COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO FIN DE GRADO

ANÁLISIS DE LA INFLUENCIA DE LA INFORMACIÓN DESESTRUCTURADA EN LA COTIZACIÓN DE EMPRESAS DEL SECTOR ENERGÉTICO

María Corella Gómez

Director: Dr. Antonio García de Garmendia

Madrid

Julio de 2023

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Análisis de la influencia de la información desestructurada en la cotización de empresas del sector energético

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2022-2023 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: María Corella Fecha: 01 / 07 / 2023

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.:  Fecha: 01 / 07 / 2023

ÍNDICE DE LA MEMORIA

RESUMEN DEL PROYECTO

En este proyecto se ha contruído un modelo de predicción de precios de acciones mediante la frecuencia de aparición de las palabras en los titulares de noticias. El objetivo del modelo es evaluar la influencia de la información desestrcuturada en la cotización de empresas del sector energético. El trabajo incluye un planteamiento teórico de la hipótesis, un caso práctico y un estudio económico de la viabilidad de convertir el modelo en una stratup fintech. Se concluye que determinados sucesos, temáticas o sentimientos reflejados en las noticias tienen el potencial de predecir las dinámicas futuras del mercado bursátil. Además, se demuestra que el modelo no solo es efectivo desde el punto de vista técnico, sino que también tiene un potencial económico significativo.

Palabras clave: Predicción de precios de acciones, Inteligencia artificial, Procesamiento del lenguaje natural (NLP), Coocurrencias de palabras, Narrativa periodística, IBEX35, IBEX-Energy, Endesa, Sector energético, Correlación de Pearson

CASO DE ESTUDIO

Los datos de entrada al modelo provienen de la recopilación de precios históricos de tres índices diferentes, IBEX35, IBEX-Energy y ELE (ticker de Endesa), y de noticias de los principales periódicos españoles durante los días de mayor variación de precios para cada uno de los índices. Se lleva a cabo un análisis del lenguaje utilizado en las noticias y su relación con las variaciones de precios haciendo uso de un código Python sencillo.

Tras analizar la frecuencia de las palabras en los titulares de noticias durante los días de mayores subidas y bajadas de los índices, se crea un 'diccionario' de palabras clave y se explora su correlación con estas variaciones de precios.

Finalmente, el modelo se aplica a un día específico, fuera del conjunto de datos de entrenamiento, con el objetivo de predecir la dirección de los precios de las acciones de los tres índices estudiados y así evaluar el rendimiento del modelo.

RESULTADOS

El coeficiente de correlación de Pearson entre la frecuencia de aparición y la variación de precio en un día confirma la relevancia de ciertas palabras en la cotización de ciertas empresas: se descubre que, para cada índice, existen palabras clave cuya aparición en las noticias está correlacionada con la variación de su precio. Durante la evaluación del rendimiento, las métricas del MSE y R^2 proporcionan evidencia de que el modelo desarrollado logra predecir con cierta precisión las variaciones en los precios de las acciones.

CONCLUSIONES

La investigación revela que la inclusión del análisis de noticias en los modelos de predicción de precios de acciones puede aportar una visión más completa y matizada del comportamiento del mercado. Este hallazgo resalta la utilidad de las técnicas avanzadas de procesamiento del lenguaje natural y aprendizaje automático en las finanzas cuantitativas, y sugiere la necesidad de explorar más a fondo la integración de factores cualitativos en los modelos financieros cuantitativos.

ABSTRACT

In this project, a stock price prediction model has been built using the frequency of occurrence of words in news headlines. The objective of the model is to evaluate the influence of uncut information on the share price of companies in the energy sector. The paper includes a theoretical approach to the hypothesis, a case study and an economic study of the feasibility of converting the model into a fintech startup. It is concluded that certain events, themes, or sentiments reflected in the news have the potential to predict future stock market dynamics. Furthermore, it is shown that the model is not only effective from a technical point of view, but also has significant economic potential.

Keywords: Stock price prediction, Artificial intelligence, Natural Language Processing (NLP), Word co-occurrences, News narrative, IBEX35, IBEX-Energy, Endesa, Energy sector, Pearson correlation.

CASE STUDY

The input data for the model comes from the collection of historical prices of three different indexes, IBEX35, IBEX-Energy and ELE (Endesa ticker), and news from the main Spanish newspapers during the days of highest price variation for each of the indexes. An analysis of the language used in the news and its relationship with price variations is carried out using a simple Python code.

After analyzing the frequency of the words in the news headlines during the days with the largest increases and decreases in the indexes, a 'dictionary' of keywords is created and their correlation with these price variations is explored.

Finally, the model is applied to a specific day, outside the training data set, in order to predict the direction of the stock prices of the three indexes studied and thus evaluate the performance of the model.

RESULTS

The Pearson correlation coefficient between the frequency of appearance and the price variation in a day confirms the relevance of certain words in the share price of certain companies: it is found that, for each index, there are keywords whose appearance in the news is correlated with the direction of their price. During the performance evaluation, the MSE and R^2 metrics provide evidence that the developed model manages to predict with some accuracy the variations in stock prices.

CONCLUSIONS

The research reveals that the inclusion of news analysis in stock price prediction models can provide a more complete and nuanced view of market behaviour. This finding highlights the utility of advanced natural language processing and machine learning techniques in quantitative finance and suggests the need to further explore the integration of qualitative factors into quantitative financial models.

CAPÍTULO 1. INTRODUCCIÓN

El primer capítulo introduce la justificación detrás de este proyecto, así como sus principales objetivos. Además, proporciona al lector una visión general de la organización y el esquema de la tesis para facilitar su seguimiento.

Se examina el sector energético en la actualidad, explorando las tendencias, desafíos y oportunidades clave en este ámbito. Paralelamente, se presenta una breve historia sobre la predicción de precios de acciones, el aprendizaje automático y el procesamiento del lenguaje natural (NLP), trazando la evolución de estas disciplinas.

Finalmente, la *Sección 1.3* se dedica a esclarecer los objetivos y el alcance de la tesis, describiendo en detalle las metas que se buscan alcanzar y delimitando el enfoque de la investigación. La *Sección 1.4* establece la alineación de este Trabajo de Fin de Grado con los Objetivos de Desarrollo Sostenible de las Naciones Unidas.

1.1 SECTOR ENERGÉTICO EN LA ACTUALIDAD

El sector energético desempeña un papel vital en la economía global. Sus actividades de generación, distribución y consumo de energía son fundamentales para el desarrollo económico y social de cualquier país. Además, el desempeño financiero de las empresas energéticas tiene una influencia directa en la actividad de otras industrias, como la producción de bienes y servicios, el transporte, la construcción y la agricultura. Por consiguiente, resulta de gran interés conseguir una herramienta que permita predecir los precios de acciones de estas empresas, con el fin de tomar decisiones informadas en el ámbito financiero.

En el escenario contemporáneo, el sector energético se encuentra en un estado de transformación radical. Este cambio está motivado por la creciente preocupación por el cambio climático y la urgencia de reducir las emisiones de gases de efecto invernadero. Esta tendencia ha catalizado un auge en la inversión en energías renovables, como la solar y la eólica, así como en políticas para fomentar su adopción. En este contexto, están emergiendo nuevas tecnologías,

como las relacionadas con la eficiencia energética [28], que permiten una mayor diversificación de la matriz energética y una reducción en la dependencia de combustibles fósiles.

Por otro lado, el sector energético se enfrenta a desafíos geopolíticos significativos. La reciente guerra en Ucrania y las tensiones entre Rusia y Occidente han añadido un nivel de complejidad adicional, ya que Rusia es uno de los principales proveedores de gas y petróleo a Europa. Esta situación ha alimentado una creciente inquietud en cuanto a la seguridad energética y ha impulsado a los países a buscar alternativas y reducir su dependencia de la Federación Rusa [53].

Dentro de este panorama de incertidumbre, las empresas energéticas que se adaptan rápidamente a estos cambios pueden tener una ventaja competitiva, mientras que aquellas que no lo hacen pueden enfrentar dificultades financieras. Por tanto, comprender y predecir las tendencias y cambios en el sector energético es esencial. Este es precisamente el enfoque y la ambición de este proyecto: proporcionar una herramienta analítica que permita predecir los precios de las acciones de las empresas energéticas con el objetivo de informar y guiar la toma de decisiones financieras.

1.2 PREDICCIÓN DE MERCADOS FINANCIEROS, APRENDIZAJE AUTOMÁTICO Y NLP

El arte y la ciencia de predecir el comportamiento de los mercados financieros tiene un largo historial de intentos, éxitos, fracasos y renovaciones teóricas. Durante décadas, la Hipótesis del Mercado Eficiente (EMH)[1] fue la teoría más aceptada sobre la predicción de acciones. La EMH sostiene que los precios de las acciones siempre reflejan toda la información disponible, lo que implicaría que las ganancias superiores al promedio del mercado son casi imposibles de conseguir de forma consistente [15]. Sin embargo, la incapacidad de esta teoría para explicar ciertas anomalías y comportamientos del mercado llevó a un replanteamiento de las metodologías para la predicción del precio de acciones.

[1] Por sus siglas en inglés, Efficient Market Hypothesis

Diversas técnicas se han explorado a lo largo de los años, desde los enfoques más tradicionales de análisis fundamental y técnico, hasta modelos más complejos y modernos que involucran la simulación de Monte Carlo y las redes neuronales. A pesar de estos avances, la predicción de acciones sigue siendo un desafío debido a la alta volatilidad y la incertidumbre inherente a los mercados financieros. En la *Figura 1.1* se representa una línea del tiempo de la evolución de las técnicas de predicción de mercados financieros:

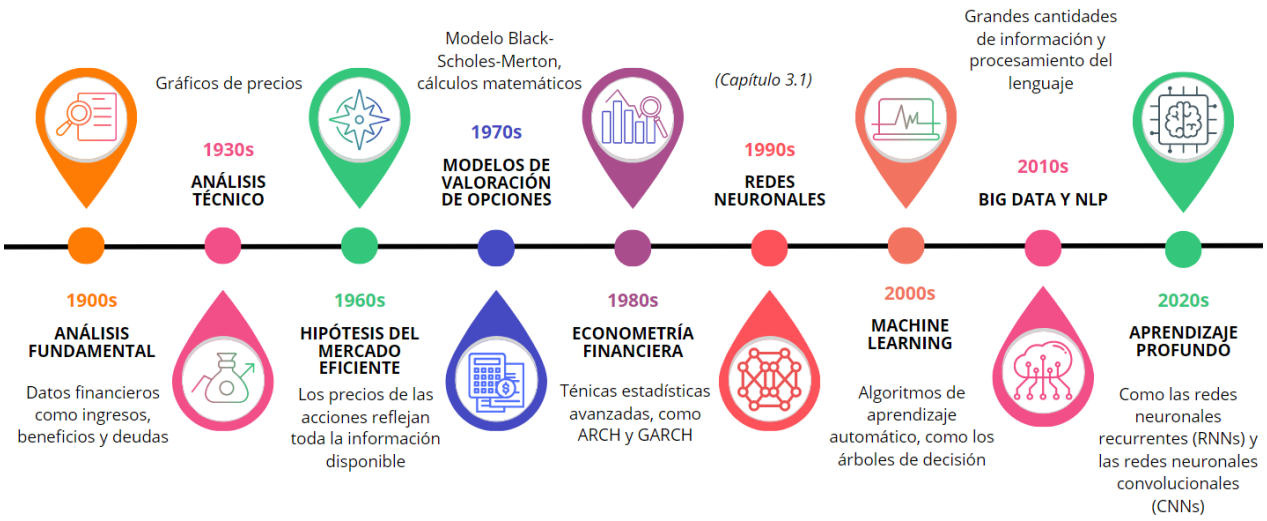

Figura 1.1. Evolución de los métodos de predicción en mercado de valores (elaboración propia, 2023)

En este contexto de búsqueda continua de mejores métodos predictivos, ha surgido una revolución en los últimos años gracias a los avances en el aprendizaje automático (ML)[2] y el Procesamiento de Lenguaje Natural (NLP)[3]. Estas disciplinas, que forman parte del amplio campo de la Inteligencia Artificial, han permitido desarrollar herramientas y enfoques novedosos para enfrentar el desafío de la predicción de acciones [16].

El aprendizaje automático se basa en la idea de que los sistemas informáticos pueden aprender de los datos, identificar patrones y tomar decisiones con mínima intervención humana. Aplicado al ámbito financiero, el ML puede ayudar a descubrir relaciones y tendencias que no son fácilmente detectables con métodos tradicionales.

[2] Por sus siglas en inglés, Machine Learning
[3] Por sus siglas en inglés, Natural Language Processing

Por otro lado, el NLP es una tecnología que permite a las máquinas entender, interpretar y responder al lenguaje humano. Asumiendo que los artículos de noticias tienen un impacto en el mercado de valores, este trabajo se presenta como un intento de estudiar la relación entre las noticias y la tendencia de las acciones, tomando datos no cuantificables como los titulares de noticias y prediciendo la tendencia futura de las acciones de una empresa con la frecuencia de aparición de las palabras en las noticias.

Según un informe elaborado por Gartner[4], las técnicas emergentes de inteligencia artificial, como el aprendizaje automático, las tecnologías de lenguaje natural y los llamados modelos base - también conocidos como modelos de lenguaje de gran tamaño - están siendo aplicadas de manera innovadora para procesar fuentes de datos a gran escala [7]. El informe asegura que, para 2026, el 20% de las empresas utilizarán análisis de cuantificación de sentimientos en grandes flujos de datos en tiempo real para responder a eventos de mercado. Esto subraya la importancia y el potencial del uso de técnicas de ML y NLP en el ámbito financiero.

1.3 OBJETIVOS Y ALCANCE

La presente investigación tiene como objetivo principal desarrollar un modelo basado en aprendizaje automático y procesamiento de lenguaje natural para predecir las fluctuaciones diarias de los precios de las acciones en función del análisis de noticias. Esta meta se desglosa en varios objetivos secundarios, que juntos conforman la estructura de esta tesis.

En primer lugar, el *Capítulo 2* presentará una revisión exhaustiva de la literatura existente, identificando las teorías, los métodos y los resultados más relevantes en el campo de predicción de precios de acciones. Esta revisión proporcionará el soporte académico necesario para entender los desafíos y las oportunidades presentes en este ámbito de estudio.

[4] Gartner es una empresa de investigación y consultoría de tecnología de la información reconocida a nivel mundial, fundada en 1979 en Estados Unidos. Gartner es una fuente influyente y respetada para las empresas que buscan entender las tendencias tecnológicas, evaluar proveedores y productos, y tomar decisiones estratégicas de IT.

En el *Capítulo 3,* se abordará la creación de un marco conceptual y se formulará la hipótesis de trabajo. Adicionalmente, en este apartado se detallará la metodología de diseño del modelo propuesto, argumentando su viabilidad teórica.

El alcance de esta investigación abarca también la implementación y evaluación del modelo en un contexto real. Por lo tanto, en el *Capítulo 4* se pondrá en práctica el marco teórico establecido anteriormente, se recolectarán y procesarán datos de noticias y precios históricos de acciones de tres índices concretos seleccionados, y se generarán predicciones de los movimientos de sus precios.

El *Capítulo 5* se dedicará al análisis exhaustivo de los resultados obtenidos por el modelo de predicción. Mediante una evaluación cuantitativa y cualitativa del desempeño del modelo, se validará la eficacia del modelo y se intentará demostrar su capacidad para predecir las variaciones en los precios de las acciones con un grado de precisión aceptable.

Tras el desarrollo, aplicación y validación del modelo, en el *Capítulo 6* se realizará una evaluación económica del proyecto. Este análisis incluirá el cálculo de los costes de implementación y operación de la plataforma de predicción, así como una estimación de los ingresos esperados. Se buscará evidenciar el potencial del proyecto para generar un rendimiento económico significativo.

