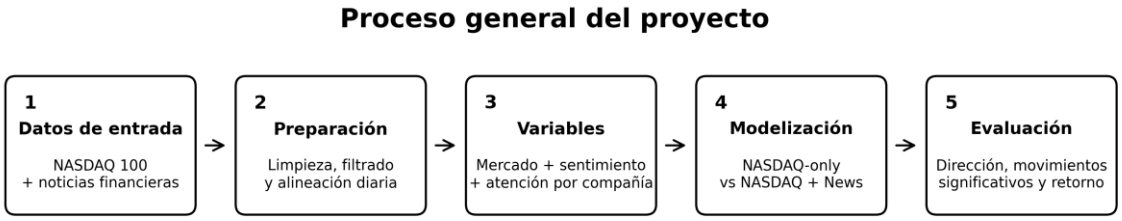


Ilustración 1- Proceso general seguido para transformar datos de mercado y noticias financieras en modelos predictivos comparables.

4. Resultados

El experimento principal se realiza con horizonte de predicción a 3 días. El conjunto de entrenamiento cubre el periodo entre septiembre de 2010 y diciembre de 2019, mientras que el periodo de prueba comprende desde enero de 2020 hasta diciembre de 2023. Esta división es exigente porque incluye años especialmente inestables, como la pandemia, la corrección tecnológica de 2022 y la recuperación posterior del mercado.

En predicción direccional, los mejores modelos se sitúan en torno al 55-57 % de accuracy. El modelo XGBoost entrenado únicamente con variables de noticias alcanza un 57,30 % de accuracy y un ROC-AUC del 54,67 %. En la comparación directa entre NASDAQ-only y NASDAQ + News, la mejora en accuracy es reducida, alrededor de 0,5 puntos porcentuales en algunos modelos. Sin embargo, el F1-score mejora de forma más clara en determinadas configuraciones: XGBoost pasa de 68,87 % a 70,36 %, e Hist. Gradient Boosting ajustado pasa de 70,10 % a 71,69 %.

En el modelo híbrido, la incorporación de noticias también mejora las métricas principales. La accuracy direccional pasa de 50,51 % a 52,47 %, el F1-score de 61,24 % a 62,86 % y el ROC-AUC de 50,57 % a 54,26 %. Además, la detección de movimientos significativos obtiene un ROC-AUC aproximado del 63,28 %, lo que indica que las variables de noticias pueden ser más útiles para identificar episodios de mayor intensidad que para anticipar todos los movimientos diarios del índice.

En regresión del retorno, los resultados son más limitados, algo esperable en una tarea de predicción financiera a corto plazo. Los valores de R^2 son cercanos a cero o negativos en la mayoría de los modelos. Aun así, algunos modelos basados únicamente en noticias obtienen una accuracy direccional derivada del retorno superior al 57 %, lo que muestra que la información textual puede ser útil para algunas tareas, aunque no sea suficiente para estimar con precisión la magnitud del retorno futuro.

5. Conclusiones

Los resultados muestran que predecir el NASDAQ sigue siendo una tarea difícil. Las métricas obtenidas no permiten hablar de una capacidad predictiva elevada ni de una solución general al problema de predicción bursátil. Sin embargo, sí permiten extraer una conclusión: las variables derivadas de noticias financieras mejoran algunos modelos y métricas frente al modelo técnico, aunque la mejora es moderada y no aparece de forma uniforme.

La principal aportación del proyecto está en haber construido un proceso completo y reproducible para transformar una base textual de gran tamaño en variables financieras utilizables por modelos predictivos. El trabajo muestra que el valor de las noticias no debe reducirse al sentimiento medio. También importan el volumen informativo, la polarización, la dispersión del tono, los cambios recientes en la cobertura y la atención sobre compañías concretas.

6. Referencias

[1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. Journal of Finance, 25(2), 383-417.

[2] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139-1168.

[3] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.

[4] Remy, P., & Ding, X. Financial News Dataset from Bloomberg and Reuters. GitHub repository: philipperemy/financial-news-dataset.

MARKET SIGNAL EXTRACTION FROM FINANCIAL NEWS: NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING APPLIED TO NASDAQ PREDICTION

Author: Castilla Montes, María.
Supervisor: Muñoz San Roque, Antonio.
Co-supervisor: Mestre Marcos, Guillermo..
Collaborating Entity: ICAI.

ABSTRACT

This Final Degree Project studies whether financial news, once transformed into quantitative variables through natural language processing techniques, can improve the prediction of NASDAQ movements compared with a model based only on historical market data. To do so, a daily dataset is built by combining market variables with information extracted from financial news, including sentiment, news volume, tone dispersion, polarization and company-level attention. The results show moderate improvements in some metrics and model configurations, particularly in F1-score and in the detection of significant market movements.

Keywords: NASDAQ, financial news, natural language processing, FinBERT, machine learning, financial prediction.

1. Introduction

Financial market prediction is a particularly complex problem. Asset prices are not driven only by their own historical behaviour, but also by expectations, changes in market narratives, corporate results, regulatory decisions and news that may affect investors' perception. Traditional quantitative models, which typically focus on metrics like past returns, volatility, trading volume, or moving averages, often miss a key source of market insights: the information contained in news, reports, and other written content.

This project starts from that idea and analyses whether financial news can be transformed into useful variables for predicting NASDAQ behaviour. The objective is not to claim that news can automatically anticipate the market, but to assess whether it adds useful information compared with a model that only uses historical data from the index.

The NASDAQ is especially relevant for this analysis due to its sector composition, with a strong presence of technology and growth-oriented companies. In these firms, valuation is highly influenced by expectations about future earnings, innovation, regulation and the evolution of the technology cycle. As a result, changes in media coverage, the tone of financial news or the attention given to specific companies may help explain the aggregate behaviour of the index.

2. Project Definition

The aim of the project is to compare two predictive approaches. The first one, called NASDAQ-only, uses only market-based variables, including past returns, volatility, volume, moving averages, momentum and market regime indicators. The second one, called NASDAQ + News, starts from the same market information but adds variables built from financial news.

The textual data used in the project comes from the Financial News Dataset from Bloomberg and Reuters, published by Philippe Remy and Xiao Ding. This dataset contains financial news from Bloomberg and Reuters. Based on this source, a process of cleaning, filtering and daily aggregation is carried out in order to transform unstructured text into variables that can be used by machine learning models.

The project is structured around three predictive tasks: directional prediction of the index, detection of significant market movements and estimation of future returns. This makes it possible to study not only whether the NASDAQ goes up or down, but also the intensity of the movement and the expected return over different time horizons.

3. Methodology

The methodology is divided into several stages. First, historical NASDAQ data is processed and technical market variables are generated. Then, the financial news dataset is cleaned and filtered, keeping the texts that are suitable for financial sentiment analysis.

To measure news sentiment, the project uses FinBERT, a BERT-based model adapted to the financial domain. Unlike methods based only on dictionaries of positive or negative words, FinBERT considers the linguistic context of each text and assigns probabilities to three sentiment classes: positive, negative and neutral. A numerical sentiment score is then calculated for each news item as the difference between the positive and negative probabilities.

Once the sentiment of each individual article is obtained, the variables are aggregated by date. This produces daily measures such as average sentiment, sentiment dispersion, sentiment strength, polarization, positive-to-negative ratio, news volume, number of companies mentioned and recent changes in media coverage. In addition, company-specific variables are created to capture attention around relevant NASDAQ constituents such as AAPL, AMZN, MSFT, NVDA and TSLA.

The comparison between models respects the chronological order of the data. This is essential in a financial prediction problem, since the model must be trained on past information and evaluated on later dates. Several types of models are tested, including linear models, tree-based models, ensemble methods, boosting algorithms and a hybrid combination architecture. The evaluation considers classification metrics, regression metrics and metrics related to the detection of significant market movements.

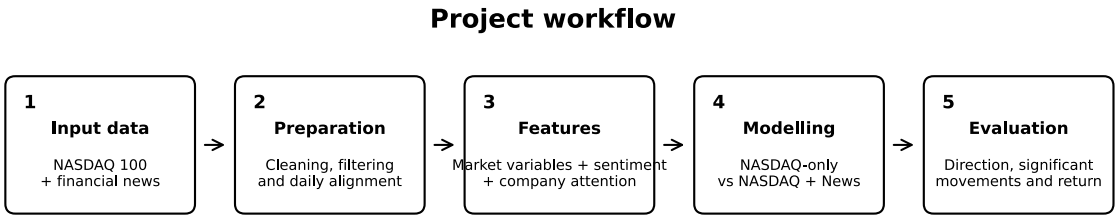


Figure 1- General workflow used to transform market data and financial news into comparable predictive models

4. Results

The main experiment is carried out using a 3-day prediction horizon. The training period covers September 2010 to December 2019, while the test period runs from January 2020 to December 2023. This is a demanding split, as the test period includes highly unstable years such as the pandemic, the 2022 technology correction and the subsequent market recovery.

In directional prediction, the best models achieve accuracy values around 55-57 %. The XGBoost model trained only with news-derived variables reaches 57.30 % accuracy and a ROC-AUC of 54.67 %. In the direct comparison between NASDAQ-only and NASDAQ + News, the improvement in accuracy is limited, around 0.5 percentage points in some models. However, the F1-score improves more clearly in certain configurations: XGBoost increases from 68.87 % to 70.36 %, and adjusted Hist. Gradient Boosting increases from 70.10 % to 71.69 %.

In the hybrid model, the inclusion of news also improves the main metrics. Directional accuracy increases from 50.51 % to 52.47 %, F1-score from 61.24 % to 62.86 %, and ROC-AUC from 50.57 % to 54.26 %. In addition, the detection of significant market movements reaches an approximate ROC-AUC of 63.28 %, suggesting that news-derived variables may be more useful for identifying periods of higher market intensity than for anticipating every daily movement of the index.

The regression task produces more limited results, which is expected in short-term financial prediction. R^2 values are close to zero or negative for most models. Even so, some models trained only with news variables achieve a directional accuracy derived from the predicted return above 57 %. This suggests that textual information may be useful for certain tasks, although it is not enough to estimate the exact magnitude of future returns with high precision.

5. Conclusions

The results confirm that predicting the NASDAQ remains a challenging endeavor. The metrics obtained do not suggest a strong predictive capacity or a general solution to stock market prediction. However, they do support a more nuanced conclusion: variables derived from financial news can improve some models and metrics compared with the technical market model, although the improvement is moderate and does not appear consistently across all configurations.

The main contribution of the project is not only the final predictive performance, but the construction of a complete and reproducible process that transforms a large textual dataset into financial variables that can be used by predictive models. The project also shows that the value of news should not be reduced to average sentiment alone. News volume, polarization, tone dispersion, recent changes in coverage and company-level attention also provide relevant information.

The main limitations are that the improvements are small and, although additional loss-based comparisons were included, most differences are not statistically significant. In addition, daily aggregation may lose part of the immediate market reaction to certain news items. Future work could therefore include intraday data, additional statistical validation and more advanced techniques for detecting financial events or changes in market narratives.

6. References

[1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. Journal of Finance, 25(2), 383-417.

[2] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance, 62(3), 1139-1168.

[3] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.

[4] Remy, P., & Ding, X. Financial News Dataset from Bloomberg and Reuters. GitHub repository: philipperemy/financial-news-dataset.

Índice de la memoria

Índice de figuras

Índice de tablas

1. INTRODUCCIÓN

El análisis financiero tradicional suele partir de variables cuantitativas como precios, retornos, volumen, volatilidad o medias móviles. Sin embargo, la valoración de los activos depende no solo de la información ya incorporada en los precios, sino también de las expectativas sobre acontecimientos futuros. Parte de esa información aparece en forma de noticias financieras, que pueden modificar la percepción de los inversores sobre empresas, sectores o riesgos de mercado.

Este Trabajo Fin de Grado analiza si esas noticias, transformadas en variables cuantitativas mediante procesamiento de lenguaje natural, pueden mejorar la predicción del NASDAQ frente a un modelo basado únicamente en datos históricos del índice. Una parte clave del trabajo ha sido construir bien la base de datos: limpiar las noticias, extraer variables textuales, alinearlas con la serie del NASDAQ y comprobar si realmente añaden información frente a las variables técnicas.

1.1 MOTIVACIÓN DEL PROYECTO

El proyecto surge del interés por conectar el análisis financiero con técnicas de procesamiento de lenguaje natural. En finanzas se trabaja habitualmente con datos estructurados, pero mucha información del mercado aparece en formato textual. Las noticias financieras contienen contexto, tono, frecuencia e intensidad informativa, elementos que no siempre quedan reflejados de forma directa en las series históricas de precios.

El NASDAQ resulta especialmente interesante para este análisis por su composición sectorial, dominada por compañías tecnológicas. En este tipo de empresas, la valoración depende en gran medida de expectativas sobre crecimiento futuro, innovación, regulación y beneficios esperados. Por ello, noticias sobre inteligencia artificial, semiconductores,

resultados empresariales o regulación tecnológica pueden influir en la percepción de los inversores sobre el índice o sobre algunos de sus componentes principales.