Finalmente, el *Capítulo 7* recogerá las conclusiones del proyecto, así como posibles líneas de trabajo futuro. Este último capítulo cerrará el ciclo de la investigación, reafirmando la contribución de esta tesis al campo de estudio y planteando nuevos desafíos para futuras investigaciones.

1.4 ALINEACIÓN CON LOS OBJETIVOS DE DESARROLLO SOSTENIBLE

Los objetivos de desarrollo sostenible (ODS) son un conjunto de metas globales adoptadas por la ONU en 2015 para abordar los desafíos más urgentes en materia de desarrollo sostenible. El sector energético es crítico para el desarrollo sostenible, ya que es esencial para el crecimiento económico y la reducción de la pobreza, y a su vez, juega un papel importante en la lucha contra el cambio climático.

El presente Trabajo de Fin de Grado puede contribuir a los ODS de la siguiente manera:

ODS 7: Energía asequible y no contaminante. Al analizar las noticias y comunicados de prensa relacionados con empresas energéticas, el proyecto puede ayudar a identificar tendencias y oportunidades en el sector de la energía renovable, forzando la implantación de una energía cada vez más sostenible y ampliamente disponible. El creciente uso de energía renovable y sostenible podría fomentar, a su vez, el consumo y la producción responsables (**ODS 12**). Una de las metas dentro del séptimo objetivo es la siguiente: "De aquí a 2030, ampliar la infraestructura y mejorar la tecnología para prestar servicios energéticos modernos y sostenibles para todos en los países en desarrollo" [35]. El predictor de precios jugaría un papel fundamental en la mejora de tecnología e infraestructura.

ODS 8: Trabajo decente y crecimiento económico. Al proporcionar una metodología para predecir los precios de las acciones de las empresas energéticas, el proyecto podría ayudar a los inversores y las empresas a tomar decisiones informadas que promuevan el aumento de las inversiones, de la producción de bienes y servicios, del gasto y del consumo energético. El aumento de estos indicadores de la economía de un país implica crecimiento económico, que a su vez está directamente relacionado con la disminución del desempleo (Ley de Okun).

ODS 9: Industria, innovación e infraestructura: Al contribuir al desarrollo de una metodología precisa y confiable de predicción de precios de activos, el proyecto podría ayudar a las empresas a tomar decisiones sobre inversiones en investigación y desarrollo en tecnologías de energía limpia y sostenibles. Según las Naciones Unidas, "la innovación y el progreso tecnológico son claves para descubrir soluciones duraderas para los desafíos económicos y medioambientales, como el aumento de la eficiencia energética y de recursos" [35]. Por lo tanto, la herramienta permitiría mayor inversión en I+D, lo cual se traduce en innovación y desarrollo de infraestructura en el sector.

CAPÍTULO 2. ESTADO DE LA CUESTIÓN

La predicción del precio de las acciones ha suscitado recientemente un interés significativo entre inversores y analistas profesionales. La complejidad del mercado y de la fluctuación de los precios de las acciones se encuentra intrínsecamente ligada a diversos factores, incluyendo eventos políticos, noticias del mercado, informes de ganancias trimestrales y comportamientos de comercio contradictorios. Ante este panorama, no es extraño que los inversores recurran a indicadores técnicos vinculados con las acciones como herramientas de apoyo en sus decisiones. Sin embargo, y pese a que dichos indicadores están disponibles, la tarea de predecir las tendencias del mercado suele ser un desafío, dada la volatilidad y variabilidad que caracterizan a los mercados bursátiles.

A pesar de los recientes avances, la hipótesis del mercado eficiente es una de esas teorías que sigue generando debate en la vanguardia de la economía y las finanzas. Esta hipótesis, propuesta inicialmente por el economista Eugene Fama en 1970, sostiene que todos los participantes en un mercado financiero tienen acceso a la misma información en el mismo momento, por lo que los precios de los activos financieros siempre reflejan la información disponible más actual y relevante [15]. Esto implica que ningún inversionista podría superar sistemáticamente el rendimiento del mercado en su totalidad, dado que todos tienen la misma capacidad de procesar información y ajustar los precios de los activos en consecuencia.

La hipótesis del mercado eficiente se basa en la idea del "paseo aleatorio", que es una metáfora financiera para describir el patrón de los precios de las acciones. Según esta metáfora, los cambios futuros en el precio de una acción son independientes de los cambios pasados y son impredecibles, ya que cada nueva pieza de información afecta al precio de una manera imprevisible e independiente.

Sin embargo, esta teoría ha sido objeto de mucha controversia. Muchos expertos financieros argumentan que los mercados financieros no siempre son eficientes, y que existen "anomalías" en el mercado que pueden permitir a algunos inversores lograr rendimientos superiores al promedio del mercado. Como se explica a continuación, numerosos estudios empíricos cuestionan la validez de la hipótesis del mercado eficiente, demostrando que los precios de las acciones pueden verse influenciados por factores históricos, como los patrones de precios anteriores de un activo, así como por factores irracionales y psicológicos.

Algunos estudios que han cuestionado la validez de la hipótesis del mercado eficiente toman un enfoque puramente cuantitativo. Por ejemplo, un estudio realizado en los mercados de valores de Kuwait y Arabia Saudí entre 1985 y 1989 reveló que los 35 valores sauditas mostraban una desviación significativa del paseo aleatorio. Mediante regresiones lineales, se identificaron factores institucionales que estaban contribuyendo a la ineficiencia del mercado, como la falta de liquidez, la fragmentación del mercado, y los retrasos en la negociación y la notificación [6].

Otro estudio evaluó cómo los componentes temporales y permanentes de los movimientos de los precios de las acciones pueden estar relacionados con las perturbaciones macroeconómicas de oferta y demanda, haciendo uso de la regresión múltiple. Gallagher et al. encontraron que los choques de demanda tienen solo efectos temporales sobre los precios reales de las acciones, mientras que los choques de oferta pueden afectar el nivel de los precios reales de las acciones de forma permanente [19].

Haciendo uso del mismo método de predicción, un análisis multivariante de la eficiencia del mercado de acciones en la Bolsa de Atenas (ASE) volvió a refutar la hipótesis del mercado eficiente. Se concluyó que la ASE es informacionalmente ineficiente, lo que implica que los precios de las acciones pasadas contienen cierta información sobre los futuros movimientos de los precios en los que los inversores pueden actuar [24].

Además de las técnicas de regresión, otros métodos cuantitativos han sido utilizados para predecir los precios de las acciones. Una de las técnicas más tradicionales en la serie temporal es el Modelo Autorregresivo Integrado de Media Móvil (ARIMA)[5]. En este sentido, un estudio realizado por Ariyo et al. aplicó el modelo ARIMA a la predicción de los precios de las acciones utilizando datos de la Bolsa de Valores de Nueva York (NYSE) y la Bolsa de Valores de Nigeria (NSE). En este estudio, los investigadores presentaron un proceso exhaustivo de construcción de un modelo predictivo ARIMA, cuyos resultados revelaron que el modelo tiene un gran potencial para la predicción a corto plazo y puede competir favorablemente con otras técnicas existentes para la predicción de precios de acciones [3].

[5] Del inglés, Autoregressive Integrated Moving Average

También se han llevado a cabo estudios que combinan dos enfoques distintos en su modelo de predicción. Un ejemplo de ello es un estudio de Drachal, que combina la estadística bayesiana con un modelo de mezclas finitas dinámicas (DFM)[6] en un intento de predecir los precios de los principales combustibles fósiles: petróleo, gas natural y carbón [12]. El modelo de mezclas finitas dinámicas es un tipo de modelo de mezcla estadística que permite la existencia de diferentes patrones en los datos. El DFM supone que los datos provienen de varias distribuciones diferentes, cada una representando un "componente" diferente del conjunto de datos, y se usa para representar la heterogeneidad en los datos. La estadística bayesiana, por otro lado, se utiliza para realizar la inferencia sobre los parámetros del modelo de mezclas finitas dinámicas. La inferencia bayesiana permite utilizar información a priori (conocimientos previos) sobre los parámetros y actualizar esta información a medida que se obtienen nuevos datos. El uso simultáneo de ambas técnicas conforma una herramienta valiosa para la predicción de precios y ofrece mayor precisión que otros modelos.

Si nos adelantamos un poco en el tiempo, en las últimas décadas se han desarrollado nuevas metodologías basadas en la inteligencia artificial y el aprendizaje automático. Debido a su buen rendimiento en otros campos, estas técnicas han despertado un gran interés en el campo de predicción del mercado de valores [8]. Algunas de las metodologías más prometedoras incluyen son los distintos tipos de redes neuronales artificiales, los bosques aleatorios, las máquinas de vectores soporte, los árboles de decisión, los árboles potenciados por gradiente y los vecinos más próximos (k-nearest neighbors) [46]. Estas metodologías representan alternativas novedosas y ventajosas, pero se debe tener en cuenta que el rendimiento de los modelos de aprendizaje automático también está vinculado a la calidad de los datos de entrada empleados para construir el modelo.

En la línea del aprendizaje automático, estudios más avanzados se han centrado en el desarrollo de modelos predictivos mediante redes neuronales. Un ejemplo notable de esto es el trabajo de Wang et al., quienes emplearon Redes Neuronales Recurrentes (RNN) para analizar las fluctuaciones de los precios del petróleo [48]. Las RNN son una clase especial de redes neuronales que añaden 'memoria' a la red mediante bucles en las neuronas. Esta característica permite que la RNN tenga en cuenta la información de los datos temporales previos para hacer una predicción, lo cual es particularmente útil para las series temporales como los precios de

[6] Del inglés, Dynamic Factor Model

las acciones. En cuanto al algoritmo "Backpropagation Through Time" (BPTT), se trata de una variante de la popular técnica de "Backpropagation" utilizada para entrenar redes neuronales estándar. En esencia, BPTT implica desenrollar toda la secuencia de datos temporales y aplicar la retropropagación estándar. A través de este proceso, se actualizan los pesos de la red neuronal, mejorando la precisión del modelo con cada iteración de entrenamiento.

De manera similar, Ticknor introduce un enfoque eficaz basado en Redes Neuronales Artificiales (ANN) con regularización [47]. Las ANN son sistemas computacionales inspirados en las redes neuronales biológicas que constituyen el cerebro humano. Estas redes se componen de nodos o "neuronas" conectadas que trabajan conjuntamente para generar una salida a partir de múltiples entradas, y tienen la capacidad de aprender y mejorar con el tiempo al ajustar las ponderaciones de las conexiones entre las neuronas. En el estudio de Ticknor, se utiliza un método de regularización bayesiana para evitar el sobreajuste de la ANN. El sobreajuste se produce cuando un modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad para generalizar a partir de nuevos datos. La regularización bayesiana ayuda a controlar este fenómeno al añadir una penalización al tamaño de los pesos en la ANN.

En la misma línea de estudio, García-Quintero et al. emplean una Red Neuronal Artificial (RNA) de tipo feedforward para entrenar un modelo utilizando datos históricos de los precios del oro, con el objetivo de evaluar su capacidad para predecir precios futuros [20]. Las redes neuronales feedforward, también conocidas como perceptrones multicapa, son una de las estructuras más comunes de RNA. En ellas, la información se mueve en una única dirección, desde la capa de entrada, pasando por las capas ocultas, hasta la capa de salida, sin bucles de retroalimentación. Este tipo de RNA destaca por su capacidad para aprender y modelar relaciones no lineales. Los resultados del estudio demuestran que el modelo logra una alta precisión en la predicción de precios, lo que indica que la RNA puede ser una herramienta útil para la toma de decisiones en el mercado bursátil.

Como se puede observar por los estudios ya mencionados, debido a la formación en Economía de los investigadores, la mayoría de los estudios realizados en el campo de la predicción se basan en metodologías estadísticas de series temporales que utilizan datos históricos para pronosticar la evolución de los precios de las acciones. A pesar de la variedad de formatos de los datos, los cuales abarcan desde precios de índices bursátiles hasta rendimientos, volatilidad y tasas de interés, los datos de entrada empleados son puramente cuantitativos.

Sin embargo, la teoría de las finanzas conductuales sostiene que los inversores toman decisiones basadas en emociones y estados de ánimo, y, por lo tanto, cabe la posibilidad de que tomen decisiones irracionales. Esto quiere decir que el mercado no es "totalmente eficiente", como sugiere la hipótesis del mercado eficiente. Según las finanzas conductuales, el mercado se ve afectado por un concepto conocido como "sentimiento del inversor", que refleja el estado de ánimo o la actitud del público hacia un activo. Este descubrimiento ha llevado a la reciente inclusión de métodos cualitativos en modelos financieros basados en la predicción de precios.

En los últimos años, la incorporación del sentimiento del inversor como una característica en los modelos predictivos se ha vuelto cada vez más popular. Muchos artículos reportan mejores resultados mediante la fusión de diferentes modelos [33]. Esta inclusión de datos cualitativos ofrece una visión más completa de los factores que afectan a los precios de las acciones, permitiendo desarrollar modelos predictivos más sofisticados y precisos.

Algunos estudios se basan en la teoría de que la información contenida en la narrativa difusa y otras fuentes de información desestructurada, como las redes sociales o las noticias financieras, puede ser utilizada para predecir los precios del mercado de valores. En el contexto de las empresas energéticas, los autores Afkhami et al. utilizan datos de búsqueda de Google para identificar las palabras clave que mejor predicen la volatilidad del precio de los combustibles, como petróleo, gasolina convencional y gas natural [1]. Los resultados sugieren que existen determinadas palabras clave con una gran capacidad predictiva de los precios. Esas palabras clave están más relacionadas con eventos geopolíticos y económicos, que con eventos meteorológicos o de demanda.

Del mismo modo, diversos investigadores han buscado comprobar si las redes sociales como Twitter afectan a la volatilidad, al volumen de negociación y a los precios diarios de las acciones de determinadas empresas. Un estudio reciente de Zaman et al. concluye que los tweets de Elon Musk aumentaron los precios de Bitcoin [51]. Siguiendo con la misma red social, Constantino et al. examinan el impacto que tiene el sentimiento de los tweets en la volatilidad de las acciones de energías renovables [11]. Mediante una técnica de procesamiento de lenguaje natural, los autores demuestran que las variables de sentimiento tienen información relevante adicional que no puede ser capturada por las variables financieras estándar. En esta misma línea se han llevado a cabo numerosos estudios, y todos ellos han revelado que existe una elevada correlación entre las redes sociales y las fluctuaciones del mercado bursátil de las energéticas [14][25][49].

El impacto que tienen determinados datos de texto en los precios de las acciones de este sector queda claramente reflejado en el estudio realizado por Liu [29]. El autor comprueba empíricamente que el sentimiento negativo asociado a la pandemia del COVID-19 afecta negativamente a los precios de energía en el mercado. Paralelamente, los resultados exponen la relación directa existente entre los precios del sector energético y los precios del mercado total en China, lo cual demuestra que el desempeño financiero de las empresas energéticas tiene un impacto directo en la economía global. Este último hallazgo se convierte en un pilar fundamental para el presente Trabajo de Fin de Grado, el cual enfoca su atención en el sector energético, dada su probada influencia significativa en otros sectores económicos.

Otros estudios, como el de Banerjee et al., han utilizado minería de texto para demostrar que el sentimiento de las noticias relacionadas con la Organización de Países Exportadores de Petróleo (OPEC) tiene un impacto significativo en los rendimientos de las acciones de las empresas energéticas en los Estados Unidos [4]. Del mismo modo, estudios más específicos estiman el impacto de las noticias de catástrofes naturales en los mercados de capitales estadounidenses [17]. Los resultados sugieren que los huracanes tienen un impacto significativo en los retornos de las acciones de las empresas energéticas y que el impacto puede ser positivo o negativo, dependiendo de la gravedad y la ubicación del huracán. Tras una catástrofe, los precios de las renovables registran las mayores ganancias, mientras que las acciones de empresas del carbón pierden valor.