El trabajo compara dos enfoques. El primero, denominado NASDAQ-only, utiliza únicamente indicadores técnicos del índice, como retornos pasados, volatilidad, medias móviles, momentum, volumen y variables de régimen. El segundo, denominado NASDAQ + News, parte de la misma base técnica, pero incorpora variables procedentes de noticias financieras: sentimiento, volumen informativo, frecuencia de menciones, polarización, dispersión del tono, intensidad y variables específicas por compañía.

En este estudio, las noticias no se reducen al sentimiento medio. También se considera cuántas noticias se publican, sobre qué compañías se concentran, cómo varía el tono informativo y si aumenta de forma repentina la atención sobre determinados temas. Así, el proyecto combina predicción financiera mediante aprendizaje automático y procesamiento de lenguaje natural aplicado a texto económico.

El problema se aborda desde tres perspectivas: predicción direccional del índice, detección de movimientos significativos y estimación del retorno futuro. El objetivo no es plantear una herramienta de inversión, sino analizar si las noticias añaden señal frente a la información que ya recoge la evolución histórica del NASDAQ.

2. ESTADO DEL ARTE

La predicción de mercados financieros se ha estudiado desde enfoques muy distintos. Durante mucho tiempo, una de las ideas centrales ha sido que los precios incorporan rápidamente la información disponible y que, por tanto, anticipar movimientos futuros de forma sistemática es especialmente difícil. Esta visión está relacionada con la hipótesis de los mercados eficientes formulada por Fama (1970), que sostiene que los precios reflejan la información disponible en cada momento. Por ello, la utilidad de las noticias financieras como fuente de información predictiva no puede darse por supuesta y debe ser evaluada empíricamente.

Sin embargo, que los mercados sean difíciles de predecir no significa que toda la información se incorpore de forma instantánea, perfecta y homogénea. En la práctica, los inversores interpretan la información de manera diferente, reaccionan con distintos tiempos y pueden verse influidos por el tono con el que se presentan los acontecimientos. En este contexto se sitúa el presente proyecto: analizar si la información textual, correctamente procesada, puede aportar información complementaria a la contenida en los datos históricos del mercado.

2.1 PREDICCIÓN FINANCIERA Y APRENDIZAJE AUTOMÁTICO

Los modelos financieros han utilizado tradicionalmente datos estructurados: precios, retornos, volumen, volatilidad, factores macroeconómicos o ratios financieros. A partir de ellos se han desarrollado enfoques econométricos, modelos de series temporales y estrategias de análisis técnico. Estos métodos tienen la ventaja de trabajar con datos limpios y directamente medibles, pero también presentan una limitación clara: suelen capturar lo que ya ha ocurrido en el mercado, no necesariamente cómo se está interpretando la información nueva.

El aprendizaje automático amplía este enfoque porque permite modelar relaciones no lineales, interacciones entre variables y patrones difíciles de recoger mediante una especificación econométrica sencilla. Trabajos como el de Gu, Kelly y Xiu (2020) muestran que los métodos de machine learning pueden aportar valor en problemas de asset pricing, precisamente porque son capaces de capturar relaciones complejas entre predictores. Aun así, el problema sigue siendo complejo, porque los mercados son ruidosos y cambian con rapidez. Lo que sí permite es justificar el uso de modelos más flexibles cuando aumenta el número de variables y las relaciones entre ellas no son necesariamente lineales.

En este TFG, esta idea se aplica a la predicción del NASDAQ. El modelo parte de variables clásicas derivadas del índice, como retornos pasados, volatilidad, medias móviles, momentum o volumen. Este bloque funciona como referencia técnica y permite medir hasta qué punto la propia serie histórica contiene información útil para anticipar la dirección, la magnitud o el retorno futuro del índice.

2.2 EL PAPEL DE LAS NOTICIAS EN LOS MERCADOS

Aunque la relación entre noticias y mercados se ha estudiado desde hace años, el NLP permite ahora trabajar con volúmenes de texto mucho mayores y de forma más sistemática. Antes, incorporar noticias a un modelo exigía leerlas manualmente, clasificarlas o trabajar con muestras reducidas. Actualmente, las técnicas de NLP permiten transformar miles o millones de textos en variables cuantitativas, lo que abre la puerta a analizar cómo se relaciona el lenguaje financiero con los movimientos del mercado.

Tetlock (2007) fue uno de los trabajos más influyentes en esta línea. Analizó el contenido de una columna financiera del Wall Street Journal y encontró una relación entre el pesimismo mediático y la presión bajista posterior en los precios, además de vínculos con el volumen negociado. La idea de este TFG no es replicar ese enfoque concreto, sino tomar como punto de partida que el tono de los medios puede medirse y compararse con la evolución del mercado.

Heston y Sinha (2017) avanzan en una dirección todavía más cercana a este proyecto. Utilizan una base de datos compuesta por más de 900.000 noticias para estudiar si las noticias financieras permiten predecir retornos futuros. Sus resultados muestran que las noticias diarias tienen capacidad predictiva en horizontes cortos, especialmente de uno o dos días, mientras que las agregaciones semanales pueden mantener efecto durante periodos más largos. Esta conclusión es importante porque refuerza una decisión metodológica del trabajo: el horizonte temporal no es un detalle menor. Una variable procedente de noticias puede perder parte de su utilidad si se mide en una ventana inadecuada.

También es relevante el trabajo de Ke, Kelly y Xiu (2019), que plantea la predicción de retornos a partir de texto con un enfoque supervisado. Su aportación principal es que no basta con usar diccionarios genéricos o contar palabras positivas y negativas; el objetivo debe ser construir representaciones textuales adaptadas al problema financiero concreto. Este punto conecta directamente con el planteamiento de este TFG: las noticias no se introducen como texto plano, sino como un bloque de variables diseñado para medir tono, volumen, intensidad, dispersión y atención informativa.

2.3 SENTIMIENTO FINANCIERO Y PROCESAMIENTO DE LENGUAJE NATURAL

El análisis de sentimiento consiste en identificar si un texto transmite una orientación positiva, negativa o neutral. En finanzas, esta tarea es más delicada que en otros dominios porque el lenguaje económico tiene significados propios. Palabras que en el lenguaje general podrían parecer negativas no siempre lo son en un contexto financiero, y viceversa. Términos relacionados con pasivos, depreciación, riesgo o pérdidas pueden tener matices distintos según el tipo de documento, el sector o la noticia concreta.

Loughran y McDonald (2011) muestran precisamente este problema al estudiar textos financieros y cuestionar el uso de diccionarios generales para medir tono en documentos corporativos. Su trabajo es importante porque justifica la necesidad de herramientas

específicas para lenguaje financiero. No es lo mismo analizar opiniones de consumidores que noticias sobre resultados empresariales, tipos de interés, regulación o revisiones de beneficios.

En esta línea aparecen modelos como FinBERT, propuesto por Araci (2019), que adapta la arquitectura BERT al dominio financiero. La ventaja de este tipo de modelos es que no se basan únicamente en la frecuencia de palabras positivas o negativas, sino que analizan el contexto lingüístico en el que aparece cada término, permitiendo una interpretación más precisa del tono financiero del texto. Para este proyecto, FinBERT resulta especialmente adecuado porque permite obtener una medida de sentimiento más alineada con el lenguaje financiero que la que se obtendría mediante modelos genéricos.

Aun así, reducir las noticias al sentimiento sería insuficiente. Que una noticia tenga un tono positivo o negativo es relevante, pero también importa cuántas noticias se publican, sobre qué compañías se concentran, si existe un aumento repentino de cobertura o si el mercado recibe mensajes contradictorios. Por eso, este trabajo entiende el bloque de noticias de forma amplia. Incluye sentimiento, pero también volumen informativo, número de menciones, dispersión del tono, polarización e intensidad informativa. Esta decisión permite estudiar no solo si el tono importa, sino también si la atención informativa contiene información útil para los modelos.

2.4 DEL TEXTO A VARIABLES PREDICTIVAS

Uno de los principales retos de este tipo de proyectos no es únicamente elegir el modelo, sino construir bien la base de datos. Las noticias deben limpiarse, normalizarse, fecharse correctamente y agregarse al mismo nivel temporal que los precios del índice. Además, hay que evitar que el modelo utilice información que no habría estado disponible en el momento de la predicción. Este punto es especialmente sensible en finanzas, porque cualquier fuga temporal puede inflar artificialmente los resultados y hacer que el modelo parezca mejor de lo que realmente es.

Por ese motivo, la evaluación debe respetar la estructura temporal de los datos. En un problema financiero no tiene sentido mezclar aleatoriamente observaciones de distintos años, porque el entrenamiento podría incorporar información de periodos posteriores. La comparación debe seguir una lógica cronológica: el modelo se entrena con información pasada y se evalúa en fechas posteriores.

Este criterio se incorpora en el diseño del TFG. La comparación entre modelos se realiza separando claramente el bloque de mercado y el bloque de noticias. El modelo NASDAQ-only utiliza exclusivamente información histórica del índice, como retornos pasados, volatilidad, medias móviles, momentum, volumen y variables de régimen. El modelo NASDAQ + News añade variables extraídas del corpus periodístico. De esta forma, la diferencia entre ambos permite evaluar de manera más limpia si las noticias mejoran la predicción respecto al modelo técnico.

2.5 *APORTACIÓN DEL PROYECTO FRENTE A TRABAJOS PREVIOS*

La mayoría de los trabajos previos se centran en una de estas dos líneas: modelos financieros basados en datos de mercado o análisis de sentimiento aplicado a textos económicos. Este TFG combina ambas partes en un proceso completo aplicado al NASDAQ. La aportación no está solo en entrenar modelos, sino en construir una comparación ordenada entre dos formas de entender la predicción: una basada únicamente en la dinámica histórica del índice y otra que incorpora información procedente de noticias financieras.

Además, el análisis no se limita a la dirección del índice, sino que incorpora también la detección de movimientos significativos y la estimación del retorno futuro. Esto permite evaluar el comportamiento del NASDAQ desde tres perspectivas complementarias: signo, intensidad del movimiento y rentabilidad esperada. Esta división es importante porque un modelo puede ser razonable anticipando la dirección, pero no la magnitud; o puede detectar mejor episodios relevantes aunque no mejore de forma clara la predicción diaria general.

El estado del arte muestra que la predicción financiera con texto es una línea de investigación activa, pero compleja. Los resultados dependen del horizonte temporal, de la calidad del dato, del tipo de texto, de la forma de agregación y del método de validación. Por eso, este proyecto adopta una posición prudente: no presupone que las noticias vayan a mejorar siempre la predicción, sino que construye un marco empírico para comprobar bajo qué condiciones pueden aportar información útil frente a un modelo basado únicamente en datos históricos del índice.

3. OBJETIVOS Y ALCANCE DEL TRABAJO

El objetivo de este trabajo es evaluar si las noticias financieras, transformadas en variables cuantitativas mediante procesamiento de lenguaje natural, mejoran la predicción del NASDAQ frente a un modelo basado únicamente en datos históricos del índice. La finalidad no es sustituir el análisis financiero tradicional, sino comprobar si las noticias financieras pueden incorporarse de forma útil a un proceso de modelización cuantitativa.

El proyecto se plantea desde una perspectiva aplicada: construir el dataset, generar variables de mercado y de noticias, definir las variables objetivo, entrenar modelos y evaluar los resultados respetando el orden temporal de los datos. En problemas financieros, esta fase es especialmente importante, porque una mala construcción del dato puede generar resultados poco fiables.

3.1 OBJETIVO GENERAL

El objetivo general del TFG es comparar dos enfoques predictivos:

- Un modelo NASDAQ-only, basado únicamente en información histórica del índice: retornos pasados, volatilidad, medias móviles, momentum, volumen y variables de régimen.
- Un modelo NASDAQ + News, que incorpora además variables derivadas de noticias financieras: sentimiento, volumen informativo, número de compañías mencionadas, intensidad, polarización, dispersión del tono y variables específicas por compañía.

Esta comparación permite analizar si las noticias aportan información útil frente al modelo técnico. Si la mejora es reducida o inestable, el resultado también es relevante, porque ayuda a delimitar el alcance real de este tipo de información en predicción financiera.

3.2 OBJETIVOS ESPECÍFICOS

Para alcanzar el objetivo general, el trabajo se divide en los siguientes objetivos específicos:

- Construir una base de datos diaria que integre información del NASDAQ y noticias financieras del periodo de estudio.
- Transformar el texto de las noticias en variables cuantitativas mediante técnicas de procesamiento de lenguaje natural.
- Generar variables de mercado a partir de la serie histórica del índice.
- Construir variables derivadas de noticias, incluyendo sentimiento, volumen informativo, dispersión, polarización, intensidad y atención por compañía.
- Definir tres tareas predictivas: predicción direccional del índice, identificación de movimientos significativos y estimación del retorno futuro.
- Entrenar y comparar distintos modelos de aprendizaje automático, respetando una separación temporal entre entrenamiento y prueba.
- Analizar las limitaciones del enfoque, especialmente las relacionadas con ruido de mercado, calidad de las noticias, horizonte temporal y estabilidad de los modelos.