La combinación de procesamiento del lenguaje y aprendizaje automático podría ser prometedora debido a su doble capacidad para extraer información relevante de texto no estructurado y aprender de los patrones dentro de estos datos para hacer predicciones. Sin embargo, aún no se han desarrollado muchas herramientas de predicción que utilicen la fusión de ambas técnicas. Uno de los escasos ejemplos es YUKKA Lab, una empresa de inteligencia artificial y análisis de datos que ofrece soluciones para el mercado financiero. Se especializa en el análisis de datos de mercado, el análisis de sentimiento, la extracción de información y la creación de modelos predictivos. Esta compañía de software ha sido incluida por los analistas de Gartner en sus informes de Cool Vendors, es decir, ha sido identificada como una de las empresas emergentes innovadoras en el área del Big Data en al año 2022 [7]. La plataforma de YUKKA procesa miles de millones de noticias en bruto y, mediante técnicas de extracción de información y detección de sentimientos, puede reconocer los acontecimientos en los que están implicadas entidades relevantes, además de obtener información procesable sobre las

tendencias y permitir a los inversores predecir los puntos de inflexión en los mercados. También permite detectar riesgos potenciales e identificar lo antes posible las próximas oportunidades de venta.

En definitiva, la aplicación de análisis semántico a los datos financieros es una tarea no trivial, ya que las noticias y las redes sociales contienen un número significativo de sarcasmos, metáforas y términos específicos del ámbito energético [33]. Sin embargo, y a pesar del creciente interés por incorporar datos cualitativos en el análisis de los mercados financieros, es evidente que la investigación es aún insuficiente en lo que respecta a modelos de predicción que utilizan exclusivamente datos cualitativos como insumo. Es más, hasta donde llega el conocimiento del autor, no se ha explorado la predicción del mercado de valores utilizando como variable predictiva el análisis de la frecuencia de aparición de palabras en los titulares de las noticias. Más aún, los estudios existentes tienden a enfocarse en noticias y datos que están intrínsecamente relacionados con el sector que se está estudiando. En este sentido, se observa una marcada ausencia de investigaciones que busquen predecir los precios de las acciones del sector energético utilizando noticias de carácter general y no necesariamente ligadas directamente al sector.

Esta omisión en la literatura científica destaca una oportunidad única para explorar y posiblemente abrir un nuevo campo de estudio. La perspectiva de construir un modelo de predicción que se base enteramente en la frecuencia de aparición de palabras en titulares de noticias de carácter general, y que busque prever los movimientos en los precios de las acciones de un sector tan vital como el energético, aporta un enfoque totalmente innovador y desafiante.

Este horizonte inexplorado no solo desafía las aproximaciones tradicionales de análisis de los mercados financieros, sino que además plantea una pregunta intrigante: ¿Podría ser que las claves para predecir el comportamiento del sector energético, e incluso de la economía global, se encuentren escondidas en nuestras noticias diarias? Este es el desafío que se abordará en el presente Trabajo de Fin de Grado.

CAPÍTULO 3. PLANTEAMIENTO DEL MODELO

En este capítulo, se presenta el modelo teórico utilizado para predecir los precios de las acciones basado en el análisis de la narrativa de las noticias. Este modelo se fundamenta en el enfoque de aprendizaje automático supervisado, y busca establecer una función de mapeo entre las características de entrada, es decir, la frecuencia de palabras clave en las noticias, y la salida, la variación de precios de las acciones. Se presentan los detalles del modelo, las hipótesis de trabajo y la metodología empleada para su implementación.

3.1 MODELO TEÓRICO

El modelo propuesto se basa en la intersección entre dos disciplinas: la ingeniería financiera y la ciencia del lenguaje natural. El objetivo central de este trabajo es proponer una metodología matemática y estadística para entender cómo la narrativa, tal como se refleja en la frecuencia de aparición de las palabras en las noticias, puede usarse para predecir los precios de las acciones. Para construir y describir el modelo, se hace uso de dos enfoques distintos pero complementarios: la regresión lineal y las redes neuronales. Ambos enfoques permiten representar relaciones entre variables y hacer predicciones a partir de los datos, pero difieren en cuanto a su capacidad para capturar y modelar patrones de relaciones lineales y no lineales, respectivamente.

3.1.1 MODELO DE REGRESIÓN LINEAL

En esta sección se presenta el modelo teórico más simplificado para el análisis de las fluctuaciones en los precios de las acciones basado en la narrativa periodística.

Empezamos con la suposición de que la frecuencia de aparición de palabras clave en las noticias, que designaremos como vector Alfa (α), tiene un impacto en la variación de los precios de las acciones, que llamaremos ΔP.

El conjunto de características de entrada α se define como un vector de características $\alpha = [\alpha_1, \alpha_2, ..., \alpha_n]$, donde cada α_i es la frecuencia relativa de aparición de una palabra clave específica

en las noticias recopiladas de un día. Por ejemplo, α_1 podría representar la frecuencia relativa de aparición de la palabra "coronavirus", α_2 la de la palabra "guerra", y así sucesivamente.

El valor de salida ΔP es una variable continua que representa la variación de precios de las acciones de una empresa en un período de tiempo determinado, en este caso, en el mismo día en el que se publican las noticias recopiladas. La tarea es encontrar una función f tal que $f(\alpha) \approx \Delta P$, que mapee el conjunto de características de entrada α al valor de salida ΔP.

La relación entre las palabras clave y la variación de los precios se formaliza a través de un modelo de regresión lineal múltiple. En este enfoque, se asume una relación lineal entre el vector de frecuencias de aparición de palabras clave (α) y la variación del precio del activo (ΔP). El modelo se construye de la siguiente manera [18]:

$$\Delta P = \beta_0 + \sum(\beta_i \alpha_i) + \varepsilon \tag{3.1}$$

donde:

- ΔP es la variación en los precios de las acciones,
- β_0 es el coeficiente de intercepción o término independiente, que indica el valor esperado de ΔP cuando todos los α_i son iguales a cero,
- β_i son los coeficientes asociados a cada palabra clave, que determinan cuánto influye una unidad de cambio en la frecuencia relativa de la palabra clave i-ésima (α_i) sobre ΔP,
- ε es el error aleatorio que tiene una distribución normal con media 0. Este error capta otros factores no considerados por el modelo pero que pueden influir en la variación de precios de las acciones.

A continuación, se muestra la *Figura 3.1*, cuyo objetivo no es más que el de facilitar la visualización del modelo de regresión lineal que se acaba de describir. Las cajas verdes representan la frecuencia de aparición de cada palabra clave, en una caja azul se muestra la variación en los precios de las acciones, y en una caja roja se incluye el término de error. Este último indica que hay otros factores no representados en el modelo que también pueden afectar a la variación de precios. Por último, cada flecha representa la influencia de una palabra clave

en la variación de precios, y el coeficiente beta correspondiente sobre cada flecha (β_1, β_2, ..., β_n) indica cuánto contribuye cada una de las palabras.

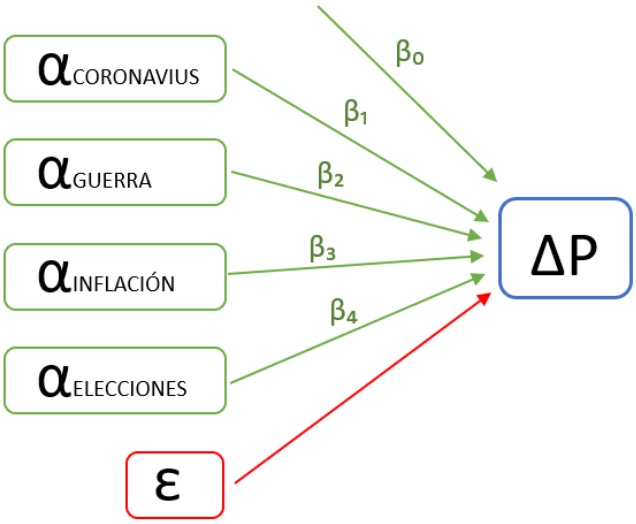

Figura 3.1. Modelo visual de Regresión Lineal (elaboración propia, 2023)

La tarea de aprendizaje automático consiste en encontrar los coeficientes β que minimizan la función de coste [30], es decir, que minimizan el error cuadrático medio (MSE)[7]:

$$MSE = \frac{\sum(y_i - f(\alpha_i))^2}{N} \tag{3.2}$$

donde y_i es el valor observado de ΔP, $f(\alpha_i)$ es el valor pronosticado de ΔP por el modelo, y N es el número total de observaciones.

Para comprender mejor el concepto de Error Cuadrático Medio, primero necesitamos entender qué es el error en una regresión lineal. Para ello, se utiliza como ejemplo la *Figura 3.2*, en la cual se representa una línea de regresión (ilustrada en azul) para intentar predecir los datos existentes (representados por los puntos verdes). Este modelo lineal posee un cierto grado de error para cada estimación (indicado en rojo) que se puede expresar como:

$$Error = valor\ real - valor\ estimado$$

[7] Del inglés, Mean Squared Error

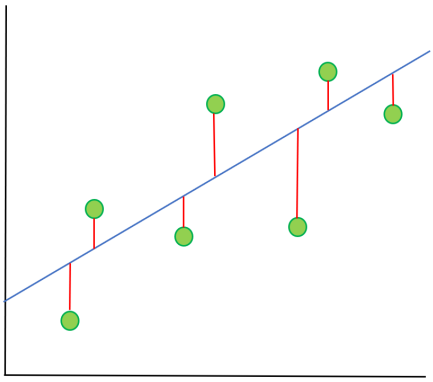

Figura 3.2. Error de la Regresión Lineal (elaboración propia, 2023)

Al elevar el error al cuadrado, garantizamos que el error sea siempre positivo, lo que nos indica que el error perfecto es 0. Tras calcular el error cuadrático de cada estimación individual, podremos calcular el error medio, que es el promedio de estos errores cuadráticos.

Una vez introducido el error cuadrático, estamos ahora en una mejor posición para entender cómo se busca minimizar el MSE mediante el método de los Mínimos Cuadrados Ordinarios (OLS)[8]. En esencia, lo que el método OLS busca reducir es la Suma de los Errores al Cuadrado (SSE)[9], no el MSE como tal [21]. El SSE se define como:

$$SSE = \sum(y_i - f(\alpha_i))^2 \tag{3.3}$$

Formalmente, el problema de optimización es:

$$\text{minimizar SSE respecto a } \beta = [\beta_0, \beta_1, ..., \beta_n]$$

Para resolver el problema, se toma la derivada de la función de coste (en este caso, la SSE) respecto a cada uno de los β_i y se iguala a cero. Resolviendo este sistema de ecuaciones, se obtienen los valores de β que minimizan la SSE.

No obstante, si detectamos multicolinealidad entre las frecuencias de aparición de las palabras clave, es decir, una correlación fuerte entre ellas, debemos recurrir a métodos de regularización como la Regresión de Crestas o LASSO[10] [34]. Estos dos métodos de regularización, además de lidiar con la multicolinealidad en la regresión lineal, se utilizan para prevenir el sobreajuste.

[8] Del inglés, Ordinary Least Squares
[9] Del inglés, Sum of Squared Errors
[10] Del inglés, Least Absolute Shrinkage and Selection Operator

Por un lado, la Regresión de Crestas agrega un término de penalización a la función de coste de la regresión lineal. La idea es que, al aumentar esta penalización, los coeficientes de regresión se reducen, lo que conduce a un modelo más simple y menos propenso al sobreajuste. La función de coste para la regresión de crestas es:

$$\min \ [\sum(y_i - f(\alpha_i))^2 + \lambda\sum\beta_i^2] \tag{3.4}$$

donde λ es un parámetro de regularización que controla la magnitud de la penalización.

Por otro lado, el método LASSO también agrega un término de penalización a la función de coste, pero en este caso, la penalización es la suma de los valores absolutos de los coeficientes de regresión. Este enfoque puede llevar a que algunos coeficientes de regresión sean exactamente cero, lo que equivale a excluir esas variables del modelo. Esta es la razón por la que LASSO también se puede utilizar para la selección de características.

La función de coste para LASSO es:

$$\min \ [\sum(y_i - f(\alpha_i))^2 + \lambda\sum|\beta_i|] \tag{3.5}$$

Al igual que con la regresión de crestas, λ es un parámetro de regularización que controla la magnitud de la penalización.

Para ambos métodos, un valor de λ demasiado grande puede llevar a un modelo demasiado simple que subajuste los datos, mientras que un valor de λ demasiado pequeño puede no tener mucho efecto en la prevención del sobreajuste [31]. El valor óptimo de λ generalmente se selecciona a través de la validación cruzada.

La validación cruzada es el último concepto que se va a explicar en la sección de modelo teórico de regresión lineal. Se trata de una técnica estadística robusta esencialmente empleada para evaluar el rendimiento predictivo de un modelo. Su objetivo es asegurar que el modelo no se esté ajustando demasiado a los datos de entrenamiento, fenómeno conocido como sobreajuste, el cual puede llevar a un pobre rendimiento del modelo al enfrentarse a datos nuevos o desconocidos.

En la validación cruzada, el conjunto de datos se divide en 'k' subconjuntos o 'folds'. Después, el modelo se entrena 'k' veces, cada vez utilizando un subconjunto distinto como conjunto de validación y el resto de los datos como conjunto de entrenamiento. Por lo tanto, cada

subconjunto tiene la oportunidad de ser el conjunto de validación una vez, mientras que los datos restantes sirven para entrenar el modelo. Esta idea se ve claramente reflejada en la *Figura 3.3*. Finalmente, el rendimiento del modelo en cada una de las 'k' iteraciones se promedia para obtener una medida de la eficacia general del modelo.

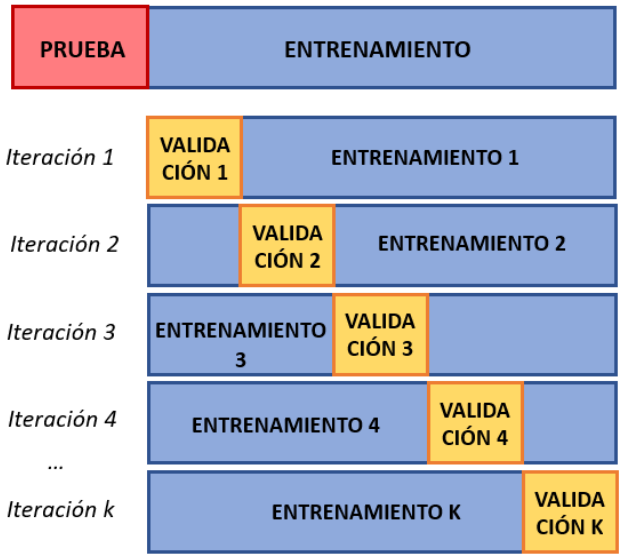

Figura 3.3. División de Datos para Validación Cruzada (elaboración propia, 2023)

La elección de 'k' es un equilibrio entre el sesgo y la varianza del estimado del error de predicción. Un valor comúnmente utilizado es k=10, también conocido como validación cruzada de 10 folds [27]. Sin embargo, cuando los datos son escasos, se puede utilizar una variante llamada 'validación cruzada de exclusión única' (LOO)[11], en la que 'k' es igual al tamaño de la muestra.

Al aplicar la validación cruzada en el contexto de la Regresión de Crestas y LASSO, se obtiene una medida más fiable del rendimiento del modelo al variar el parámetro de penalización λ. Esto permite identificar el valor óptimo de λ que minimiza el error de validación, proporcionando el mejor equilibrio entre sesgo y varianza para las predicciones.

Este modelo simple de regresión lineal es nuestro punto de partida y proporciona una buena base para entender la relación entre el número de veces que aparecen las palabras en las noticias

[11] Del inglés, Leave One Out

de un día, y la variación de los precios de las acciones de una entidad para ese mismo día. Sin embargo, es importante reconocer que este modelo tiene sus limitaciones.

Primero, asume que esta relación es lineal y constante en el tiempo, lo cual puede no ser cierto en los mercados financieros, que son conocidos por su complejidad y dinamismo. Segundo, puede haber retrasos en la respuesta del mercado a las noticias, que nuestro modelo actual no puede captar.

Para abordar estas limitaciones, en la próxima sección se avanza hacia un enfoque más sofisticado, utilizando redes neuronales que pueden captar relaciones no lineales y aprender representaciones más complejas de las palabras clave.

3.1.2 MODELO DE REDES NEURONALES

Como se ha visto hasta ahora, el proceso de aprendizaje del modelo se realiza utilizando un conjunto de datos históricos que consiste en una serie temporal de precios de acciones y las noticias asociadas a cada período de tiempo. Aunque en la práctica el modelo se desarrollará y aplicará utilizando una regresión lineal, el modelo teórico se puede ampliar a través de una visión más sofisticada y potente del aprendizaje automático: las redes neuronales. En este enfoque, en lugar de confiar en un mapeo lineal de las entradas a las salidas, se puede considerar una serie de transformaciones no lineales que permiten modelar interacciones más complejas. Ahora, la idea central de nuestro modelo es que la variación en los precios de las acciones puede ser predicha como una función no lineal de las frecuencias relativas de las palabras clave en las noticias financieras, y que esta función puede ser aprendida a partir de los datos mediante el entrenamiento de una red neuronal artificial.

Las redes neuronales artificiales son sistemas de cálculo inspirados en la estructura del cerebro humano. Una red neuronal consiste en un conjunto de nodos interconectados, o "neuronas", que trabajan juntas para generar una salida a partir de un conjunto de entradas. La arquitectura de la red se estructura en capas: una capa de entrada que recibe los datos, una o más capas ocultas que procesan los datos, y una capa de salida que produce el resultado.

En términos matemáticos, cada neurona i en una capa oculta l realiza un cálculo que puede ser representado como [37]:

$$h_i^l = \sigma\left(\sum_j w_{ij}^l \times h_j^{l-1} + b_i^l\right) \quad (3.6)$$

donde:

- h_i^l es la activación de la neurona i en la capa l,
- w_{ij}^l es el peso sináptico que conecta la neurona j en la capa l-1 con la neurona i en la capa l,
- b_i^l es el sesgo de la neurona i en la capa l,
- σ es la función de activación, que introduce no linealidad en el modelo. Comúnmente se usan funciones como la sigmoide, la tangente hiperbólica o la unidad lineal rectificada (ReLU)[12].

Los pesos sinápticos w y los sesgos b son los parámetros del modelo, que se aprenden durante el entrenamiento. Este proceso se lleva a cabo a través de un algoritmo de optimización, como el descenso de gradiente estocástico, el cual se definirá más adelante, en esta misma sección.

La aplicación de redes neuronales como método de aprendizaje automático permite capturar interacciones más complejas entre las palabras que aparecen en los titulares y subtitulares de las noticias, y adaptar de manera más flexible el modelo a los datos, lo cual puede mejorar la precisión de las predicciones.

Tras esta breve introducción al funcionamiento general de las redes neuronales, se procede a su aplicación en el contexto de estudio teórico. Aunque el desarrollo de este método de aprendizaje automático es un proceso iterativo y experimental que requiere probar diferentes configuraciones para obtener el mejor rendimiento posible, a continuación, se lleva a cabo un diseño preliminar básico para determinar la arquitectura de la red neuronal del modelo:

Supongamos que contamos con un conjunto de *m* palabras clave, es decir, $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_m\}$, y queremos predecir la variación en el precio de las acciones ΔP para un día dado.

[12] Del inglés, Rectified Linear Unit

La red neuronal propuesta tiene una estructura multicapa [10], consistente en una capa de entrada que recibe las frecuencias relativas de las palabras clave, una o más capas ocultas que procesan los datos, y una capa de salida que produce el resultado.

- La **capa de entrada** tiene *m* nodos, correspondientes a las *m* palabras clave extraídas de las noticias de un día. Cada nodo i en esta capa recibe la frecuencia relativa de la palabra clave α_i

- Las **capas ocultas** se componen de un número de neuronas que puede ser afinado durante el proceso de desarrollo del modelo. Estas capas procesan las entradas a través de una serie de transformaciones no lineales. Cada neurona i en una capa oculta l realiza un cálculo anteriormente definido en la *Ecuación 3.6*

- La **capa de salida** está compuesta por una sola neurona, que representa la predicción de la variación en el precio de las acciones ΔP. Esta capa produce el resultado final del modelo, y su cálculo puede ser representado como:

$$\Delta P = \sigma(\sum_i w_i x_i + b) \tag{3.7}$$

donde los x_i son las salidas de las neuronas en la última capa oculta, y los w_i son los pesos sinápticos que conectan las neuronas en la última capa oculta con la neurona en la capa de salida.

Tal y como se muestra en la anterior *Ecuación 3.6*, cada neurona en las capas ocultas realiza una combinación lineal ponderada de sus entradas, seguida de una transformación no lineal a través de la función de activación. El número de capas ocultas y el número de neuronas en cada una de estas capas son hiperparámetros que deben ser afinados. Se puede comenzar suponiendo que hay una sola capa oculta con un número de neuronas igual a la media geométrica del número de nodos en la capa de entrada y la capa de salida. Esta es una heurística comúnmente usada que puede proporcionar un buen punto de partida [23][44]. Por lo tanto, si tenemos 50 palabras clave en un día, lo que da un total de 50 nodos de entrada, podríamos empezar con una capa oculta de aproximadamente 7 nodos (la raíz cuadrada de 50*1, ya que tenemos un solo nodo de salida). La arquitectura de la red neuronal descrita quedaría como se muestra en la *Figura 3.4*.

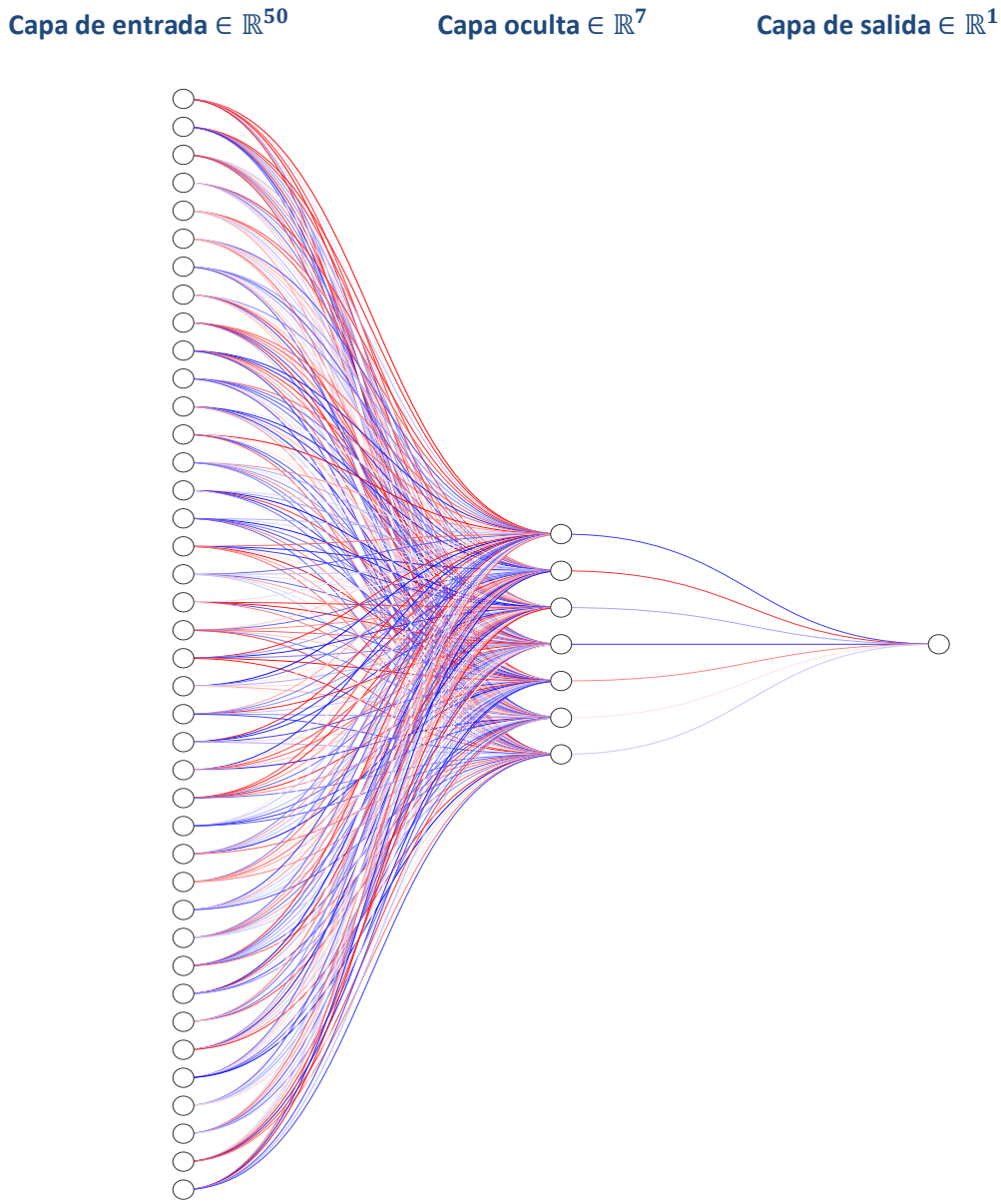

Figura 3.4: Arquitectura preliminar de la red neuronal (elaboración propia, herramienta NN-SVG, 2023)

En el diagrama de la red neuronal de la *Figura 3.4*, las curvas representan las conexiones o pesos entre las neuronas. Los pesos en una red neuronal determinan la contribución de una entrada o neurona a la siguiente y, por lo tanto, son un componente crucial de la red. Al colorear las curvas en función de sus pesos, se facilita la identificación de las conexiones más fuertes y más débiles en la red. En este caso, el color azul se utiliza para representar los pesos negativos, que implican que un aumento en la entrada correspondiente resultará en una disminución en la salida de la neurona, y viceversa. El color rojo se utiliza para los pesos positivos, que indican que un aumento en la entrada correspondiente resultará en un aumento en la salida de la

neurona. El color negro se utiliza como color por defecto, para los pesos que están cerca de cero y tienen poco o ningún efecto en la salida de la neurona.

Además, cabe mencionar que al elaborar la figura se han utilizado las curvas de Bézier, una herramienta matemática comúnmente utilizada en gráficos por ordenador para generar curvas suaves y controlables [22]. En el diagrama, las curvas de Bézier trazan las trayectorias de las conexiones entre las neuronas, dando una sensación de "flujo" a través de la red, lo que ayuda a visualizar cómo la información se propaga desde la entrada hasta la salida.

El número de capas ocultas y el número de neuronas en cada capa luego pueden ser ajustados a través de un proceso de validación cruzada, donde se entrena la red con diferentes configuraciones y se selecciona la que produzca el menor error de validación, tal y como se ha explicado en la *Sección 3.1.1*.

Además de la selección del número de capas y neuronas, el modelo requiere una serie de decisiones detalladas y técnicas que incluyen la elección de la función de activación y el ajuste de los hiperparámetros del algoritmo de optimización.

Para la función de activación, la Unidad Lineal Rectificada (ReLU) es una elección común para las capas ocultas debido a su simplicidad y efectividad. Esta función de activación no lineal tiene la forma:

$$f(x) = \max(0, x) \tag{3.8}$$

Es decir, devuelve el valor de entrada si este es positivo y cero en caso contrario. La función ReLU es popular porque, además de ser fácil de calcular, favorece la dispersión en la red, lo que puede ayudar a evitar problemas de sobreajuste [37]. Para la capa de salida no se necesita una función de activación, ya que la ausencia de esta permite que la red produzca una salida no acotada, requisito imprescindible para la regresión.

En cuanto a la optimización, el uso del Descenso de Gradiente Estocástico (SGD)[13] y su variante mejorada, Adam[14], se hace crucial [2]. Ambos se centran en minimizar la función de coste, que en nuestro caso está definida por el Error Cuadrático Medio (MSE) presentado en la

[13] Del inglés, Stochastic Gradient Descent
[14] Acrónimo para "Adaptive Moment Estimation"

sección 3.1.1, el cual proporciona una medida cuantitativa de la discrepancia entre las predicciones del modelo y los valores reales. Estos algoritmos actualizan iterativamente los pesos de la red en la dirección contraria al gradiente de la función de coste, permitiendo una vía de mejora continua del modelo.

Adam, como mejora del SGD, incorpora el concepto de Momento y la Tasa de Aprendizaje Adaptativa (AdaGrad). Esto significa que cada actualización de los pesos en la red se realiza utilizando un promedio ponderado de los gradientes pasados y su cuadrado, proporcionando así ajustes más precisos.

La tasa de aprendizaje es el hiperparámetro más importante de cualquier algoritmo de optimización, pues controla la velocidad a la que el algoritmo ajusta los pesos de la red. La elección de este parámetro se realiza mediante la validación cruzada, optimizando así el equilibrio entre aprendizaje efectivo y riesgo de sobreajuste.

Aun así, el proceso que realmente desempeña un papel crucial en la prevención del sobreajuste es la regularización. Las técnicas de regularización, como L_1 y L_2, se utilizan para agregar una penalización a la función de coste, de manera que el algoritmo evite ajustarse excesivamente a los datos de entrenamiento [39]. En concreto, la regularización L_1 añade un término que es proporcional al valor absoluto de los coeficientes de los pesos (la norma L_1), mientras que la regularización L_2 añade un término que es proporcional al cuadrado de los coeficientes de los pesos (también conocida como la norma L_2).

Si suponemos que λ_1 y λ_2 son los hiperparámetros para las regularizaciones L_1 y L_2 respectivamente, la función de coste se modifica de la siguiente manera:

$$J'(\theta) = J(\theta) + \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2{}^2 \tag{3.9}$$

donde $J(\theta)$ es la función de coste original (MSE), $\|\theta\|_1$ es la norma L_1 de los pesos, y $\|\theta\|_2{}^2$ es la norma L_2 de los pesos.

Otra técnica de regularización específica de las redes neuronales es el Dropout [39]. Durante cada iteración del entrenamiento, una proporción aleatoria de las neuronas y sus conexiones correspondientes son "apagadas" temporalmente Esto ayuda a asegurar que la red neuronal no dependa demasiado de ninguna característica de entrada particular, mejorando así la generalización y reduciendo el sobreajuste. Matemáticamente, esto se modela multiplicando la

salida de cada neurona por una variable aleatoria de Bernoulli, que toma el valor 1 con probabilidad 1-p y 0 con probabilidad p, donde p es el parámetro de Dropout. Si suponemos que en este modelo usamos un valor de Dropout de 0.5 en las capas ocultas, esto significa que aproximadamente la mitad de las neuronas se "apagan" durante cada iteración de entrenamiento.

Finalmente, el tamaño de lote y el número de épocas son otros hiperparámetros importantes en nuestro modelo. El tamaño de lote (Batch Size) controla la cantidad de ejemplos de entrenamiento que se utilizan para calcular cada actualización de los pesos, mientras que el número de épocas (Epochs) se refiere a la cantidad de veces que el algoritmo pasa por todo el conjunto de datos de entrenamiento. En un escenario de experimentación teórica, podríamos suponer que usamos un tamaño de lote de 32, pues se trata de un valor bastante común en muchas aplicaciones, ya que representa un buen equilibrio entre la eficiencia computacional y la capacidad de generalización del modelo. También podríamos suponer que entrenamos el modelo durante 100 épocas, un valor bastante típico que se utiliza a menudo en la literatura y en la práctica [10].