3.3 ALCANCE DEL PROYECTO

El alcance del proyecto se centra en la predicción diaria del NASDAQ a partir de datos históricos del índice y noticias financieras. No se pretende construir una estrategia de inversión real ni incorporar costes de transacción, restricciones de liquidez o gestión de cartera.

La unidad de análisis es diaria: las noticias y los datos de mercado se agregan por fecha para generar observaciones comparables. Esta decisión permite trabajar con una estructura coherente con los precios diarios del índice, aunque deja fuera la reacción intradía del mercado.

El trabajo se centra en el NASDAQ como índice agregado, no en la predicción individual de cada compañía. Aunque se construyen variables específicas por empresa, el objetivo final es estudiar el comportamiento del índice en conjunto.

También es importante aclarar que el bloque de noticias no se limita al sentimiento positivo o negativo. Incluye variables de cobertura, intensidad informativa, volumen, dispersión, polarización y atención por compañía. Por tanto, el análisis considera no solo el tono de las noticias, sino también la cantidad y concentración de información disponible en cada fecha.

3.4 PREGUNTA DE INVESTIGACIÓN

La pregunta que guía el trabajo es:

"¿Aportan las noticias financieras, transformadas en variables cuantitativas mediante técnicas de procesamiento de lenguaje natural, información predictiva adicional para anticipar los movimientos del NASDAQ frente a un modelo basado únicamente en datos históricos del índice?"

De ella se derivan tres preguntas secundarias:

- ¿Mejora la incorporación de noticias la predicción de la dirección del NASDAQ?
- ¿Ayudan las variables procedentes de las noticias a identificar movimientos significativos del índice?
- ¿Aportan información útil para estimar el retorno futuro, más allá de las variables técnicas de mercado?

Estas preguntas sirven como hilo conductor entre la construcción del dataset, la modelización y la lectura final de los resultados.

4. DATOS Y CONSTRUCCIÓN DEL DATASET

La construcción del dataset ha sido una de las fases más importantes del trabajo. No se partía de una tabla preparada para modelizar, sino de dos fuentes de naturaleza distinta: una serie financiera estructurada del NASDAQ 100 y una base de datos textual de noticias financieras. El objetivo de este capítulo es explicar cómo ambas fuentes se limpiaron, transformaron y alinearon temporalmente hasta obtener una base de datos diaria preparada para modelos de aprendizaje automático.

La parte de mercado estaba formada por fechas, precios, volumen y retornos. En cambio, la base de noticias exigió un procesamiento más amplio: filtrado por idioma, control de calidad de los textos, análisis de cobertura temporal, extracción de sentimiento, agregación diaria y construcción de variables por compañía. Esta fase era necesaria para convertir texto no estructurado en variables comparables con los datos financieros.

4.1 DATOS DE MERCADO: NASDAQ 100

El bloque de mercado se construyó a partir de la evolución diaria del NASDAQ 100. La base de datos inicial del índice contiene 4.031 observaciones entre el 4 de enero de 2010 y el 12 de enero de 2026, con variables como precio ajustado, cierre, apertura, máximo, mínimo, volumen, retorno diario y retorno objetivo.

La serie muestra una tendencia creciente a largo plazo, aunque con correcciones relevantes en torno a 2020, 2022 y algunos tramos posteriores. Esta evolución permite contextualizar el problema: el índice presenta una trayectoria positiva, pero también periodos de caída y aumento del riesgo que hacen que la predicción no pueda reducirse a asumir una subida constante.

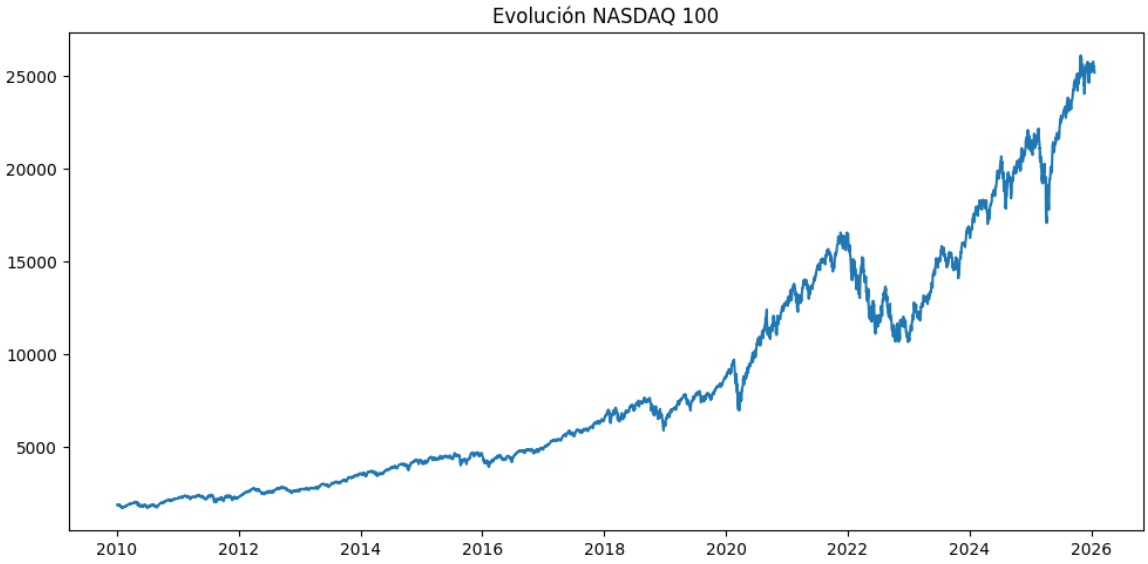


Ilustración 1- Evolución histórica del NASDAQ 100

A partir de esta serie se calcularon estadísticas descriptivas básicas. La rentabilidad media diaria fue de 0,0729 %, con una desviación típica diaria de 1,3112 %. En términos anualizados, la rentabilidad estimada fue del 18,36 %, con una volatilidad anualizada del 20,81 %. El máximo drawdown observado fue de -35,56 %. Estos valores reflejan un índice con crecimiento relevante, pero también con episodios de pérdida suficientemente intensos como para justificar un análisis predictivo más completo.

La volatilidad móvil calculada en ventanas de 30 días muestra fases de estabilidad junto a periodos de aumento acusado del riesgo, especialmente en torno a 2020. También aparecen repuntes en 2011, 2018, 2022 y 2025.

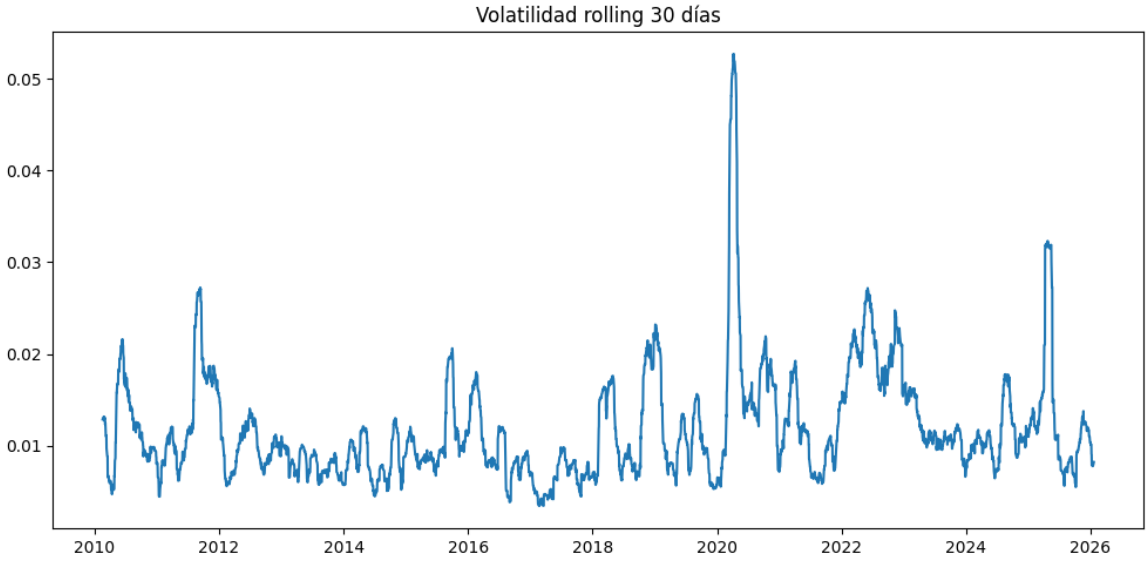


Ilustración 2- Volatilidad móvil a 30 días del NASDAQ 100.

A partir de los datos de mercado se generaron variables técnicas para el modelo NASDAQ-only: retornos pasados, volatilidad, medias móviles, distancia respecto a medias móviles, momentum, volumen y variables de régimen. Este bloque constituye la referencia frente a la que se compara el modelo que incorpora noticias financieras.

4.2 BASE DE DATOS DE NOTICIAS FINANCIERAS

La base de datos de noticias utilizada procede del repositorio Financial News Dataset from Bloomberg and Reuters, publicado por Philippe Remy y Xiao Ding. Se trata de una base de datos de noticias financieras procedentes de Bloomberg y Reuters, utilizada en este proyecto como fuente principal para construir variables relacionadas con el flujo informativo del mercado.

La base de datos original tenía un tamaño aproximado de 25 GB, por lo que fue necesario trabajar por fases. Primero se realizó un análisis exploratorio para estudiar la distribución temporal de las noticias, los idiomas presentes, la extensión de los textos, las palabras más

frecuentes y las categorías temáticas. Después se aplicaron filtros de calidad y se transformó el corpus en variables agregadas por fecha.

La distribución anual muestra una cobertura irregular. Entre 2010 y 2015 aparecen volúmenes muy altos, con años que superan el millón de noticias. A partir de 2016 el volumen desciende, con mínimos en 2021 y 2022, y una recuperación parcial en 2023. Esta irregularidad es relevante porque parte del volumen de noticias puede estar condicionado por la propia fuente de datos y no solo por el interés real del mercado.

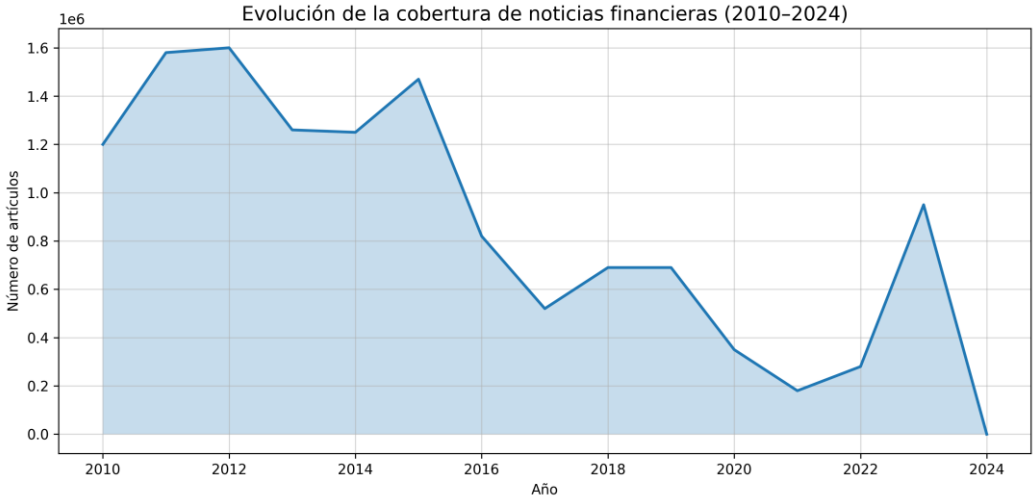


Ilustración 3- Número de noticias por año entre 2010 y 2024

El volumen diario también presenta diferencias importantes. La mayoría de los días se concentra en niveles moderados, pero existen jornadas con un flujo informativo extremo. Estos casos se identificaron a partir del percentil 99 del volumen diario. En lugar de tratarlos únicamente como anomalías, se consideraron relevantes porque pueden reflejar momentos de mayor atención del mercado.

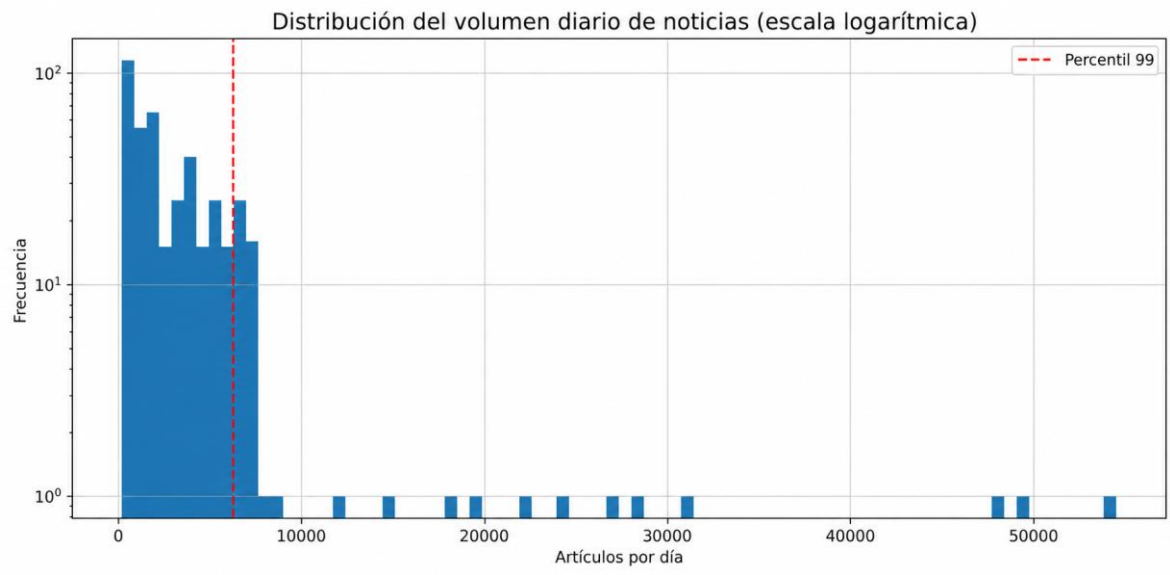


Ilustración 4- Distribución diaria del volumen de noticias

El volumen diario también presenta diferencias importantes. La mayoría de días se concentra en niveles moderados, pero existen jornadas con un flujo informativo extremo. Estos casos se identificaron a partir del percentil 99 del volumen diario. En lugar de tratarlos únicamente como anomalías, se consideraron relevantes porque pueden reflejar momentos de mayor atención del mercado.

4.3 PREPROCESADO Y FILTRADO

El preprocesado tuvo un papel decisivo en la calidad del dataset final. En primer lugar, se filtraron los textos por idioma. Como el análisis de sentimiento se realizó con un modelo financiero en inglés, se conservaron únicamente las noticias en ese idioma. Esta decisión permitió mantener coherencia entre el corpus utilizado y el modelo de procesamiento lingüístico.

Después se revisó la extensión de los textos. Los textos demasiado breves se descartaron porque no contenían suficiente información para extraer sentimiento de forma fiable. También se controlaron registros de calidad insuficiente y casos extremos que podían

introducir ruido en el análisis. Con ello se obtuvo una base de datos más coherente para la extracción de variables.

Además del filtrado básico, se estudiaron patrones temporales del flujo informativo. El volumen de noticias es mayor entre semana y cae durante el fin de semana, un patrón coherente con el funcionamiento de los mercados y de la prensa financiera. Estos análisis no se incorporan como objetivo principal del proyecto, pero ayudan a entender la estructura temporal del corpus.

4.4 CLASIFICACIÓN TEMÁTICA Y CONTENIDO DEL CORPUS

Las noticias se agruparon en varias categorías temáticas: tecnología, resultados empresariales, macroeconomía, fusiones y adquisiciones (M&A), crisis y regulación. La distribución muestra que tecnología representa el 41,3 % de la cobertura y resultados empresariales el 29,2 %. Les siguen macroeconomía, con un 10,0 %, M&A, con un 9,9 %, regulación, con un 6,2 %, y crisis, con un 3,4 %.

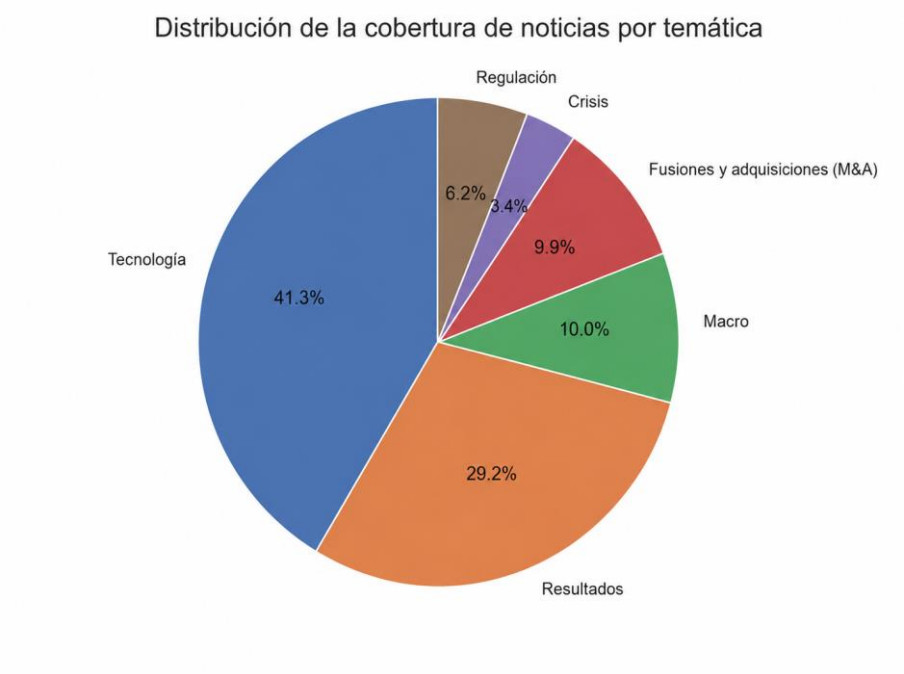


Ilustración 5- Peso relativo de cada categoría temática en la cobertura de noticias.

Esta distribución encaja con el objetivo del proyecto. El NASDAQ 100 está muy ligado a compañías tecnológicas, por lo que tiene sentido que tecnología y resultados empresariales sean las categorías dominantes. Aun así, las categorías macroeconómicas, regulatorias o de crisis también son relevantes, porque pueden afectar a la valoración del índice y a las expectativas de los inversores.

4.5 ATENCIÓN INFORMATIVA POR COMPAÑÍA

Además del análisis agregado, se construyeron variables por compañía para capturar la atención informativa sobre algunos de los principales valores del índice. Esta decisión se justifica por la propia naturaleza del NASDAQ 100: aunque el índice recoge la evolución de un conjunto amplio de compañías no financieras cotizadas en Nasdaq, su comportamiento está muy condicionado por los valores de mayor peso.

En este contexto, la atención se aproxima mediante el número de noticias asociadas a cada ticker bursátil. El análisis muestra que la cobertura no se reparte de forma uniforme: un grupo

reducido de valores concentra una parte significativa del flujo de noticias, mientras que el resto recibe una cobertura más limitada. Este patrón es coherente con la estructura del índice, donde algunas compañías tienen un peso elevado tanto en capitalización como en seguimiento mediático.

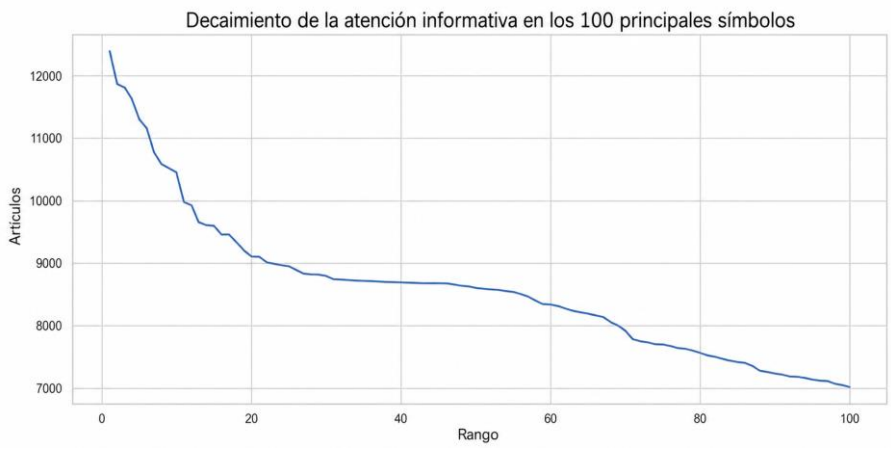


Ilustración 6- Concentración de la atención informativa por compañía.

En el dataset final se incorporaron variables específicas para compañías como AAPL, AMZN, AVGO, COST, MSFT, NVDA, PEP y TSLA. Para cada una se incluyeron medidas de sentimiento medio, volumen de noticias, fuerza del sentimiento, polarización y medias móviles.

4.6 EXTRACCIÓN DE SENTIMIENTO CON FINBERT

La extracción de sentimiento es una de las fases centrales del proyecto, ya que permite transformar cada noticia financiera en una representación numérica utilizable por los modelos predictivos. Para ello se utilizó FinBERT, un modelo basado en la arquitectura BERT y adaptado específicamente al lenguaje financiero (Araci, 2019).

La elección de FinBERT se justifica porque el lenguaje económico tiene matices propios. Un modelo generalista puede interpretar de forma incorrecta determinados términos financieros, mientras que FinBERT ha sido entrenado para distinguir mejor el tono de textos

relacionados con mercados, empresas, resultados, beneficios, pérdidas, riesgo o regulación. Por tanto, resulta más adecuado que un enfoque basado únicamente en diccionarios de palabras positivas y negativas.

El procesamiento se aplicó a nivel de noticia individual. Para cada texto, FinBERT devuelve tres probabilidades asociadas a las clases positive, negative y neutral. A partir de estas probabilidades se calculó una puntuación continua de sentimiento mediante la siguiente expresión:

$$sentimiento = P(positivo) - P(negativo)$$

Esta formulación permite convertir la salida del modelo en una variable numérica interpretable. Los valores positivos indican predominio de tono favorable, los valores negativos reflejan tono desfavorable y los valores cercanos a cero suelen corresponder a noticias neutrales o con equilibrio entre componentes positivos y negativos.

Dado el tamaño de la base de datos textual, el procesamiento se realizó por lotes. Esta decisión permitió aplicar el modelo a una base de noticias de gran tamaño sin cargar todo el corpus en memoria al mismo tiempo. Antes de aplicar FinBERT, los textos se filtraron y limpiaron para conservar únicamente noticias en inglés y con longitud suficiente para que el análisis de sentimiento fuera fiable.

El resultado de esta fase fue una base de datos enriquecida a nivel de noticia, en la que cada registro conserva su fecha y añade las probabilidades de sentimiento positivo, negativo y neutral, junto con la puntuación continua de sentimiento. Esta tabla no se utilizó directamente en los modelos, ya que la unidad de análisis del proyecto es diaria. Por ello, fue necesario agregar posteriormente las noticias por fecha.

A partir de las puntuaciones individuales se construyeron variables diarias de sentimiento. Entre ellas se incluyen el sentimiento medio, la mediana del sentimiento, la desviación típica, la fuerza del tono, la polarización y el ratio entre noticias positivas y negativas. Estas variables permiten capturar dimensiones distintas del flujo informativo. El sentimiento

medio resume el tono general del día; la dispersión mide si las noticias son homogéneas o contradictorias; la fuerza del tono recoge la intensidad del mensaje; y la polarización identifica días en los que conviven noticias con orientaciones opuestas.

De esta forma, el análisis de noticias no se limita a clasificar cada texto como positivo o negativo. El objetivo es construir un conjunto de variables que representen cómo llega la información financiera al mercado en cada fecha: con qué tono, con qué intensidad, con qué dispersión y con qué grado de concentración informativa.

4.7 CONSTRUCCIÓN DEL DATASET FINAL

Una vez extraído el sentimiento de las noticias, el siguiente paso fue construir una base de datos diaria que integrara toda la información disponible. Esta fase no consistió solo en unir tablas, sino en alinear por fecha los datos de mercado, las variables agregadas de noticias y los indicadores específicos por compañía.

El proceso se realizó de forma progresiva para mantener la trazabilidad de cada transformación. Primero se preparó la base diaria de mercado a partir del NASDAQ 100. Sobre esta serie se calcularon variables técnicas como retornos pasados, volatilidad, medias móviles, distancia respecto a dichas medias, momentum, volumen y variables de régimen. Este bloque constituye la información utilizada por el modelo NASDAQ-only.

Después se construyó el bloque agregado de noticias. La base de datos enriquecida con FinBERT se agrupó por fecha y, para cada día, se calcularon variables como volumen de noticias, sentimiento medio, mediana, dispersión, fuerza del tono, polarización, ratio positivo/negativo, número de compañías mencionadas y variaciones respecto a días anteriores. También se añadieron medias móviles para suavizar parte del ruido diario y recoger tendencias recientes del flujo informativo.

Por último, se generó un bloque específico por compañía. Se identificaron noticias asociadas a tickers relevantes del entorno NASDAQ, como AAPL, AMZN, AVGO, COST, MSFT,

NVDA, PEP y TSLA. Para cada compañía se calcularon variables de sentimiento medio, volumen de noticias, fuerza del tono, polarización y medias móviles. Esta capa permite capturar si la atención informativa se concentra en valores concretos con peso relevante dentro del índice.

El proceso dio lugar a tres tablas principales antes de la modelización:

- Una base de datos diaria de mercado y conteo inicial de noticias, con 4.031 filas y 12 columnas, que cubre el periodo entre el 4 de enero de 2010 y el 12 de enero de 2026.
- Una tabla de variables agregadas de noticias, con 3.538 filas y 17 columnas, que resume el estado diario del corpus periodístico.
- Una tabla de variables específicas por compañía, con 3.538 filas y 65 columnas, que recoge indicadores de sentimiento y atención informativa para las compañías seleccionadas.