Al finalizar el entrenamiento, la red aprende a mapear las frecuencias de aparición de las palabras en las noticias de un día, a una predicción del cambio en el precio de las acciones de ese mismo día. En la práctica, esto significa que se podría alimentar el modelo con las características correspondientes a un día específico y obtener una predicción de cómo cambiará el precio de las acciones.

Para finalizar la *Sección 3.1*, se presenta una tabla comparativa, *Tabla 3.1*, que resume las principales diferencias y similitudes entre los dos enfoques teóricos que se han desarrollado alrededor del modelo de estudio:

Características	***Regresión Lineal***	***Redes Neuronales***
Formulación del modelo	$\Delta P = \beta_0 + \beta_1\alpha_1 + \cdots + \beta_n\alpha_n$	Varía dependiendo de la arquitectura de la red
Función de coste	MSE	MSE
Método de aprendizaje	OLS. Si multicolinealidad→ Regresión de Crestas/ LASSO	SGD. Más precisión→ Adam
Entrenamiento	Optimización directa de la función de coste	Optimización iterativa
Hiperparámetros	λ (para Regresión de Crestas y LASSO)	Número de capas, neuronas por capa, tasa de aprendizaje, regularización (L1 y L2/ Dropout), tamaño del lote, número de épocas
Función de Activación	No aplica	ReLU en capas ocultas, ninguna en capa de salida
Validación	Validación cruzada	Validación cruzada
Manejo de variables	Debe seleccionarse manualmente (ingreso directo)	Puede aprender interacciones automáticamente
Flexibilidad	Modelo lineal	Modelo no lineal, permite modelar relaciones complejas

Tabla 3.1. Comparativa entre Regresión Lineal y Redes Neuronales (elaboración propia, 2023)

3.2 HIPÓTESIS DE TRABAJO

La hipótesis de trabajo que rige este estudio sostiene que la narrativa periodística relevante incide de manera sustancial en las fluctuaciones de los precios de las acciones. En un sentido más específico, la hipótesis alternativa (H1) plantea que determinados sucesos, temáticas o sentimientos reflejados en las noticias tienen el potencial de predecir las dinámicas futuras del mercado bursátil. Por consiguiente, la hipótesis nula (H0) mantiene que no existe una relación significativa entre las características derivadas de las noticias y las variaciones en los precios de las acciones.

Esta investigación parte del reconocimiento de que el mercado de valores es un sistema intrincado, altamente sensible a factores externos. En este contexto, se postula que el análisis de la narrativa periodística puede ofrecer una mirada alternativa y valiosa para comprender y

prever las tendencias de mercado. Así, se presupone que las noticias de relevancia, en particular aquellas vinculadas con la industria energética y las empresas específicas de interés en este estudio, contienen información de valor que puede ser capitalizada para anticipar las fluctuaciones de precios de las compañías de este sector.

La hipótesis se establece en términos concretos, sugiriendo que la frecuencia de aparición de palabras clave y los sentimientos manifestados en las noticias presentan correlación con las variaciones de los precios de las acciones. Se espera encontrar que ciertas palabras o tópicos, tales como "crisis", "innovación" o "acuerdos comerciales", estén asociados a tendencias de mercado específicas. Adicionalmente, se postula que el análisis de la narrativa periodística puede brindar una ventaja predictiva respecto a los enfoques tradicionales basados en análisis técnicos o fundamentales. Se espera que la incorporación de información textual en el modelo predictivo optimice su capacidad para captar señales sutiles y tendencias emergentes, que resulten difíciles de detectar a través de métodos puramente numéricos, tradicionales.

Tanto la hipótesis alternativa (H1) como la hipótesis nula (H0) pueden ser formuladas en términos de un modelo de regresión lineal. En un sentido general, un modelo de regresión lineal puede ser expresado como:

$$\Delta P = \alpha\beta + \varepsilon \tag{3.10}$$

tal y como se vio en la *Sección 3.1.1*, donde la variable dependiente coincide con la variación del precio de la acción, la matriz de variables independientes α coincide con las frecuencias de las palabras clave, β es el vector de coeficientes que se estiman a partir de los datos, y ε es el vector de términos de error.

Entonces, H0 se puede expresar como la afirmación de que todos los coeficientes en β son cero (es decir, las variables independientes no tienen ningún efecto sobre la variable dependiente), mientras que H1 sería la afirmación de que al menos uno de los coeficientes en β es distinto de cero (es decir, al menos una de las palabras clave tiene un efecto significativo sobre los precios de las acciones).

En un sentido más específico, si denotamos por α_i la serie temporal de frecuencias de aparición de la palabra clave k_i y por β_i el coeficiente correspondiente, entonces la contribución de la palabra clave k_i al modelo de regresión lineal puede ser expresada como $\alpha_i\beta_i$. De acuerdo con

H1, se espera que al menos una de estas contribuciones sea significativa, es decir, que la frecuencia de aparición de al menos una palabra clave pueda predecir de manera efectiva las fluctuaciones en los precios de las acciones.

Por lo tanto, el análisis de los datos recogidos se llevará a cabo mediante la técnica estadística de la regresión lineal múltiple, y los resultados serán evaluados a través de las pruebas de significancia de los coeficientes, como la prueba t de Student. Si al menos uno de los coeficientes β_i es significativamente distinto de cero, entonces se puede rechazar H0 y aceptar H1.

Para validar estas hipótesis, se ha llevado a cabo un riguroso trabajo de campo que incluye la recopilación de datos históricos de precios de acciones, la extracción y procesamiento de titulares y subtitulares de noticias relevantes, y la construcción de un modelo predictivo basado en regresiones lineales. Mediante el análisis de las características extraídas de las noticias y su correlación con las variaciones de precios, se pretende demostrar la viabilidad y eficacia del enfoque propuesto en este estudio.

El trabajo de campo no solo busca validar las hipótesis anteriormente mencionadas, sino también demostrar la capacidad de los modelos basados en aprendizaje automático para el análisis y la predicción de mercados financieros. Se postula que este enfoque permite descubrir patrones sutiles y dinámicos en los datos que son difíciles de detectar mediante métodos tradicionales. A través del presente estudio, se busca proporcionar una nueva perspectiva sobre la predicción de precios de acciones y contribuir a la comprensión de la complejidad del mercado de valores.

3.3 IMPLEMENTACIÓN DEL MODELO

La implementación del modelo es un proceso metódico que se extiende a través de varias fases, cada una con su propio conjunto de desafíos y oportunidades. La metodología utilizada en este estudio se basa en una serie de transformaciones y operaciones en el corpus de datos, que se representan esquemáticamente en la *Figura 3.5*. Aunque en este capítulo se ofrece una descripción teórica y más superficial de la metodología empleada, es importante destacar que el *Capítulo 4* se dedicará a explorar el enfoque de trabajo de campo y la aplicación práctica de

estas técnicas, permitiendo una inmersión más profunda en los aspectos prácticos de la implementación del modelo.

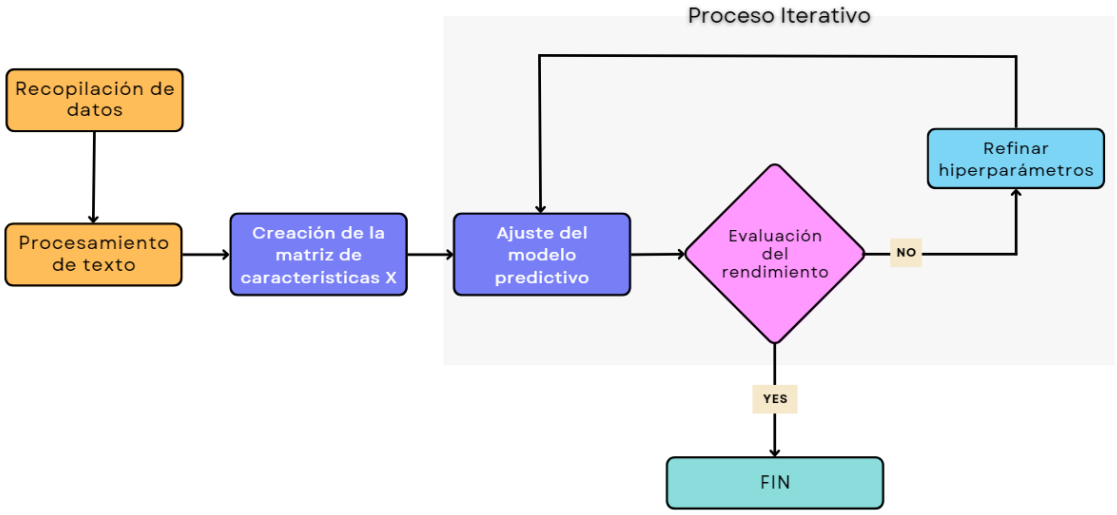

Figura 3.5. Diagrama de Flujo de la Implementación del Modelo (elaboración propia, 2023)

Comenzamos con la recopilación de datos, lo cual constituye la piedra angular de cualquier proyecto de análisis de datos. Para este estudio, los dos principales tipos de datos recogidos fueron los precios históricos de las acciones y la narrativa periodística, en concreto, titulares y subtitulares de las noticias. Estos conjuntos de datos son representados respectivamente por $P = \{p_1, p_2, ..., p_s\}$, la serie temporal de los precios de las acciones, y $T = \{t_1, t_2, ..., t_n\}$, la colección de noticias relevantes.

A continuación, procedemos a la etapa de procesamiento de texto. Se trata de una fase fundamental, ya que permite extraer características significativas de las noticias, que se pueden utilizar para el análisis posterior. Concretamente, las noticias en T son procesadas para identificar un conjunto de palabras relevantes, llamadas palabras clave $K = \{k_1, k_2, ..., k_m\}$, así como para determinar sus correspondientes frecuencias de aparición $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_m\}$. Esta operación se puede conceptualizar como la aplicación de una función $\varphi: T \rightarrow K \times \alpha$ que mapea cada noticia a un conjunto ordenado de palabras clave y sus frecuencias.

El siguiente paso es la creación de la matriz de características X. En esta etapa, se combinan las series temporales de los precios de las acciones P y las frecuencias de las palabras clave α para crear una matriz que encapsula toda la información relevante para el análisis. Cada fila de X corresponde a un instante temporal, en concreto a un día, mientras que cada columna representa una característica distinta: la primera columna el precio de la acción, el resto de las

columnas, desde la 2 hasta la m+1, las frecuencias de las palabras clave que aparecen en las noticias de ese día.

$$X_t = \begin{pmatrix} p_1 & \alpha_{1,1} & \cdots & \alpha_{m,1} \\ \vdots & \vdots & \vdots & \vdots \\ p_s & \alpha_{1,s} & \cdots & \alpha_{m,s} \end{pmatrix} \tag{3.11}$$

Con la matriz de características X preparada, el siguiente paso es ajustar nuestro modelo predictivo. Si elegimos un enfoque de modelo de regresión lineal, buscamos un vector de coeficientes β tal que Xβ es una buena aproximación para P. Esto se puede representar como una función ψ: X → β que mapea las características a los coeficientes. En el caso de la red neuronal, buscamos una matriz de pesos W que minimiza la discrepancia entre las salidas del modelo y los precios de las acciones. Esto se puede representar como una función θ: X → W que mapea las características a los pesos.

Una vez ajustado el modelo, el último paso en la implementación es evaluar su rendimiento. Para ello, calculamos una serie de métricas, como el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE)[15], y el coeficiente de determinación (R^2). Este análisis se realiza tanto en el conjunto de entrenamiento como en el conjunto de validación, lo que nos permite evaluar no solo en qué medida el modelo se ajusta a los datos, sino también su capacidad para generalizar a nuevos datos.

La implementación del modelo es inherentemente iterativa. El modelo se ajusta y evalúa repetidamente, refinando los hiperparámetros en cada iteración para optimizar su rendimiento. A través de este enfoque metódico y riguroso, se busca construir un modelo robusto y preciso que pueda predecir eficazmente las fluctuaciones en los precios de las acciones a partir de la narrativa periodística.

En este tercer capítulo, se ha establecido un sólido marco teórico para el modelo predictivo, explorando desde una sencilla regresión lineal hasta un enfoque más complejo utilizando redes neuronales. La formulación matemática avanzada ha desempeñado un papel primordial en la

[15] Del inglés, Root Mean Squared Error

presentación de un modelo coherente y robusto, así como en la formulación de las hipótesis de trabajo.

El proceso detallado de implementación del modelo, aunque teórico en esta etapa, ha delineado el camino para su aplicación práctica. Este recorrido, que comienza con la recopilación de datos y concluye con la evaluación del modelo, será explorado más a fondo y en un escenario práctico en el próximo capítulo.

Por lo tanto, el *Capítulo 3* ha sentado las bases teóricas y ha trazado la hoja de ruta para la aplicación práctica del modelo en el *Capítulo 4*. Este será el escenario donde podremos probar las hipótesis y validar la eficacia del modelo en la predicción de los precios de las acciones, enfrentando los desafíos prácticos que puedan surgir.

CAPÍTULO 4. APLICACIÓN PRÁCTICA DEL MODELO

Después de la construcción teórica del modelo predictivo en el *Capítulo 3*, nos adentramos ahora en el mundo práctico del aprendizaje automático y la analítica de datos, plasmando de forma experimental las ideas y métodos que se han discutido hasta ahora. Nos encontramos en el punto del proyecto donde la teoría se encuentra con la práctica.

La intención es demostrar la funcionalidad del modelo, su utilidad en escenarios reales y su capacidad para ofrecer predicciones significativas. En la *Sección 4.1* se desglosa el proceso práctico paso a paso para el desarrollo del modelo predictivo, abordando desde la recopilación de datos hasta el entrenamiento. Posteriormente, en la *Sección 4.2*, el modelo se pone a prueba en un escenario específico para evaluar su desempeño.

Este capítulo está diseñado para proporcionar un panorama completo del viaje desde el desarrollo teórico inicial del modelo hasta su implementación y aplicación práctica, ayudando a la comprensión de cómo un modelo teórico toma forma en la realidad.