La unión final se realizó por fecha, combinando el bloque de mercado, el bloque agregado de noticias y el bloque específico por compañía. A partir de esta unión se generaron dos versiones principales del dataset de modelización. La primera, final_model_dataset, contiene 2.789 observaciones y 83 columnas, con variables objetivo a 3 y 5 días. La segunda, final_model_dataset_1d, contiene 2.789 observaciones y 86 columnas, incorporando también variables objetivo a 1 día. Ambas versiones cubren el periodo comprendido entre el 23 de abril de 2010 y el 8 de diciembre de 2023.

La reducción del número de observaciones respecto a las bases de datos iniciales se debe a la alineación entre fechas de mercado y noticias, el cálculo de variables retardadas, la creación de medias móviles, la disponibilidad efectiva de noticias procesadas y la eliminación de registros sin información suficiente en alguno de los bloques. Esta depuración fue necesaria para obtener una base de datos final coherente, sin valores no válidos y preparada para la evaluación temporal posterior.

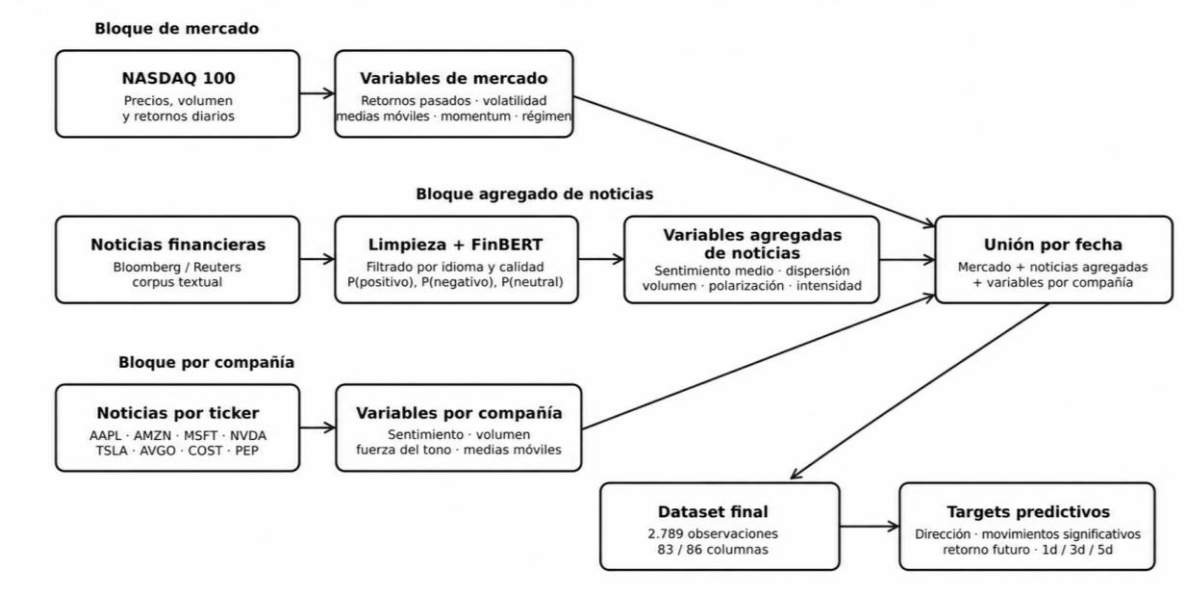


Ilustración 7-Esquema de construcción el dataset final

4.8 VARIABLES FINALES Y OBJETIVOS DE PREDICCIÓN

Una vez construido el dataset final, las variables se organizaron según su uso en la comparación entre modelos. El modelo NASDAQ-only utiliza únicamente el bloque de mercado, formado por variables como retornos pasados, volatilidad, medias móviles, momentum, volumen y variables de régimen. El modelo NASDAQ + News utiliza ese mismo bloque y añade las variables derivadas de noticias, tanto agregadas por fecha como específicas por compañía.

Esta estructura permite que la comparación entre ambos enfoques sea directa: la diferencia entre los modelos no está en el periodo analizado ni en la variable objetivo, sino en la incorporación o no del bloque de noticias. De esta forma, puede evaluarse si la información periodística mejora el comportamiento del modelo respecto a una referencia construida solo con datos históricos del índice.

El dataset final se preparó para tres tareas predictivas. La primera es la predicción direccional, mediante las variables `target_up_1d`, `target_up_3d` y `target_up_5d`, que

indican si el retorno futuro del NASDAQ es positivo en horizontes de 1, 3 y 5 días. Esta tarea se plantea como un problema de clasificación binaria.

La segunda tarea es la detección de movimientos significativos, mediante las variables `target_bigmove_1d`, `target_bigmove_3d` y `target_bigmove_5d`. Estas variables identifican los casos en los que el movimiento futuro del índice supera un determinado umbral de magnitud, permitiendo analizar no solo la dirección del movimiento, sino también su relevancia.

La tercera tarea es la estimación del retorno futuro, mediante las variables `target_return_1d`, `target_return_3d` y `target_return_5d`. En este caso, el objetivo es una variable continua, por lo que se plantea como un problema de regresión.

Esta división permite estudiar el problema desde tres perspectivas complementarias: signo del movimiento, presencia de movimientos significativos y magnitud del retorno esperado.

4.9 PREVENCIÓN DE FUGA DE INFORMACIÓN Y CIERRE DEL DATASET

Uno de los principales riesgos del proyecto era la fuga de información temporal. Para evitarlo, las variables predictoras se construyeron utilizando únicamente información disponible hasta la fecha correspondiente, y la evaluación se planteó con separación cronológica entre entrenamiento y prueba. No se utilizaron particiones aleatorias, ya que en series temporales financieras podrían mezclar observaciones futuras con pasadas y dar una estimación demasiado optimista del rendimiento.

También se controló la alineación entre variables y objetivos de predicción. En los horizontes de 1, 3 y 5 días, las variables objetivo se desplazaron correctamente para que el modelo no utilizara información perteneciente al periodo que intentaba anticipar.

5. METODOLOGÍA DE MODELIZACIÓN

Una vez construido el dataset, el siguiente paso fue diseñar una metodología que permitiera comparar los modelos de forma consistente. La pregunta principal del trabajo no era únicamente qué algoritmo obtiene mejores métricas, sino si las variables derivadas de noticias financieras mejoran la predicción respecto a un modelo basado solo en datos históricos del NASDAQ. Por ello, la metodología se organizó alrededor de una comparación controlada entre modelos NASDAQ-only y NASDAQ + News.

5.1 PLANTEAMIENTO GENERAL

El proceso de modelización se estructuró en torno a dos enfoques. El primero, NASDAQ-only, utiliza únicamente información histórica del índice: retornos pasados, volatilidad, medias móviles, *momentum*, volumen y variables de régimen. Este enfoque representa el rendimiento que puede obtenerse empleando solo variables tradicionales de mercado.

El segundo enfoque, NASDAQ + News, parte del mismo bloque de mercado, pero añade variables procedentes de noticias financieras. En este bloque se incluyen tanto variables agregadas del flujo informativo como variables específicas por compañía. De esta forma, la diferencia entre ambos modelos no está en el periodo analizado ni en la variable objetivo, sino en la incorporación del bloque de noticias.

Este diseño permite evitar una comparación poco equilibrada. El modelo técnico no se plantea como una referencia débil, sino como un modelo construido con variables de mercado razonables. Así, cualquier mejora del modelo NASDAQ + News puede interpretarse de forma más prudente como una mejora asociada a la incorporación de información textual.

5.2 OBJETIVOS DE PREDICCIÓN

El proyecto se evaluó mediante tres tareas predictivas complementarias.

La primera es la predicción direccional del índice, cuyo objetivo es anticipar si el retorno futuro del NASDAQ será positivo o negativo en el horizonte considerado. Esta tarea se formula como un problema de clasificación binaria.

La segunda es la detección de movimientos significativos. No todas las variaciones del índice tienen la misma relevancia financiera, por lo que se definió una variable específica para identificar desplazamientos de mayor magnitud. Esta tarea permite analizar si las noticias ayudan especialmente en episodios de mayor intensidad de mercado.

La tercera es la estimación del retorno futuro, planteada como un problema de regresión. En este caso, el modelo no solo intenta anticipar el signo del movimiento, sino aproximar su magnitud.

Las tres tareas se evaluaron en horizontes de 1, 3 y 5 días. Esta decisión permite analizar si las variables derivadas de noticias tienen un efecto más inmediato o si funcionan mejor en ventanas algo más amplias.

5.3 SEPARACIÓN TEMPORAL Y VALIDACIÓN

La validación de los modelos se realizó respetando el orden temporal de los datos. En lugar de mezclar observaciones de distintos años de forma aleatoria, se separó el conjunto de entrenamiento del conjunto de prueba siguiendo un criterio cronológico.

Esta decisión es especialmente importante en predicción financiera. Una partición aleatoria podría mezclar información futura con datos pasados y generar una estimación demasiado optimista del rendimiento. Por ello, todos los modelos se evaluaron con una separación temporal coherente con la naturaleza del problema.

5.4 MODELOS UTILIZADOS

Se probaron modelos de distinta complejidad para comparar alternativas con diferentes niveles de flexibilidad e interpretabilidad. La selección incluye modelos lineales, métodos basados en árboles, algoritmos de *boosting*[1] y una arquitectura híbrida de combinación.

En primer lugar, se utilizaron modelos lineales, como regresión logística para clasificación y modelos lineales regularizados para regresión. Estos modelos sirven como referencia por su sencillez e interpretabilidad, y permiten comprobar si existe una relación lineal básica entre las variables y los objetivos de predicción.

Después se incorporaron modelos basados en árboles y métodos de *ensemble*[2], como *Random Forest*, *Extra Trees* y *Gradient Boosting*. *Random Forest* fue propuesto por *Breiman* (2001) y se basa en combinar múltiples árboles de decisión entrenados con componentes aleatorias para reducir varianza y mejorar la estabilidad de la predicción. En concreto, utiliza una estrategia de bagging, que consiste en entrenar distintos árboles sobre subconjuntos aleatorios de los datos y combinar posteriormente sus predicciones. De esta forma, al promediar modelos individuales que pueden ser inestables por separado, se obtiene una predicción más robusta y menos dependiente de las particularidades de una muestra concreta. Estos modelos son adecuados para datos tabulares porque pueden capturar relaciones no lineales, interacciones entre variables y efectos de umbral.

También se probaron modelos de *boosting* más avanzados, como *XGBoost* y *LightGBM*. *XGBoost*, presentado por Chen y Guestrin (2016), es un sistema escalable de *tree boosting*

[1] Algoritmos de boosting: técnicas de aprendizaje automático que combinan varios modelos simples, normalmente árboles de decisión, entrenados de forma secuencial. Cada nuevo modelo intenta corregir los errores cometidos por los anteriores, de modo que el conjunto final obtiene una predicción más precisa y robusta que la de cada modelo individual.

[2] Métodos de ensemble: técnicas de aprendizaje automático que combinan varios modelos para generar una predicción conjunta más estable y precisa que la obtenida por un único modelo. Su objetivo es reducir el error, mejorar la generalización y aumentar la robustez del sistema predictivo.

diseñado para mejorar el rendimiento predictivo mediante la combinación secuencial de árboles. *LightGBM*, propuesto por Ke et al. (2017), desarrolla una implementación eficiente de *gradient boosting* especialmente orientada a datos tabulares de gran tamaño. Estos modelos se incluyeron porque suelen ofrecer buen rendimiento en problemas con muchas variables estructuradas.

Por último, se incluyeron redes neuronales sencillas, como el perceptrón multicapa, como parte del conjunto de modelos candidatos. Su uso permite contrastar si una arquitectura no lineal distinta a los árboles mejora la predicción, aunque en este trabajo no se plantea como la solución principal.

Todas las implementaciones se realizaron en Python, utilizando principalmente librerías estándar de aprendizaje automático como *scikit-learn*, *XGBoost* y *LightGBM*. *Scikit-learn* se empleó para los modelos lineales, Random Forest, Extra Trees, Gradient Boosting y MLP, siguiendo el marco de trabajo propuesto por Pedregosa et al. (2011).

5.5 ARQUITECTURA DE COMBINACIÓN DE MODELOS

Además de los modelos individuales, se evaluó una arquitectura de combinación mediante *stacking*. Este enfoque utiliza las predicciones de varios modelos base como variables de entrada para un modelo final, denominado modelo híbrido. La idea es que distintos algoritmos pueden capturar patrones diferentes y que una combinación supervisada puede aprovechar mejor esa información que una selección manual de un único modelo.

En este trabajo, el *stacking* se utilizó como extensión del análisis principal, no como sustituto de la comparación entre NASDAQ-only y NASDAQ + News. Los modelos base se entrenaron con distintos bloques de variables y sus salidas se emplearon para construir una predicción final.

Para evitar sobreajuste y fuga temporal, la combinación se diseñó respetando la separación cronológica de los datos, de modo que el modelo híbrido no utilizara información posterior al periodo de entrenamiento.

La arquitectura final se organizó en tres niveles:

- Modelos base entrenados con variables de mercado.
- Modelos base entrenados con variables de mercado y noticias.
- Modelo híbrido encargado de combinar las predicciones anteriores.

De esta forma, el *stacking* permite comprobar si la combinación de enfoques mejora el rendimiento predictivo sin perder la lógica central del trabajo: evaluar el valor incremental de las noticias frente a una referencia basada únicamente en datos de mercado.

5.6 *MÉTRICAS DE EVALUACIÓN*

Las métricas se seleccionaron en función de la tarea predictiva. En la predicción direccional se utilizaron *accuracy*, *precision*, *recall*, *F1-score* y ROC-AUC. La *accuracy* mide el porcentaje global de aciertos:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

mientras que *precision, recall y F1-score* permiten evaluar mejor el equilibrio entre falsos positivos y falsos negativos:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

El ROC-AUC se empleó para medir la capacidad discriminativa del modelo.

Para la detección de movimientos significativos se utilizaron *precision*, *recall*, *F1-score* y *balanced accuracy*. Esta última métrica resulta útil cuando las clases no están equilibradas, ya que tiene en cuenta el rendimiento del modelo en ambas clases y evita que una clase mayoritaria domine la evaluación:

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

En la regresión del retorno futuro se utilizaron MAE, RMSE, R^2 y correlación entre retorno real y predicho. El MAE mide el error medio absoluto:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

el RMSE penaliza más los errores grandes:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

el R^2 resume la varianza explicada:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

La correlación permite comprobar si el modelo captura parte de la relación entre predicción y retorno observado. Además, se calculó la *accuracy* direccional derivada de la regresión, comparando el signo del retorno predicho con el signo del retorno real.

5.7 INTERPRETACIÓN DE VARIABLES

Además de evaluar el rendimiento predictivo, se analizaron las variables más relevantes en los modelos basados en árboles. El objetivo era comprobar qué tipo de información utilizaban los modelos con mayor peso relativo, diferenciando entre variables de mercado y variables derivadas de noticias.

En el bloque de mercado, la interpretación se centró en variables como retornos recientes, volatilidad y distancia a medias móviles. En el bloque de noticias, se prestó atención a variables como volumen informativo, sentimiento medio, polarización, cambios en la cobertura y variables específicas por compañía.

Esta interpretación se realizó con cautela, ya que la importancia de una variable no implica causalidad. Que un predictor resulte útil para el modelo no significa que provoque directamente el movimiento del mercado, pero sí ayuda a valorar si el modelo está utilizando información coherente con la lógica financiera del problema.

5.8 CRITERIO DE COMPARACIÓN FINAL

La comparación final se realizó evaluando el modelo NASDAQ-only y el modelo NASDAQ + News bajo las mismas condiciones: mismo periodo, mismas variables objetivo y mismas métricas. La diferencia entre ambos resultados constituye la base principal para valorar si las noticias aportan información útil frente a las variables históricas del índice.

Cuando la mejora aparece solo en determinados horizontes, tareas o métricas, la interpretación debe ser prudente. Por ello, los resultados no se entienden como una prueba general de capacidad predictiva, sino como una evaluación empírica del valor incremental de las noticias dentro del diseño definido en este trabajo.

6. RESULTADOS

Este capítulo presenta los resultados obtenidos tras entrenar y evaluar los modelos descritos en la metodología. El análisis se centra en la comparación principal del TFG: comprobar si las variables derivadas de noticias financieras mejoran el rendimiento de los modelos frente a una referencia basada únicamente en datos históricos del NASDAQ.

Los resultados se interpretan con prudencia, ya que el periodo de prueba incluye años especialmente inestables, como 2020, 2022 y 2023. Por ello, el objetivo no es obtener una tasa de acierto artificialmente elevada, sino analizar si la incorporación de noticias produce mejoras consistentes, aunque moderadas, respecto al modelo técnico.

6.1 CONFIGURACIÓN FINAL DEL EXPERIMENTO

El experimento principal se realizó con un horizonte de predicción de 3 días. Esta elección permite capturar un posible efecto de las noticias en una ventana algo más amplia que el movimiento diario, pero todavía cercana al momento de publicación. Tras el filtrado y la eliminación de valores no válidos, el dataset final quedó formado por 2.769 observaciones.

La separación temporal se fijó el 1 de enero de 2020. El conjunto de entrenamiento comprende el periodo entre el 22 de septiembre de 2010 y el 31 de diciembre de 2019, con 1.797 observaciones. El conjunto de prueba cubre desde el 2 de enero de 2020 hasta el 8 de diciembre de 2023, con 972 observaciones.

Esta división resulta exigente porque el periodo de prueba incluye la pandemia, la corrección tecnológica de 2022 y la recuperación posterior del mercado. Además, la distribución de los objetivos cambia entre ambos periodos: el porcentaje de retornos positivos a 3 días pasa del 60,99 % en entrenamiento al 55,86 % en prueba, mientras que los movimientos significativos

aumentan del 24,49 % al 41,15 %. Esto confirma que el periodo de prueba fue más volátil y menos parecido al tramo utilizado para entrenar los modelos.

6.2 RESULTADOS EN PREDICCIÓN DIRECCIONAL DEL NASDAQ

La primera tarea evaluada fue la predicción direccional del NASDAQ a 3 días, entendida como la capacidad del modelo para anticipar si el retorno futuro del índice será positivo o negativo. La Tabla 1 recoge los modelos más representativos de esta tarea, seleccionados por su rendimiento relativo y por su utilidad para comparar los enfoques News-only, NASDAQ-only y NASDAQ + News.

Modelo	**Bloque de variables**	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
XGBoost	News-only	57,30 %	57,96 %	85,82 %	69,19 %	54,67 %
LightGBM ajustado	NASDAQ-only	56,79 %	56,58 %	97,42 %	71,58 %	52,44 %
Hist. Gradient Boosting ajustado	NASDAQ + News	56,28 %	56,16 %	99,08 %	71,69 %	52,75 %
XGBoost	NASDAQ + News	55,97 %	56,38 %	93,55 %	70,36 %	54,26 %
LightGBM	NASDAQ + News	56,07 %	56,37 %	94,48 %	70,61 %	52,28 %

Tabla 1-Resultados principales en predicción de dirección a 3 días.

El mejor resultado en accuracy lo obtiene XGBoost entrenado únicamente con variables de noticias, con un 57,30 % y un ROC-AUC del 54,67 %. Aunque el resultado se sitúa por encima del 50 %, la capacidad discriminativa sigue siendo moderada y debe compararse con cautela frente a referencias ingenuas, como predecir siempre la clase mayoritaria. Por tanto, no debe interpretarse como una capacidad predictiva elevada, sino como un comportamiento razonable dentro de una tarea financiera especialmente difícil.

En la comparación entre NASDAQ-only y NASDAQ + News, la mejora en accuracy es reducida. En algunos modelos, como Hist. Gradient Boosting ajustado, la incorporación de noticias mejora la accuracy en torno a 0,5 puntos porcentuales, mientras que en otros casos la diferencia es prácticamente nula o incluso negativa. Por ejemplo, LightGBM ajustado

obtiene mejores resultados en la versión NASDAQ-only que en la versión con noticias, lo que muestra que el bloque textual no aporta una mejora uniforme en todos los modelos.

Para comprobar si estas diferencias eran estadísticamente relevantes, se realizó una comparación adicional de las pérdidas de clasificación entre los modelos NASDAQ-only y NASDAQ + News / híbrido. La Tabla 2 muestra la diferencia media de pérdida, el estadístico DM y el p-value asociado para los principales modelos comparados.

Modelo	Comparación	Diferencia accuracy	Diferencia media de pérdida	DM statistic	p-value	Conclusión
Hist. Gradient Boosting ajustado	NASDAQ-only vs NASDAQ + News	0,005144	0,005144	0,502645	0,615214	No significativa
Stacked market+news meta-model	NASDAQ-only base vs modelo híbrido stacked	0,002058	0,002058	0,151311	0,879730	No significativa
Random Forest ajustado	NASDAQ-only vs NASDAQ + News	0,000000	0,000000	NaN	NaN	No significativa
Extra Trees ajustado	NASDAQ-only vs NASDAQ + News	0,000000	0,000000	NaN	NaN	No significativa
XGBoost ajustado	NASDAQ-only vs NASDAQ + News	0,000000	0,000000	0,000000	1,000000	No significativa
LightGBM ajustado	NASDAQ-only vs NASDAQ + News	-0,011317	-0,011317	-1,697940	0,089519	No significativa
Logistic Regression ajustado	NASDAQ-only vs NASDAQ + News	-0,022634	-0,022634	-0,790292	0,429357	No significativa

Tabla 2- Comparación estadística entre modelos NASDAQ-only y modelos con noticias

Los resultados de esta comparación indican que ninguna de las diferencias observadas resulta estadísticamente significativa a niveles convencionales. En el caso de Hist. Gradient Boosting ajustado, la mejora de accuracy es positiva, pero el p-value de 0,615 no permite rechazar la hipótesis de igualdad de rendimiento. De forma similar, el modelo híbrido stacked presenta una ligera mejora frente al modelo base, pero con un p-value de 0,880, por

lo que la diferencia debe interpretarse como descriptiva y no como evidencia estadística concluyente.

La mejora resulta algo más visible en términos de F1-score. El Hist. Gradient Boosting ajustado pasa de 70,10 % en NASDAQ-only a 71,69 % en NASDAQ + News, y XGBoost aumenta de 68,87 % a 70,36 %. Esto sugiere que las variables de noticias pueden ayudar a equilibrar precision y recall en determinadas configuraciones, especialmente cuando el modelo tiende a identificar con mayor frecuencia los días de retorno positivo. Sin embargo, esta mejora en F1-score no se traduce necesariamente en una mejora significativa de la accuracy global.

En conjunto, los modelos se sitúan alrededor del 55-57 % de accuracy. Estos valores son razonables para una predicción financiera de corto plazo, pero no permiten hablar de una predicción robusta del mercado. La aportación del bloque de noticias debe entenderse como un valor incremental moderado, visible en algunas métricas y modelos, pero no como una mejora concluyente ni estadísticamente significativa en todos los casos.

Por tanto, la conclusión de esta sección debe formularse con cautela: las variables derivadas de noticias financieras muestran cierta capacidad informativa y pueden mejorar algunas configuraciones concretas, especialmente en F1-score, pero la evidencia estadística disponible no permite afirmar que el modelo con noticias sea sistemáticamente superior al modelo basado únicamente en datos históricos del NASDAQ.

6.3 COMPARACIÓN ENTRE MODELO SIN NOTICIAS Y MODELO CON NOTICIAS

La comparación central del trabajo no consiste solo en identificar el modelo con mayor accuracy, sino en observar qué cambia al añadir variables de noticias sobre una base común de mercado. La siguiente tabla resume la diferencia entre NASDAQ-only y NASDAQ + News en varios modelos representativos.

Modelo	**Métrica**	**NASDAQ-only**	**NASDAQ + News**	**Diferencia**
XGBoost	Accuracy	55,45 %	55,97 %	+0,51 p.p.
Hist. Gradient Boosting ajustado	Accuracy	55,76 %	56,28 %	+0,51 p.p.
Logistic Regression	Accuracy	45,16 %	47,63 %	+2,47 p.p.
XGBoost	F1	68,87 %	70,36 %	+1,49 p.p.
Hist. Gradient Boosting ajustado	F1	70,10 %	71,69 %	+1,59 p.p.
LightGBM	F1	69,99 %	70,61 %	+0,62 p.p.

Tabla 3-Diferencia al añadir noticias al modelo técnico.

Los resultados muestran mejoras moderadas al incorporar noticias, especialmente en F1-score. En XGBoost, el F1 aumenta de 68,87 % a 70,36 %, mientras que en Hist. Gradient Boosting ajustado pasa de 70,10 % a 71,69 %. En accuracy, las mejoras son menores, en torno a 0,51 puntos porcentuales para XGBoost e Hist. Gradient Boosting ajustado. Por tanto, las noticias parecen mejorar algunas métricas bajo determinadas configuraciones, aunque la magnitud del efecto es reducida.

Esta lectura debe matizarse. La comparación es descriptiva y no incorpora un contraste formal de significación estadística, como el test de Diebold-Mariano. Por ello, pequeñas diferencias de accuracy o F1-score no deben interpretarse automáticamente como superioridad concluyente del modelo con noticias.

Además, la distribución de clases condiciona la interpretación de la accuracy. En el conjunto de prueba, el 55,86 % de las observaciones tiene dirección positiva, por lo que un modelo

con sesgo hacia subidas puede obtener una accuracy relativamente alta sin mostrar una capacidad predictiva especialmente fuerte. En este contexto, el F1-score aporta una lectura complementaria al equilibrar precision y recall.

En conjunto, el bloque de noticias no produce una mejora automática ni uniforme. Su aportación aparece en algunas métricas y modelos, pero también puede introducir ruido o perder estabilidad según la configuración utilizada.

6.4 RESULTADOS DEL MODELO HÍBRIDO

Además de los modelos individuales, se evaluó una arquitectura híbrida que combina predicción direccional, detección de movimientos significativos y regresión del retorno. El objetivo era comprobar si la combinación de tareas mejoraba la comparación entre NASDAQ-only y NASDAQ + News, especialmente al integrar señales de distinta naturaleza dentro de una misma estructura predictiva.