4.1 TRABAJO DE CAMPO

El trabajo de campo realizado para este estudio se llevó a cabo en múltiples etapas meticulosamente planificadas y ejecutadas, cuyo objetivo común y último era construir un modelo predictivo sólido y fiable que pudiera desentrañar y aprovechar la relación intrínseca entre la narrativa de las noticias y las variaciones de precios de las acciones. El trabajo implicó tanto la recopilación y procesamiento de datos como la construcción y ajuste de un modelo de aprendizaje supervisado. A continuación, el diagrama de flujo de la *Figura 4.1* detalla las etapas principales, así como los programas o páginas web utilizados en cada una de ellas:

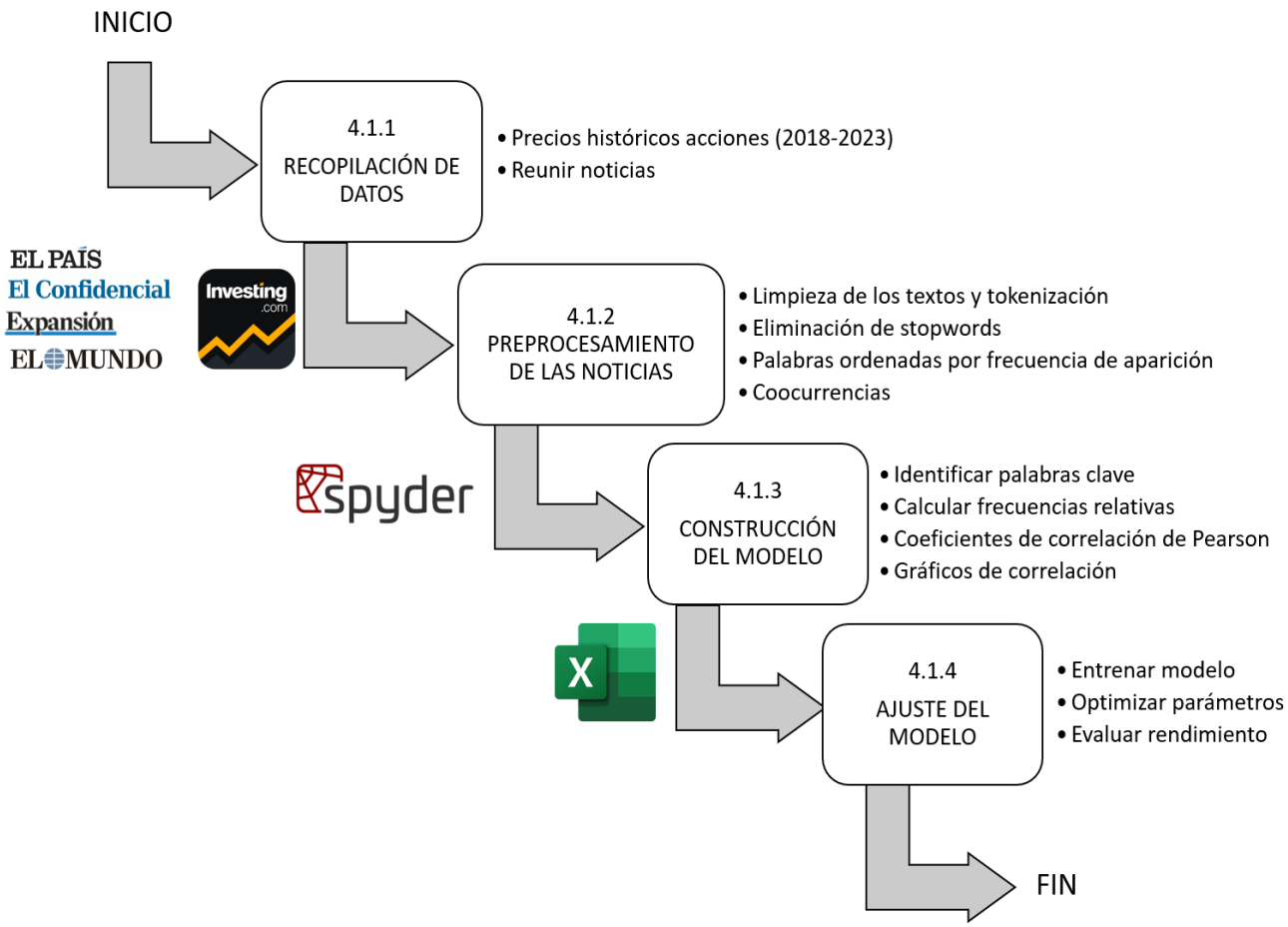

Figura 4.6. Diagrama de Flujo del Trabajo de Campo (elaboración propia, 2023)

1.1 RECOPILACIÓN DE DATOS

La recopilación de datos es la primera etapa llevada a cabo en este proceso. El conjunto de datos utilizado en el proyecto consta de dos partes principales: los precios históricos de las acciones de cada uno de los índices estudiados, y los titulares y subtitulares de las noticias diarias publicadas en los principales periódicos españoles.

4.1.1.1 RECOPILACIÓN DE PRECIOS HISTÓRICOS DE ACCIONES

Son tres los índices elegidos para proporcionar la base del análisis de las variaciones de precios: IBEX35, IBEX35-Energy y la empresa energética Endesa (ELE). La elección de recopilar los datos históricos de estos tres índices se fundamenta en la intención de aplicar y comparar el modelo en diferentes niveles de generalidad y especificidad.

El IBEX35, como índice bursátil más amplio y generalizado en España, permite evaluar y comprender las fluctuaciones generales del mercado. Al incluirlo en el análisis, se obtiene una visión panorámica de la economía y se pueden identificar patrones y tendencias más amplias. Por otro lado, el IBEX35-Energy se selecciona para abordar el sector energético en su conjunto, pues este se enfoca específicamente en las empresas del sector energético que cotizan en la bolsa española. La consideración de este índice permite capturar las dinámicas particulares y las influencias específicas relacionadas con el sector energético en el mercado accionario. Finalmente, la elección de incluir a la empresa energética Endesa permite explorar la variación de precios a nivel más granular y específico. De esta forma, se pueden examinar factores y eventos concretos que afectan a esa compañía en particular, como cambios regulatorios, anuncios de resultados financieros u otros eventos que aparentemente no guardan relación con el sector.

El siguiente Diagrama de Venn apilado ilustra la relación entre los tres niveles de análisis y cómo cada nivel de especificidad se encuentra contenido dentro del siguiente nivel más general.

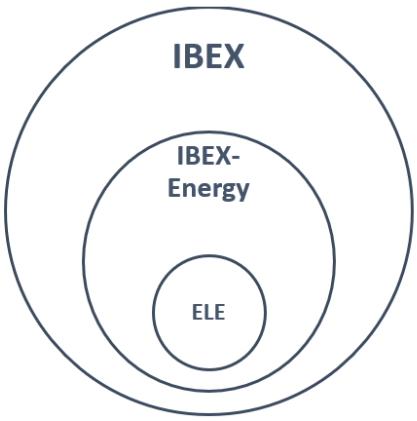

Figura 4.7. Diagrama de Venn apilado de los índices bursátiles modelados

En primer lugar, se recogieron datos de precios diarios de las acciones de los tres índices estudiados, de una fuente financiera pública confiable, Investing.com. Se trata de una plataforma popular dedicada a proporcionar información financiera, noticias y análisis del mercado global, y es ampliamente utilizada por inversores y traders para mantenerse al tanto de los mercados financieros, incluyendo acciones, divisas, materias primas, bonos y criptomonedas.

Los datos se recopilaron mediante la opción que ofrece Investing.com de datos históricos con una periodicidad diaria. Se descargaron tres archivos Excel, uno para cada índice (IBEX35,

IBEX-Energy y Endesa). Cada archivo presenta los datos organizados en las siguientes columnas: 'fecha', 'precio de cierre', 'precio de apertura', 'precio máximo', 'precio mínimo', 'volumen' y '% de variación'. Como aclaración, la columna de volumen registra el total de acciones negociadas durante el día, y la de la variación expresa el cambio porcentual en el precio de la acción desde el cierre de la jornada anterior.

Los datos descargados vienen por defecto en formato Comma Separated Values (CSV), lo cual implica que los campos de datos están separados por comas. Sin embargo, se identificaron algunas inconsistencias y errores de formato que requerían una limpieza de datos. Este proceso de limpieza implicó la conversión de datos a un formato más manejable y la eliminación de datos duplicados o irrelevantes.

En segundo lugar, con los datos en un formato adecuado, se procedió a identificar las 10 mayores subidas y bajadas de cada uno de los índices[16]. El objetivo de este análisis era destacar aquellos días críticos de mayor volatilidad en los precios, para así analizar las noticias publicadas en estas fechas, y finalmente poder comprender la influencia de la narrativa de las noticias en estas oscilaciones notables. Además de resaltar estas 20 variaciones más extremas en los precios, se asoció también la fecha correspondiente a cada una de estas fluctuaciones.[17]

Cabe mencionar que los datos se recolectaron en una serie temporal que abarca un período específico de 5 años, desde el 1 de enero de 2018 hasta el 31 de diciembre de 2022. La decisión de fijar un periodo de cinco años de recopilación de datos se basa en varias consideraciones, como la representatividad del periodo, la consideración de ciclos económicos y la consistencia con estudios anteriores. Se indican estas consideraciones en la *Figura 4.3*:

[16] Para encontrar las 10 mayores variaciones positivas y las 10 mayores variaciones negativas de cada índice, se recurrió a las funciones de Excel K.ESIMO.MAYOR y K.ESIMO.MENOR en la columna '% de Variación'. Estas funciones devuelven el k-ésimo valor más grande/pequeño de un conjunto de datos. Al aplicar estas funciones con k valores del 1 al 10, se lograron identificar las 10 subidas y bajadas más significativas.

[17] Se usó la función XLOOKUP de Excel. Esta función busca un valor en la primera columna de un rango de celdas (la variación anteriormente encontrada) y devuelve el valor en la misma fila de una columna diferente (la fecha correspondiente)

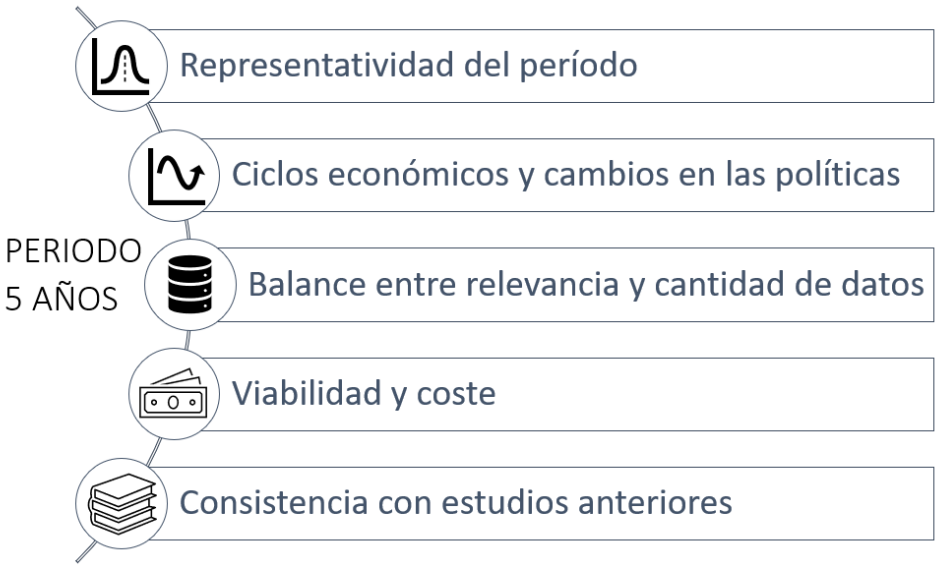

Figura 4.8. Motivos de Elección del Periodo Temporal de Datos (elaboración propia, 2023)

Un periodo de cinco años ofrece una muestra significativa y representativa de datos históricos, lo que permite capturar diversas condiciones del mercado a lo largo del tiempo. Entre estas condiciones podemos encontrar períodos de alta volatilidad, estabilidad, tendencias alcistas o bajistas, así como eventos económicos, políticos o sociales que pueden haber influido en los precios de las acciones. Por ejemplo, el presente estudio se caracteriza por cubrir un periodo de tiempo claramente marcado por dos importantes sucesos de gran impacto a nivel mundial: la pandemia del coronavirus y la guerra entre Rusia y Ucrania.

Del mismo modo, al incluir datos históricos de varios años, se logra un equilibrio entre el análisis del pasado y la consideración del entorno actual del mercado. En los mercados financieros, los datos más recientes suelen ser los más relevantes para predecir el comportamiento futuro y, por lo tanto, un periodo de cinco años puede ser un buen compromiso para obtener suficientes datos sin ir tan atrás en el tiempo hasta incluir datos pasados no relevantes.

Además, los mercados financieros suelen estar influenciados por ciclos económicos, que pueden tener una duración de varios años. Al utilizar un periodo de cinco años, se tiene en cuenta la posibilidad de capturar diferentes fases de un ciclo económico, lo que puede aportar información relevante para la construcción del modelo predictivo.

Finalmente, la elección de un periodo de cinco años también está respaldada por la consistencia con investigaciones previas que han utilizado periodos similares para analizar el

comportamiento de los precios de las acciones [41][43]. Esto facilita la comparación de resultados y la evaluación de la efectividad del modelo propuesto en relación con estudios anteriores.

4.1.1.2 RECOPILACIÓN DE NOTICIAS

Una vez determinadas las fechas correspondientes a las 10 subidas y 10 bajadas más pronunciadas en la variación del precio de cada uno de los tres índices, se procedió a la recolección de noticias correspondientes a esos días. Se recopilaron alrededor de 100 pares de titulares y subtitulares[18] por cada uno de estos días, con el objetivo de capturar la diversidad y amplitud de la cobertura mediática en momentos clave de variación en los mercados. Para ello, se hizo uso de las hemerotecas de diferentes medios de comunicación en línea de reconocido prestigio nacional, como El Mundo, El País, Expansión y El Confidencial.

A continuación, en la *Figura 4.4* se muestra un ejemplo de noticia publicada en el diario digital de El País el 10 de abril de 2020, que ilustra el concepto de título, subtítulo y cuerpo (este último no incluido en los datos de entrada al modelo):

Figura 4.9. Partes de una Noticia (elaboración propia, 2023)

[18] Texto ubicado justo debajo del título que sirve de resumen breve de lo sucedido. También llamados 'copete' o 'bajada'

El criterio adoptado para la selección de las noticias se basó en la recogida exclusiva de los títulos y subtítulos de las noticias publicadas en portada durante las horas matutinas y del mediodía de cada día crítico.

La decisión de recoger solo los titulares y subtitulares, y no el contenido completo de las noticias, se fundamentó en varios factores. Por un lado, lo que primero ven los lectores son los titulares y subtitulares, que a menudo encapsulan los aspectos más relevantes de la noticia. Segundo, estos fragmentos de texto también son más fáciles de analizar desde un punto de vista computacional, y permiten un procesamiento más eficiente de un mayor volumen de datos. Finalmente, este enfoque se alinea con la realidad de que muchos inversores no leen necesariamente la totalidad de cada artículo, pero se informan y toman decisiones a partir del título. Por lo tanto, la concentración de las noticias en estos dos elementos proporciona un reflejo más realista de cómo las noticias pueden influir en los movimientos del mercado.

De igual modo, solo se seleccionaron las noticias que aparecían en la portada de los sitios web de los medios de comunicación o en secciones fácilmente accesibles, pues suelen ser las más relevantes o de mayor impacto para la audiencia general. Como el comportamiento del lector en línea tiende a centrarse en la navegación superficial, estas noticias de fácil acceso son más propensas a ser leídas y, por lo tanto, a influir en las decisiones de inversión.

Además, se decidió excluir las noticias publicadas durante la tarde y noche, con la intención de concentrarse en la información que los inversionistas habrían tenido a su disposición durante las horas de operación de los mercados.

Por último, cabe mencionar que las noticias recopiladas no estaban explícitamente relacionadas con las empresas de interés y la industria energética, sino que más bien se recogieron noticias de carácter global. Se trata de una decisión consciente que tiene su origen en la idea de capturar la percepción general del inversor y el sentimiento del mercado. Esta elección se basa en la premisa de que los mercados financieros no son simplemente una representación de los fundamentos y actividades de las empresas individuales, sino que están influenciados por un conjunto más amplio de factores socioeconómicos y geopolíticos [5]. Por ello, los inversores no solo operan en base a las noticias específicas de las empresas en las que invierten, sino que también tienen en cuenta el clima general de los medios y los acontecimientos mundiales. Por ejemplo, una noticia de carácter general sobre la inestabilidad política en España, el cambio climático, o las tensiones comerciales internacionales, puede tener un impacto significativo en

la percepción del riesgo y, en última instancia, en las decisiones de inversión en cualquier empresa.

Al recoger y analizar noticias de carácter general, este estudio busca entender y cuantificar el impacto de estos factores aparentemente no relacionados en los precios de las acciones. Esto también permite que el modelo sea más robusto y generalizable, ya que no está limitado a un conjunto específico de noticias o empresas. Sin duda, este enfoque va a ayudar a obtener una visión más holística y matizada de las fuerzas que impulsan los movimientos del mercado.