Modelo híbrido	Accuracy dirección	Precision	Recall	F1	ROC-AUC
NASDAQ-only	50,51 %	54,44 %	69,98 %	61,24 %	50,57 %
NASDAQ + News	52,47 %	55,78 %	72,01 %	62,86 %	54,26 %

Tabla 4-Resultados del modelo híbrido.

En esta configuración, el modelo NASDAQ + News mejora al modelo NASDAQ-only en todas las métricas principales. La accuracy aumenta de 50,51 % a 52,47 %, el F1-score pasa de 61,24 % a 62,86 % y el ROC-AUC sube de 50,57 % a 54,26 %. Esta mejora es moderada, pero resulta coherente con la hipótesis del trabajo: las variables derivadas de noticias pueden aportar valor en determinadas configuraciones, especialmente cuando se combinan distintas

tareas predictivas. Aun así, el nivel de ROC-AUC indica que la capacidad discriminativa del modelo continúa siendo limitada.

Para complementar esta lectura descriptiva, se realizó una comparación estadística de las pérdidas entre el modelo NASDAQ-only y el modelo NASDAQ + News. En la tarea de predicción direccional híbrida se utilizó como función de pérdida el error 0/1, mientras que en la regresión del retorno se utilizó el error cuadrático.

Tarea	Comparación	Función pérdida	Diferencia media de la pérdida	DM statistic	p-value	Conclusión
Predicción direccional híbrida	NASDAQ-only vs NASDAQ + News	Error 0/1	0,019547	1,094635	0,273677	No significativa
Regresión del retorno	NASDAQ-only vs NASDAQ + News	Error cuadrático	0,000017	0,998472	0,318051	No significativa

Tabla 5- Comparación estadística de pérdidas en el modelo híbrido.

Los resultados muestran que, aunque el modelo NASDAQ + News obtiene mejores métricas descriptivas en la predicción direccional híbrida, la diferencia no resulta estadísticamente significativa. El p-value de 0,2737 no permite rechazar la hipótesis de igualdad de rendimiento entre ambos modelos. Por tanto, la mejora de 1,95 puntos porcentuales en accuracy debe interpretarse con cautela.

En la tarea de regresión del retorno ocurre algo similar. El modelo con noticias reduce ligeramente el error cuadrático medio, pasando de una pérdida de 0,000828 a 0,000811, pero la diferencia media de pérdida es muy pequeña y tampoco resulta estadísticamente significativa, con un p-value de 0,3181. Esto refuerza la idea de que las noticias pueden aportar una señal incremental, pero no suficiente para demostrar una superioridad estadística clara frente al modelo basado únicamente en datos de mercado.

En conjunto, el modelo híbrido ofrece una de las lecturas más favorables para la incorporación de noticias, ya que mejora de forma consistente las métricas principales frente al enfoque NASDAQ-only. Sin embargo, la validación estadística obliga a formular la conclusión con prudencia: las variables de noticias parecen aportar información adicional en esta arquitectura, pero la evidencia disponible no permite afirmar que el modelo híbrido con noticias sea significativamente superior al modelo sin noticias.

6.5 *RESULTADOS EN REGRESIÓN DEL RETORNO*

La regresión del retorno fue la tarea más exigente del trabajo, ya que no solo requiere anticipar el signo del movimiento, sino también aproximar su magnitud. En mercados financieros a corto plazo, esta tarea suele ser especialmente difícil porque los retornos diarios presentan un componente elevado de ruido y una varianza difícil de explicar con variables observables.

Modelo	Bloque de variables	RMSE	MAE	Correlación	Accuracy direccional
Random Forest Regressor	News-only	0,02765	0,02070	0,0950	57,61 %
Extra Trees Regressor	News-only	0,02770	0,02071	0,0844	57,82 %
Extra Trees Regressor	NASDAQ + News	0,02816	0,02109	0,0175	54,01 %
XGBoost Regressor	NASDAQ + News	0,02847	0,02144	0,0625	53,19 %
XGBoost Regressor	NASDAQ-only	0,02877	0,02147	0,0179	51,34 %

Tabla 6-Mejores resultados de regresión del retorno a 3 días.

Los resultados muestran que los errores entre modelos son relativamente próximos, lo que confirma la dificultad de estimar con precisión la magnitud del retorno futuro. Aun así, aparecen diferencias relevantes en la orientación de las predicciones. Los modelos entrenados únicamente con variables de noticias, especialmente Random Forest y Extra Trees, obtienen las mejores *accuracies* direccionales derivadas del retorno, con valores

superiores al 57 %. Esto indica que, aunque la magnitud exacta del retorno sea difícil de estimar, las variables de noticias pueden ayudar a aproximar parcialmente el signo del movimiento.

La comparación entre NASDAQ-only y NASDAQ + News también muestra una mejora moderada en algunos modelos de regresión. Por ejemplo, en XGBoost Regressor, al incorporar noticias, el RMSE baja de 0,02877 a 0,02847, la correlación aumenta de 0,0179 a 0,0625 y la *accuracy* direccional pasa de 51,34 % a 53,19 %. La mejora no es elevada, pero mantiene la misma lectura observada en las tareas anteriores: el bloque de noticias puede aportar información útil en determinadas configuraciones, aunque no transforma por completo la capacidad predictiva del modelo.

Para comprobar si las diferencias en error eran estadísticamente relevantes, se compararon las pérdidas cuadráticas entre los modelos NASDAQ-only y NASDAQ + News. La Tabla 7 resume los resultados principales de esta comparación.

Modelo	**Diferencia RMSE**	**Diferencia media de pérdida**	**DM statistic**	**p-value**	**Conclusión**
Hist. Gradient Boosting Regressor	-0,000450	0,000026	1,402769	0,160686	No significativa
XGBoost Regressor	-0,000304	0,000017	0,998472	0,318051	No significativa
Random Forest Regressor	-0,000107	0,000006	1,008988	0,312980	No significativa
Extra Trees Regressor	-0,000072	0,000004	0,751064	0,452614	No significativa

Tabla 7-Comparación estadística de pérdidas en regresión del retorno.

Los resultados muestran que, en la mayoría de los modelos, la incorporación de noticias reduce ligeramente el RMSE, pero las diferencias no son estadísticamente significativas. Esto ocurre en Hist. Gradient Boosting Regressor, XGBoost Regressor, Random Forest Regressor y Extra Trees Regressor, donde los *p-values* son superiores a los niveles

convencionales de significación. Por tanto, estas mejoras deben interpretarse como diferencias descriptivas, no como evidencia concluyente de superioridad estadística.

En conjunto, la regresión confirma que predecir la magnitud del retorno es más complejo que anticipar su dirección. La incorporación de noticias genera pequeñas mejoras en algunos modelos no lineales, pero estas no son estadísticamente significativas. Por ello, en esta tarea la aportación de las noticias debe interpretarse con cautela: pueden contener información útil para orientar parcialmente la predicción, pero no permiten mejorar de forma robusta la estimación cuantitativa del retorno futuro.

6.6 DETECCIÓN DE MOVIMIENTOS SIGNIFICATIVOS

La detección de movimientos significativos permite evaluar si el modelo identifica episodios de mayor magnitud, no solo si acierta la dirección del índice. Esta tarea es relevante porque, desde un punto de vista financiero, no todos los aciertos tienen el mismo valor: fallar en los días de mayor movimiento puede ser más importante que acertar variaciones pequeñas.

En el conjunto de prueba, los movimientos significativos representan el 41,15 % de las observaciones, frente al 24,49 % del conjunto de entrenamiento. Esta diferencia confirma que el periodo de prueba fue más intenso y menos parecido al tramo utilizado para entrenar los modelos.

En el modelo híbrido final, la detección de movimientos significativos alcanzó un ROC-AUC aproximado de 63,28 %, con una accuracy del 54,30 %, un *recall* del 87,67 % y un F1-score del 60,95 %. El *recall* elevado indica que el modelo identifica una parte amplia de los episodios relevantes, aunque a costa de generar falsos positivos. Por tanto, esta configuración puede ser útil para no perder días de mayor intensidad, pero no sería suficiente por sí sola para una estrategia operativa.

Este resultado es uno de los más interesantes del capítulo. Mientras que la predicción direccional general sigue siendo limitada, la detección de movimientos significativos parece

mostrar una respuesta más clara al incorporar información procedente de noticias. Esta lectura es coherente desde el punto de vista financiero: las noticias pueden no anticipar todos los movimientos pequeños del índice, pero sí reflejar mejor momentos de atención, tensión o cambio de narrativa del mercado.

6.7 IMPORTANCIA DE VARIABLES

Para interpretar el modelo NASDAQ + News, se analizaron las importancias de variables obtenidas en los modelos basados en árboles. En este tipo de algoritmos, la importancia mide la contribución relativa de cada variable a la reducción del error o a la mejora de las divisiones realizadas por el modelo. Posteriormente, estas importancias se agruparon por bloque para comparar el peso relativo de las variables de mercado y de noticias.

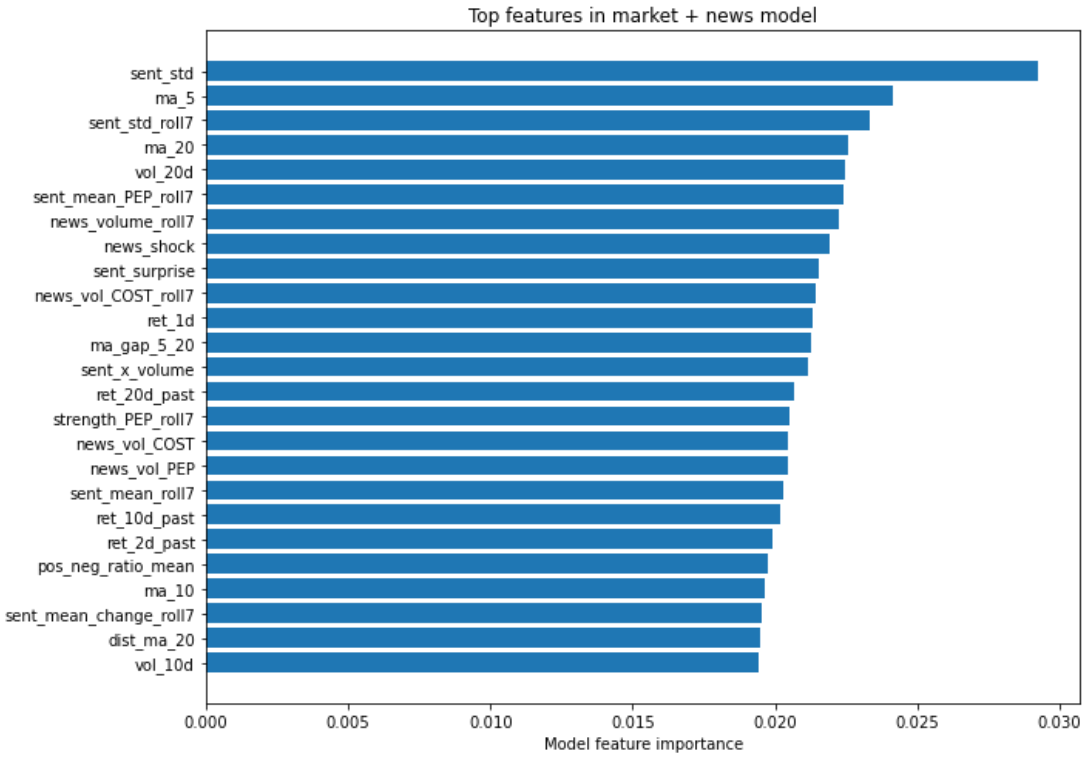


Ilustración 8-Importancia de las variables

El análisis agregado muestra que las variables derivadas de noticias representan aproximadamente el 67,7 % de la importancia total, frente al 32,3 % correspondiente a las variables de mercado. Este resultado no implica causalidad, pero sí indica que el modelo utiliza de forma relevante la información procedente de noticias durante el proceso de predicción.

Entre las variables de noticias con mayor peso aparecen `sent_std`, `sent_std_roll7`, `sent_mean_PEP_roll7`, `news_volume_roll7`, `news_shock`, `sent_surprise`, `sent_x_volume` y `pos_neg_ratio_mean`. Estas variables recogen distintas dimensiones del flujo informativo: dispersión diaria del sentimiento, evolución reciente del tono mediante medias móviles, volumen de noticias, cambios anómalos en la cobertura y relación entre noticias positivas y negativas.

También aparecen variables de mercado como `ma_5`, `ma_20`, `vol_20d`, `ret_1d`, `ret_10d_past` y distancias a medias móviles. Estas variables resumen la dinámica reciente del NASDAQ, incluyendo tendencia, volatilidad, retornos pasados y posición del índice respecto a sus niveles medios.

6.8 LECTURA GLOBAL DE LOS RESULTADOS

Los resultados confirman que la predicción del NASDAQ a corto plazo sigue siendo una tarea compleja. Los mejores modelos de clasificación direccional se sitúan en torno al 55-57 % de accuracy, valores moderados pero razonables para un problema financiero con elevada incertidumbre y un periodo de prueba especialmente volátil.