1.2 PREPROCESAMIENTO DE LOS DATOS

Después de la recopilación, las noticias pasaron por una etapa de preprocesamiento para preparar los datos para el análisis. Se utilizó el lenguaje de programación Python, con bibliotecas como *pandas* para la manipulación de datos, y *NLTK*[19] (Kit de herramientas de lenguaje natural) para el procesamiento del lenguaje natural. Se adjunta el código Python desarrollado en el *ANEXO I*. Es importante mencionar que, antes de ejecutar el programa, se debe cambiar el nombre del archivo Excel del cual se desean extraer las noticias, así como la fecha en formato DD.MM.AA para la cual queremos que el código extraiga las palabras más frecuentes. A continuación, en el diagrama de flujo de la *Figura 4.5* se ordenan las funciones llevadas a cabo a partir de este script:

[19] Del inglés, Natural Language Toolkit

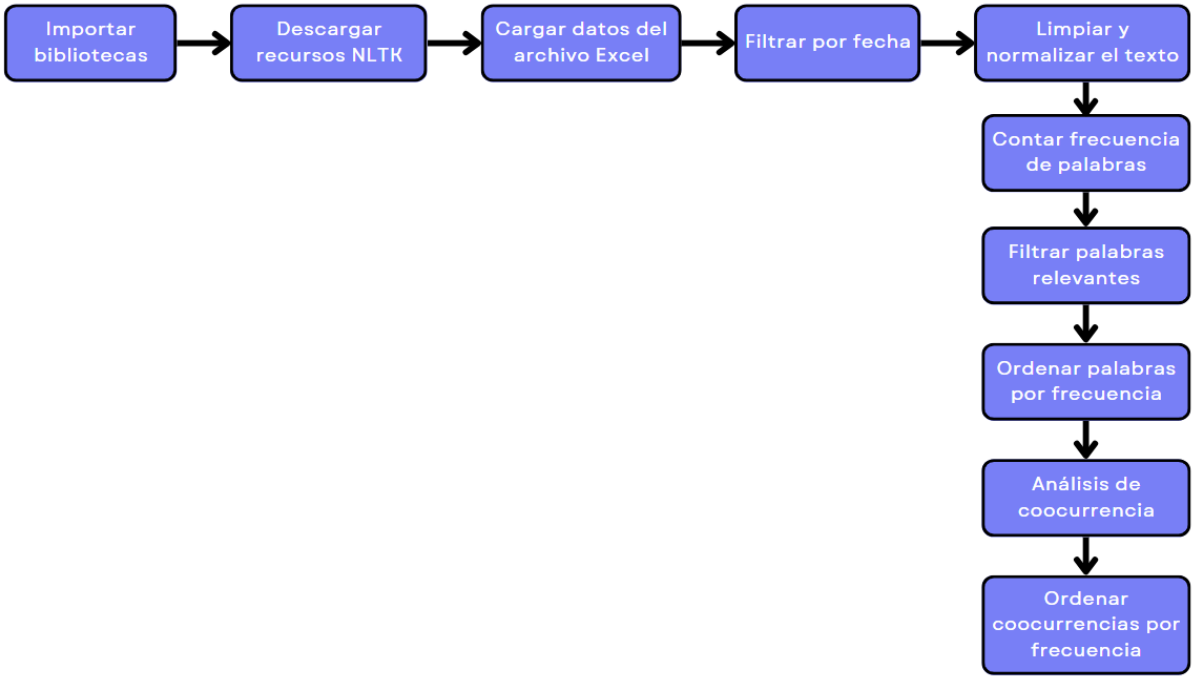

Figura 4.10. Diagrama de Flujo del Procesamiento de Noticias con Python (elaboración propia, 2023)

Tras importar las bibliotecas pandas, nltk, collections e itertools, se leyeron los datos de las noticias de un archivo Excel utilizando pandas, cargándolos en un DataFrame. Cada vez que se ejecuta el código, hay que introducir la fecha para filtrar el DataFrame con el objetivo de seleccionar solo las filas que corresponden a la fecha específica que se quiere analizar en ese momento.

En segundo lugar, se llevó a cabo una limpieza del texto, que implicó la eliminación de caracteres no alfanuméricos y la conversión de todas las letras a minúsculas.

El siguiente paso fue la eliminación de las palabras vacías. Las palabras vacías son palabras comunes en un lenguaje (como "el", "un", "y" en español) que no aportan mucha información para el análisis de texto. Para identificar y eliminar estas palabras, se utilizó la lista de palabras vacías de la biblioteca de *NLTK*, llamada *stopwords*.

A continuación, se llevó a cabo la tokenización, que es el proceso de dividir el texto en palabras individuales, o "tokens". Para esto, se utilizó la función *word_tokenize* de *NLTK*.

Por último, se realizó un conteo de la frecuencia de aparición de las palabras en las noticias de cada día. Este conteo se hizo incorporando el objeto *Counter* de la biblioteca *collections*, que proporciona una forma conveniente de contar la frecuencia de los elementos en una lista.

Mediante el conteo, se seleccionaron las palabras más relevantes (palabras clave), definidas como las palabras que aparecen más de 3 veces en las noticias de un día concreto. Finalmente, las palabras se ordenaron por su frecuencia de aparición.

Además de contar la frecuencia de las palabras, también se llevó a cabo un análisis de coocurrencia de palabras. Este análisis consiste en identificar y contar los pares o tríos de palabras que aparecen juntas en las mismas noticias. Los resultados de coocurrencia que se obtuvieron a partir del código no fueron utilizados para la construcción del modelo en sí, pues la entrada del modelo es la frecuencia de aparición de cada palabra de forma individual. Sin embargo, este último análisis ayudó a identificar las relaciones entre palabras relevantes que detonan una variación significante en los precios de las acciones.

1.3 CONSTRUCCIÓN DEL MODELO

Una vez preprocesados los datos, se procedió a la construcción del modelo. En este punto, se contaba con una serie de palabras clave con su frecuencia de aparición en cada uno de los días y la variación porcentual de precio asociada a esos días.

El primer paso en la construcción del modelo fue agrupar las palabras según su asociación con las subidas o bajadas de los precios de las acciones. Esto se realizó en función de las 10 mayores subidas y las 10 mayores bajadas del IBEX, IBEX-Energy y Endesa. El objetivo de este paso fue identificar las palabras que comúnmente aparecen en las noticias de los días con grandes variaciones de precios, distinguiendo entre las variaciones positivas y las negativas.

Para cada día seleccionado, se compilaron dos tablas de información crítica. En primer lugar, se generó un diccionario de aproximadamente 50 palabras clave, ordenadas en función de su frecuencia de aparición en las noticias. Este diccionario incluye también la fecha correspondiente y la variación del precio de la acción de la entidad en cuestión para ese día. En segundo lugar, se recopiló una tabla sobre las coocurrencias significativas entre las palabras clave más relevantes, con el objetivo de comprender el contexto en el cual estas sirven como indicadores predictivos de la subida o bajada del precio de la acción.

Para ilustrar esto, la *Tabla 4.1* y la *Tabla 4.2* muestran dos extractos del diccionario y de la tabla de coocurrencias, respectivamente, de un día que corresponde a una de las 10 mayores subidas del IBEX.

Fecha	Variación	Palabra	Frecuencia
09.11.2020	8,57%	biden	16
09.11.2020		vacuna	15
09.11.2020		coronavirus	12
09.11.2020		pfizer	10
09.11.2020		subida	10
09.11.2020		confinamiento	8
09.11.2020		eeuu	7
09.11.2020		presidente	6
09.11.2020		avances	6
09.11.2020		efectividad	5
09.11.2020		ue	5
09.11.2020		actividad	3
09.11.2020		acuerdo	3
09.11.2020		medida	3
09.11.2020		inversión	3

Tabla 4.2. Diccionario de palabras por orden de frecuencia. Subida Ibex (elaboración propia, 2023)

	Coocurrencia	Frecuencia
vacuna	coronavirus	10
ibex	biden	9
biden	coronavirus	6

Tabla 4.3: Coocurrencia de palabras clave, Subida Ibex (elaboración propia, 2023)

Empleando la misma estructura, se muestra la *Tabla 4.3* y la *Tabla 4.4* con el objetivo de representar los mismos resultados obtenidos con el código Python, pero esta vez un día en el que la variación del precio fue negativa. En ambas se observa una de las 10 mayores bajadas del precio de la acción de Endesa.

Fecha	Variación	Palabra	Frecuencia
01.03.2022	-7,48%	rusia	24
01.03.2022		ucrania	21
01.03.2022		guerra	13
01.03.2022		kiev	9
01.03.2022		sanciones	9
01.03.2022		moscú	6
01.03.2022		conflicto	6
01.03.2022		precios	6
01.03.2022		putin	6
01.03.2022		ofensiva	5
01.03.2022		crisis	4

01.03.2022	ataques	4
01.03.2022	inflación	4
01.03.2022	energía	4
01.03.2022	pérdidas	4
01.03.2022	petróleo	4

Tabla 4.4: Diccionario de palabras por orden de frecuencia. Bajada Endesa (elaboración propia, 2023)

	Coocurrencia		Frecuencia
sanciones	rusia		8
guerra	ucrania	rusia	6
rusia	kiev		5
precios	ucrania		3
precios	guerra		3

Tabla 4.5: Coocurrencia de palabras clave. Bajada Endesa (elaboración propia, 2023)

Tras obtener un diccionario de palabras para cada día, se procedió a analizar la relación entre la frecuencia de aparición de cada palabra y la variación del precio del índice o la acción mediante dos vías: una medida cuantitativa y una representación gráfica.

Para cuantificar la relación, para cada palabra se calculó el coeficiente de correlación de Pearson, estadístico que mide el grado de relación lineal entre dos variables [42].

El coeficiente de correlación de Pearson se calcula mediante la siguiente fórmula[20]:

$$r = \frac{\sum[(x_i - \bar{x})(y_i - \bar{y})]}{n\sigma_x\sigma_y} \tag{4.1}$$

Donde:

- x_i e y_i son las puntuaciones individuales en las dos variables (en este caso, frecuencia de aparición de una palabra y variación del precio de la acción)
- $\bar{x}$ e $\bar{y}$ son las medias de x e y respectivamente

[20] En la práctica se calculó con la función de Excel **COEF.DE.CORREL**, la cual realiza automáticamente los cálculos necesarios para determinar el coeficiente de correlación entre dos conjuntos de datos

- σx y σy son las desviaciones estándar de x e y respectivamente
- n es el número total de pares de puntuaciones

El resultado será un valor entre -1 y 1. Un valor de 1 significa una correlación positiva perfecta, un valor de -1 significa una correlación negativa perfecta, y un valor de 0 significa que no hay correlación entre las dos variables. En términos más concretos, si el coeficiente resulta en un valor cercano a 1 se podría inferir que, cuando la frecuencia de la palabra clave aumenta, también lo hace la variación del precio de la acción. Por otro lado, un coeficiente cercano a -1 indicaría que, a medida que la frecuencia de la palabra clave aumenta, la variación del precio de la acción tiende a disminuir.

En la *Sección 4.2*, estos coeficientes de correlación serán los "pesos" asignados a cada palabra a la hora de aplicar el modelo de predicción a un nuevo día, basándonos en su correlación histórica con la variación de los precios.

Adicionalmente al cálculo del coeficiente de correlación, para cada palabra se trazó un gráfico que representa simultáneamente dos series: por un lado, la variación del precio en función del tiempo (Serie ΔP-t), y por otro, la frecuencia relativa de aparición de la palabra en el mismo período de tiempo (Serie α-t). Este segundo enfoque de análisis permite visualizar la relación entre la frecuencia de aparición de una palabra y la variación del precio de manera más intuitiva y gráfica.

A continuación, se muestra un ejemplo en la *Tabla 4.5* de la forma en la que se presentaron los datos antes de realizar las gráficas con las que poder visualizar la correlación, en este caso para la palabra "medidas". En ella se incluyen las fechas en las que aparecía la palabra, así como el número de día al que correspondía cada una (teniendo en cuenta que se consideró el 1 de enero de 2018 como día 1) y la variación porcentual en el precio de la acción de ese día. Asimismo, la tabla contiene la frecuencia de aparición de la palabra, transformada a frecuencia relativa[21], y finalmente, el coeficiente de correlación de Pearson (*r*).

[21] Se usa la frecuencia relativa como forma de normalización. Una palabra que aparece 10 veces en un día con 100 palabras tiene más impacto y es probablemente más relevante que una palabra que aparece 10 veces en un día con 1000 palabras.

Fecha	Día	Variación	Palabra	Frecuencia	Frecuencia relativa	Coef. Correlación
13.03.2020	803	3,74%	MEDIDAS	21	0,33	-0,20
17.03.2020	807	6,41%		12	0,19	
24.03.2020	814	7,82%		7	0,11	
05.06.2020	887	4,04%		3	0,05	
02.07.2020	914	3,75%		4	0,06	
09.11.2020	1044	8,57%		3	0,05	
25.02.2022	1517	3,51%		5	0,08	
09.03.2022	1529	4,88%		8	0,13	

Tabla 4.6: Variación del precio y frecuencia relativa de aparición, "MEDIDAS" (elaboración propia, 2023)

Como ya se ha comentado, para cada palabra clave se creó un gráfico que representa la variación del precio de las acciones y la frecuencia relativa de aparición de la palabra (columnas marcadas en azul en la *Tabla 4.5*), ambas en función del tiempo (columna marcada en rojo). Estos gráficos fueron fundamentales para comprender y evaluar la utilidad de cada palabra clave en el modelo de predicción.

La superposición de las dos series de datos en un solo gráfico permite evaluar visualmente si hay una correlación aparente entre las dos. En teoría, si una palabra clave tiene un impacto significativo en las fluctuaciones del precio de las acciones, deberíamos ser capaces de observar una tendencia en el gráfico: los aumentos en la frecuencia de aparición de la palabra deberían coincidir con los movimientos notables en el precio de las acciones, ya sea hacia arriba o hacia abajo. Además, los gráficos también proporcionan un contexto valioso para interpretar el coeficiente de correlación de Pearson calculado para cada palabra. Aunque el coeficiente de correlación ofrece una medida cuantitativa de la relación entre las dos variables, los gráficos permiten una apreciación más intuitiva y visual de esta relación.

Es importante mencionar que para algunas palabras se observa una clara relación entre su frecuencia de aparición y las variaciones en el precio de las acciones, pero para otras esta relación es menos evidente. En las siguientes páginas, se presentan ejemplos de los gráficos obtenidos al construir el modelo. Solo se van a mostrar gráficos de algunas de las palabras cuya correlación es fuerte ($r > |0,5|$). Para conocer el resto de las palabras que conforman el modelo predictivo, consultar los archivos Excel adjuntos al *ANEXO II*.

4.1.1.3 MODELO PREDICTIVO IBEX 35

Teniendo en cuenta que el color verde para la serie de variación de precios indica que la palabra aparece en subidas, el color rojo indica que aparece en bajadas, y el color amarillo que aparece en ambas, a continuación, se muestran algunas de las palabras que forman parte del modelo predictivo del Ibex 35.

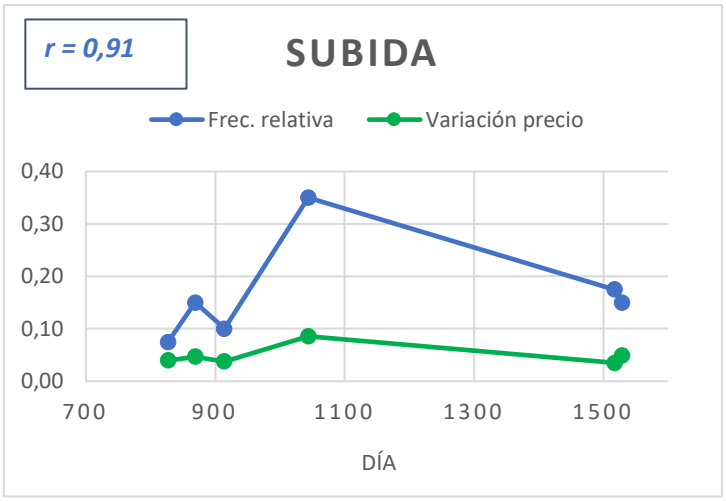

Figura 4.11. Gráfico frecuencia "Subida" y variación del Ibex (elaboración propia, 2023)

La primera palabra analizada es "subida", de la cual observamos que aparece en seis días diferentes en los que el Ibex experimentó aumentos de precio notables. Además, el coeficiente de correlación de Pearson calculado es 0,91, lo cual significa que hay una relación significativa entre las dos variables objeto de estudio, y que tienden a moverse juntas en la misma dirección. Cuando la frecuencia de aparición de esta palabra es alta, también lo es la variación del precio del índice. En este caso, el hecho de que la palabra "subida" tenga una fuerte correlación positiva con la variación del precio del índice parece lógico. Sin dejar de tener en cuenta que existen otras variables subyacentes, se podría esperar que una mayor utilización de la palabra "subida" en los titulares cause un aumento en el precio del principal índice bursátil de España.