La mejora asociada a las noticias aparece de forma selectiva, no uniforme. En predicción direccional, algunos modelos NASDAQ + News mejoran el F1-score frente a NASDAQ-only. En regresión, los modelos basados únicamente en noticias obtienen algunas de los mejores valores de la métrica accuracies direccional derivadas del retorno. Además, la detección de movimientos significativos alcanza un ROC-AUC superior al de la clasificación direccional general, lo que sugiere que la información periodística puede ser

más útil para identificar episodios de mayor intensidad que para anticipar todos los movimientos diarios del índice.

Esta lectura debe interpretarse con prudencia. Los resultados no demuestran una relación causal entre noticias y movimientos del NASDAQ, sino una asociación predictiva dentro del diseño metodológico del trabajo. Las variables derivadas del flujo informativo mejoran algunas métricas y configuraciones, pero la mejora es moderada y no aparece de forma consistente en todos los modelos.

En conjunto, el valor del proyecto no reside únicamente en la métrica final, sino en haber construido un proceso completo para transformar noticias financieras en variables cuantitativas y compararlas con un modelo técnico de mercado. El resultado es prudente, pero relevante: las noticias pueden aportar información útil para determinadas tareas predictivas, especialmente cuando se analizan movimientos significativos, aunque no eliminan la incertidumbre propia de la predicción bursátil.

7. CONCLUSIONES, LIMITACIONES Y TRABAJO FUTURO

Este trabajo partía de una pregunta concreta: comprobar si las noticias financieras, transformadas en variables cuantitativas, aportan información útil para predecir el comportamiento del NASDAQ frente a un modelo basado únicamente en datos históricos del índice. La respuesta no es absoluta, pero sí permite extraer una conclusión clara: las variables derivadas de noticias pueden mejorar algunas métricas bajo determinadas configuraciones, aunque la mejora es moderada y no aparece de forma uniforme en todos los modelos.

La principal aportación del proyecto ha sido construir un proceso completo de análisis, desde la base textual original hasta la evaluación final de modelos. No se ha trabajado únicamente con una tabla ya preparada, sino con una base de noticias de gran tamaño que ha requerido limpieza, filtrado, procesamiento mediante FinBERT y agregación diaria. A partir de este proceso se ha construido un dataset final que combina variables de mercado, variables agregadas de noticias y variables específicas por compañía.

7.1 CONCLUSIONES PRINCIPALES

La primera conclusión es que el bloque de noticias puede aportar información útil, pero no resuelve el problema de predicción del NASDAQ. En predicción direccional, los mejores modelos se sitúan en torno al 55-57 % de accuracy. Estos valores son razonables para una tarea financiera de corto plazo, pero no permiten hablar de una capacidad predictiva elevada. La mejora al añadir noticias aparece sobre todo en algunas métricas, especialmente en F1-score, y debe entenderse como un valor incremental moderado.

La segunda conclusión es que las noticias parecen ser más útiles para identificar determinados contextos de mercado que para anticipar todos los movimientos diarios. En la detección de movimientos significativos, el modelo obtiene un ROC-AUC superior al observado en la clasificación direccional general. Esto sugiere que el flujo informativo puede ser especialmente relevante en episodios de mayor intensidad, cuando aumentan la atención mediática, el volumen de noticias o la dispersión del sentimiento.

La tercera conclusión es que el bloque de noticias no debe reducirse al sentimiento medio. Algunas de las variables más relevantes están relacionadas con la dispersión del sentimiento, el volumen informativo, las medias móviles, los cambios anómalos en la cobertura y la interacción entre sentimiento y volumen. Por tanto, en finanzas no solo importa si una noticia es positiva o negativa, sino también cuánta atención genera, cómo evoluciona esa atención y si el mensaje informativo aparece concentrado o fragmentado.

La cuarta conclusión es que la comparación entre NASDAQ-only y NASDAQ + News era necesaria para interpretar correctamente los resultados. Sin una referencia basada únicamente en variables técnicas del índice, no habría sido posible valorar el efecto de incorporar información textual. El resultado muestra que añadir noticias puede mejorar determinadas métricas, pero no siempre ni en todos los modelos.

En conjunto, los resultados no demuestran una relación causal entre noticias y movimientos del NASDAQ, sino una asociación predictiva dentro del diseño metodológico del trabajo. Las variables derivadas del flujo informativo ayudan a mejorar algunas configuraciones, pero no permiten afirmar que las noticias causen directamente los movimientos observados ni que la mejora sea estable en cualquier contexto de mercado.

7.2 LIMITACIONES DEL TRABAJO

La primera limitación está en la propia naturaleza del problema. Los mercados financieros son ruidosos, cambiantes y afectados por factores que no siempre están presentes en los

datos. Tipos de interés, decisiones de bancos centrales, resultados empresariales, shocks geopolíticos o cambios de expectativas pueden modificar rápidamente el comportamiento del índice.

La segunda limitación tiene que ver con la frecuencia temporal. El trabajo opera a nivel diario, lo que permite construir una base coherente con los precios del índice, pero deja fuera la reacción intradía. Muchas noticias tienen impacto en minutos u horas, por lo que la agregación diaria puede diluir parte de esa información.

La tercera limitación está relacionada con la cobertura de noticias. Aunque la base original es amplia, no es completamente homogénea a lo largo del periodo. Hay años con mucho mayor volumen de registros que otros, y parte de esa variación puede deberse a la propia fuente de datos, no solo a cambios reales en la atención del mercado.

La cuarta limitación es la interpretación del sentimiento. FinBERT permite trabajar con un modelo adaptado al lenguaje financiero, pero el sentimiento no captura toda la relevancia de una noticia. Un texto puede ser neutral en tono y, aun así, ser importante para el mercado. También pueden existir ambigüedad, lenguaje técnico o referencias contextuales que el modelo no interprete perfectamente.

La quinta limitación está en el tamaño efectivo de la muestra para modelización. Aunque el dataset final contiene miles de observaciones, en series temporales financieras sigue siendo una muestra limitada, especialmente al separar entrenamiento y prueba. Esto obliga a interpretar con cuidado los resultados de modelos muy flexibles.

7.3 TRABAJO FUTURO

Una primera línea de trabajo futuro sería incorporar una validación estadística más formal de las diferencias entre modelos, especialmente mediante el test de Diebold-Mariano. Esto permitiría comprobar si las mejoras observadas responden a una diferencia real de capacidad predictiva o si pueden deberse al azar.

Otra línea de mejora sería trabajar con datos intradía, es decir, precios y volúmenes registrados dentro de la propia sesión de mercado. Esto permitiría estudiar con mayor precisión el efecto temporal de las noticias y analizar si determinadas publicaciones tienen impacto inmediato sobre el índice.

También sería interesante ampliar el análisis textual más allá del sentimiento. Podrían incorporarse técnicas de detección de temas, eventos, entidades o cambios de narrativa, con el objetivo de distinguir mejor entre noticias de resultados, regulación, innovación tecnológica, fusiones y adquisiciones o shocks macroeconómicos.

Una cuarta línea futura sería analizar la estabilidad del modelo por subperiodos. El periodo de prueba incluye contextos muy distintos, como la pandemia, la corrección tecnológica de 2022 y la recuperación posterior. Evaluar los modelos por etapas permitiría comprobar si las noticias son más útiles en periodos de alta volatilidad o cambios bruscos de expectativas.

Por último, podría aplicarse la misma metodología a otros índices o activos financieros, como el S&P 500, el Dow Jones o índices sectoriales. Esto permitiría comprobar si los resultados observados en el NASDAQ son propios de un índice tecnológico o si se reproducen en otros segmentos del mercado.

8. BIBLIOGRAFÍA

[1] Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383–417.

[2] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3–56.

[3] Hong, H., & Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance, 54*(6), 2143–2184.

[4] Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance, 48*(1), 65–91.

[5] Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies, 21*(4), 1455–1508.

[6] Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223–2273.

[7] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.

[8] Heston, S. L., & Sinha, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal, 73*(3), 67–83.

[9] Ke, Z. T., Kelly, B., & Xiu, D. (2019). Predicting returns with text data. *Becker Friedman Institute Working Paper No. 2019-69*.

[10] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65.

[11] 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*.

[12] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

[13] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

[14] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

[15] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[16] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*.

[17] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

[18] Remy, P., & Ding, X. (n.d.). *Financial News Dataset from Bloomberg and Reuters*. GitHub repository: philipperemy/financial-news-dataset.

[19] Yahoo Finance. (n.d.). *NASDAQ 100 historical market data*. Yahoo Finance.

[20] Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*. https://doi.org/10.48550/arXiv.1908.10063

ANEXO I: ALINEACIÓN DEL PROYECTO CON ODS

Este Trabajo Fin de Grado se alinea principalmente con los Objetivos de Desarrollo Sostenible relacionados con la innovación, la educación técnica, el crecimiento económico y la toma de decisiones basada en información. Aunque el proyecto tiene una orientación claramente financiera y tecnológica, su aportación no se limita al desarrollo de un modelo predictivo. También contribuye a explorar cómo las herramientas de análisis de datos, procesamiento de lenguaje natural e inteligencia artificial pueden utilizarse para transformar información no estructurada en conocimiento útil.

ODS 4. Educación de calidad

El proyecto se relaciona con el ODS 4 porque aplica conocimientos avanzados de Business Analytics, machine learning, procesamiento de lenguaje natural y análisis financiero en un caso práctico real. La construcción del dataset, el tratamiento de una base de noticias de gran tamaño, el uso de FinBERT y la evaluación de modelos predictivos permiten integrar competencias técnicas y analíticas de forma aplicada.

Además, el trabajo contribuye a una forma de aprendizaje orientada a problemas reales. No se limita a explicar modelos de forma teórica, sino que los utiliza dentro de un flujo completo de datos: recopilación, limpieza, transformación, modelización y evaluación. Esta aproximación encaja con la formación en competencias digitales y cuantitativas que promueve este objetivo.

ODS 8. Trabajo decente y crecimiento económico

El ODS 8 busca promover el crecimiento económico sostenido, inclusivo y sostenible, así como el empleo productivo y el trabajo decente. Este proyecto se vincula con este objetivo desde la perspectiva de la mejora en el análisis de información financiera. Los mercados dependen cada vez más de la capacidad para interpretar grandes volúmenes de datos, tanto

estructurados como no estructurados. En este contexto, desarrollar herramientas que permitan procesar noticias y convertirlas en variables analizables puede contribuir a una toma de decisiones más informada.

El trabajo no pretende automatizar decisiones de inversión ni sustituir el criterio humano, sino aportar una metodología que permita evaluar mejor el papel de la información periodística en el comportamiento del mercado. Esta visión es relevante en entornos financieros, consultoría, gestión de riesgos y análisis económico, donde la calidad de la información influye directamente en la toma de decisiones.

ODS 9. Industria, innovación e infraestructura

El ODS más directamente relacionado con este TFG es el ODS 9, centrado en industria, innovación e infraestructuras. El proyecto desarrolla una solución tecnológica basada en inteligencia artificial y procesamiento de lenguaje natural para abordar un problema complejo: estudiar si las noticias financieras contienen información predictiva sobre el NASDAQ.

La innovación del trabajo está en integrar fuentes de datos heterogéneas dentro de un mismo pipeline. Por un lado, se utilizan variables tradicionales de mercado; por otro, se transforma texto financiero en indicadores cuantitativos mediante técnicas de NLP. Esta combinación permite construir modelos más completos y estudiar el valor incremental de la información textual frente a los datos financieros clásicos.

Además, el proyecto trabaja con una base de noticias de gran tamaño, que requiere preprocesado, filtrado, agregación y extracción de sentimiento. Este proceso refleja el tipo de infraestructura analítica que cada vez es más necesaria en sectores intensivos en datos.

ODS 12. Producción y consumo responsables

Aunque el proyecto no está vinculado directamente con producción física o consumo material, sí se puede relacionar con el ODS 12 desde la perspectiva del uso responsable de

la información. En mercados financieros, la sobreabundancia de noticias puede generar ruido, reacciones exageradas o decisiones poco fundamentadas. Este trabajo propone una forma ordenada de procesar esa información, separando aquello potencialmente útiles de datos menos relevantes.

La aproximación seguida evita utilizar la inteligencia artificial como una herramienta opaca o automática. Al contrario, el trabajo compara modelos, analiza limitaciones, revisa la importancia de variables y mantiene una interpretación prudente de los resultados. Esta forma de trabajar favorece un uso más responsable de los datos y de los modelos predictivos.

ODS 17. Alianzas para lograr los objetivos

El ODS 17 se relaciona con la cooperación y la integración de conocimiento entre disciplinas. Este TFG combina varias áreas: finanzas, ciencia de datos, procesamiento de lenguaje natural, aprendizaje automático y análisis de series temporales. Esa combinación es necesaria para abordar problemas actuales, donde las soluciones no suelen pertenecer a una única disciplina.

El proyecto muestra cómo técnicas procedentes de la inteligencia artificial pueden aplicarse a un problema financiero real, manteniendo al mismo tiempo una interpretación económica de los resultados. Esta integración entre tecnología y análisis económico es una parte esencial de la transformación digital de las organizaciones.