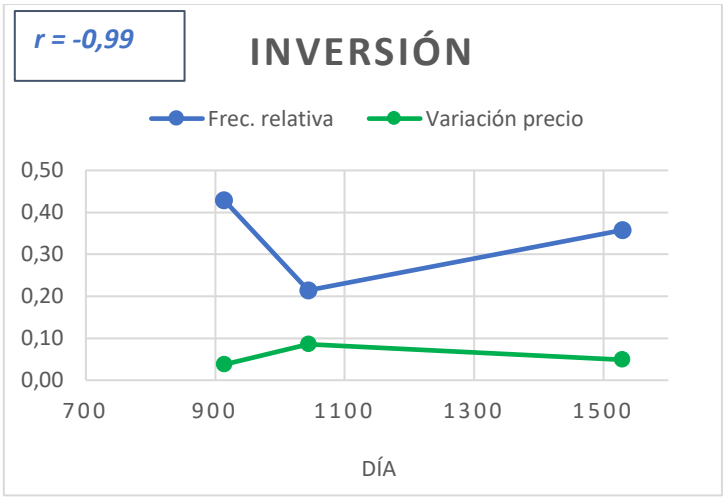

Figura 4.12. Gráfico frecuencia "Inversión" y variación del Ibex (elaboración propia, 2023)

Para la palabra "inversión", el coeficiente de correlación de Pearson calculado es -0,99, que indica una correlación negativa muy fuerte entre la frecuencia de aparición de la palabra y la variación del precio del Ibex. Al observar la representación visual de la *Figura 4.7*, notamos que a medida que aumenta la frecuencia relativa de la palabra, la variación del precio tiende a disminuir, lo cual puede parecer paradójico, pues la palabra solo aparece en días de subida. Sin embargo, una interpretación posible es que la palabra "inversión" podría utilizarse de manera más prominente en días en los que el mercado anticipa una subida del precio. En otras palabras, en días de expectativas altas y optimismo de inversión, es probable que los medios de comunicación utilicen la palabra con mayor frecuencia. Sin embargo, el mercado puede ya haber considerado estas expectativas y el aumento resultante del precio del índice puede ser menor de lo que podría haber sido en ausencia de esta anticipación. Esto resultaría en una correlación negativa entre la frecuencia relativa de la palabra "inversión" y la variación del precio.

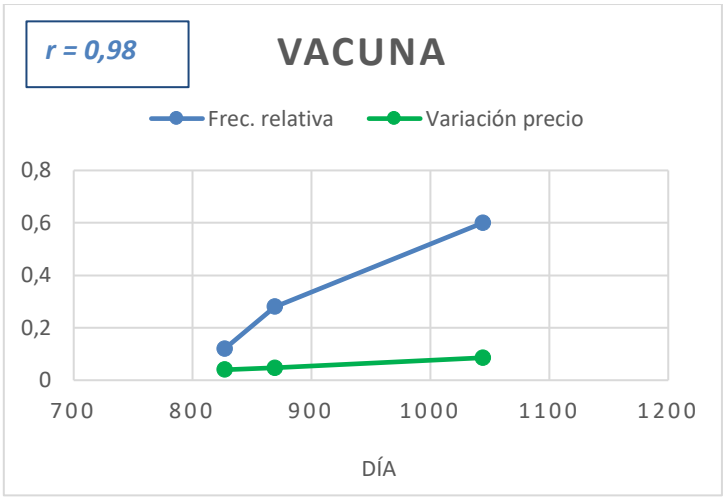

Figura 4.13. Gráfico frecuencia "Vacuna" y variación del Ibex (elaboración propia, 2023)

Otra palabra clave para el modelo predictivo del Ibex es "vacuna", cuya aparición en los titulares y subtitulares se produce en momentos de importantes subidas del índice, lo cual sugiere una relación entre la percepción positiva de la vacuna contra el COVID-19 y el incremento en el valor del índice. Las expectativas del mercado de una mejora en la economía gracias al avance de la vacunación y a la posibilidad de una recuperación más rápida de la pandemia se ven reflejadas en el aumento notable de la frecuencia relativa de la palabra “vacuna” a lo largo del tiempo.

Los días en los que aparece esta palabra están relativamente concentrados en un periodo específico, desde el 6 de abril de 2020 hasta el 9 de noviembre del mismo año, específicamente vinculado al contexto de la pandemia del coronavirus. Esto sugiere que hay un efecto de concentración temporal y, por lo tanto, la función predictiva de la palabra “vacuna” podría no ser completamente generalizable a otros periodos de tiempo. En circunstancias normales, la aparición de la palabra "vacuna" en las noticias podría no tener un impacto tan significativo en el índice.

Sin embargo, hay una lección general que se puede extraer de esto: durante periodos de crisis sanitaria a nivel mundial, noticias positivas como avances significativos en la medicina pueden tener un impacto positivo en los mercados financieros. Este tipo de noticias puede alentar el optimismo entre los inversores sobre la recuperación económica, lo que puede llevar a un aumento en los índices bursátiles.

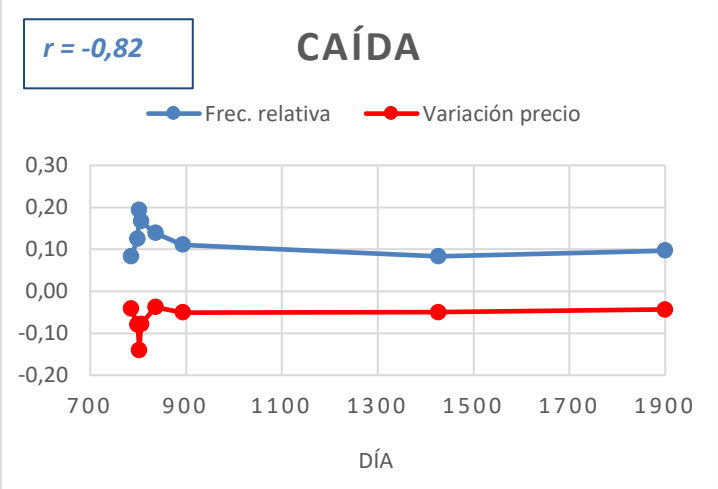

Figura 4.14. Gráfico frecuencia "Caída" y variación del Ibex (elaboración propia, 2023)

Al contrario de lo que ocurría para la palabra “subida”, la palabra "caída" aparece exclusivamente en días en los que el índice experimentó descensos notables. El gráfico de la *Figura 4.9* revela una correlación notablemente alta y negativa con la variación del precio del Ibex, lo cual es coherente con lo que se podría esperar intuitivamente.

El aumento en la frecuencia de aparición de la palabra puede que refleje una mayor atención mediática o preocupación en el mercado, lo cual augura caídas más agudas. Además, cabe mencionar que la dispersión temporal de esta palabra es amplia, lo que sugiere que su presencia en las noticias no está limitada a periodos específicos de tiempo. Por lo tanto, en principio se trata de una palabra capaz de predecir la variación del Ibex.

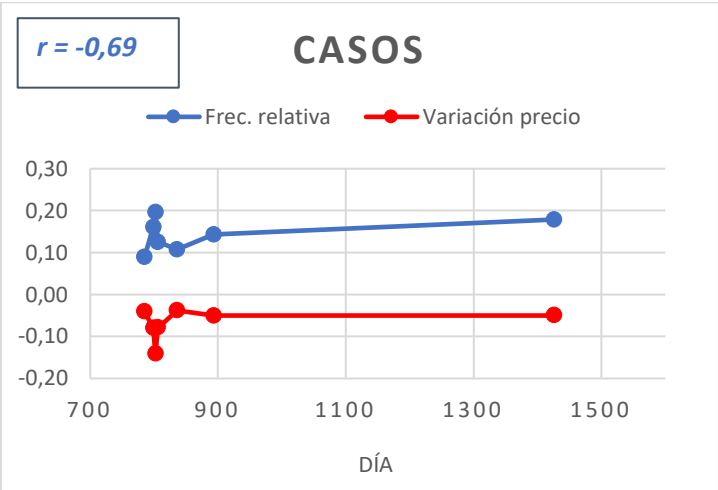

Figura 4.15. Gráfico frecuencia "Casos" y variación del Ibex (elaboración propia, 2023)

La correlación negativa de -0,69 entre la aparición de la palabra "casos" y la variación del precio del índice IBEX sugiere una relación inversa moderada: a medida que aumenta la frecuencia relativa de la palabra, el precio del índice tiende a caer.

En el contexto de los últimos cinco años (2018-2023), el uso frecuente de la palabra "casos" en las noticias probablemente esté asociado con la pandemia del coronavirus. Durante este período, se ha observado que los aumentos en los casos de COVID-19 han causado incertidumbre y volatilidad en los mercados financieros, debido a las preocupaciones sobre las posibles consecuencias económicas de las restricciones y cierres, así como al impacto en la salud y la productividad de la fuerza laboral.

Se puede observar que la palabra analizada aparece consistentemente en días de bajadas significativas del índice, lo que es coherente con la idea de que el aumento de casos de COVID-19 (o de cualquier otra enfermedad importante) puede contribuir a la inestabilidad del mercado nacional. En términos de la dispersión temporal de sus apariciones, parece que estas están relativamente concentradas, reflejando períodos de tiempo durante los cuales hubo brotes significativos de COVID-19. En definitiva, podemos decir que la aparición de "casos" está correlacionada con bajadas del índice porque la palabra "casos" está correlacionada con los brotes de coronavirus.

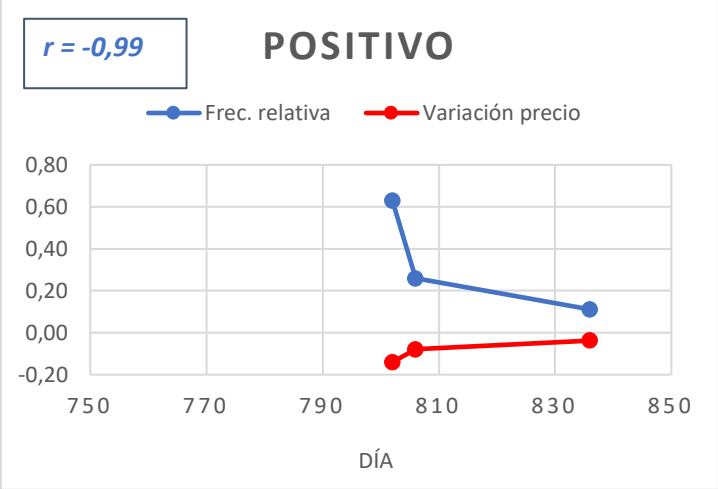

Figura 4.16. Gráfico frecuencia "Positivo" y variación del Ibex (elaboración propia, 2023)

Por último, y en línea con la palabra “casos”, encontramos en el diccionario de palabras clave el término “positivo”. Aunque generalmente se podría esperar que la palabra "positivo" esté asociada con noticias optimistas y, por ende, con subidas en el índice, en el contexto de la pandemia de coronavirus su significado cambia drásticamente. En este escenario, "positivo" se

usa comúnmente para referirse a nuevos casos confirmados de la enfermedad, una noticia desalentadora que puede generar incertidumbre en los mercados financieros.

Por lo tanto, el gráfico muestra una correlación negativa muy fuerte (-0,99) entre la frecuencia de aparición de la palabra "positivo" y la variación del precio del índice. La frecuencia de aparición de la palabra disminuye a lo largo del tiempo, patrón que podría reflejar una disminución en la intensidad de las noticias negativas relacionadas con la pandemia o quizás una disminución en el número de nuevos casos confirmados, que podría haber tenido un efecto en la mejora de la confianza del mercado y, consecuentemente, en una disminución menos pronunciada del índice.

Sin embargo, la aparición de la palabra “positivo” podría no ser completamente generalizable a otros periodos de tiempo, ya que se trata de un término con una connotación contraria en circunstancias normales. De este análisis se desprende una importante lección sobre la naturaleza contextual de las palabras en la predicción de las tendencias del mercado. La misma palabra puede tener diferentes implicaciones en diferentes contextos, y es crucial tener en cuenta estos matices al interpretar los resultados.

4.1.1.4 MODELO PREDICTIVO IBEX 35 ENERGY

Algunas de las palabras utilizadas en el modelo predictivo del Ibex-Energy son:

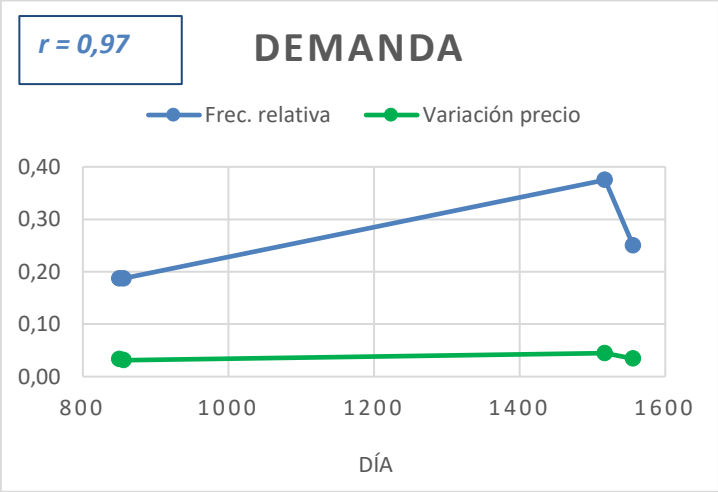

Figura 4.17. Gráfico frecuencia "Demanda" y variación del Ibex-Energy (elaboración propia, 2023)

La primera palabra que se muestra del modelo del índice energético es "demanda", con una correlación positiva muy fuerte de 0,97. A medida que la frecuencia de aparición de la palabra

"demanda" en las noticias aumenta, la variación del precio de las acciones de las empresas energéticas también tiende a subir. Este fenómeno es consistente con lo que se esperaría en una economía de mercado, donde el valor de las acciones de las empresas sube con un aumento en la demanda de sus productos o servicios (suponiendo que el mercado está reaccionando a las noticias de un aumento en la "demanda de energía").

Aun así, la palabra "demanda" puede haberse referido a otros escenarios que también podrían tener un impacto positivo en las acciones del Ibex-Energy, como la "demanda de inversión" en empresas energéticas, debida a noticias positivas sobre el crecimiento futuro de estas, o a cambios en las políticas gubernamentales que favorecen al sector. Del mismo modo, puede haberse referido a "demanda de exportaciones" de las empresas energéticas españolas que tienen una presencia significativa en el mercado internacional. Por ejemplo, tanto Endesa como Iberdrola y Naturgy operan a nivel internacional en el sector del gas y la electricidad, y en el ámbito de las energías renovables, varias empresas españolas son líderes a nivel mundial. Es el caso de Iberdrola, que es uno de los mayores productores de energía eólica del mundo. Acciona y Siemens Gamesa también tienen una presencia internacional relevante en el sector de las renovables. Por último, podemos pensar que la palabra demanda puede haberse dado en un contexto de "demanda de tecnologías limpias o renovables" debido al esfuerzo por combatir el cambio climático.

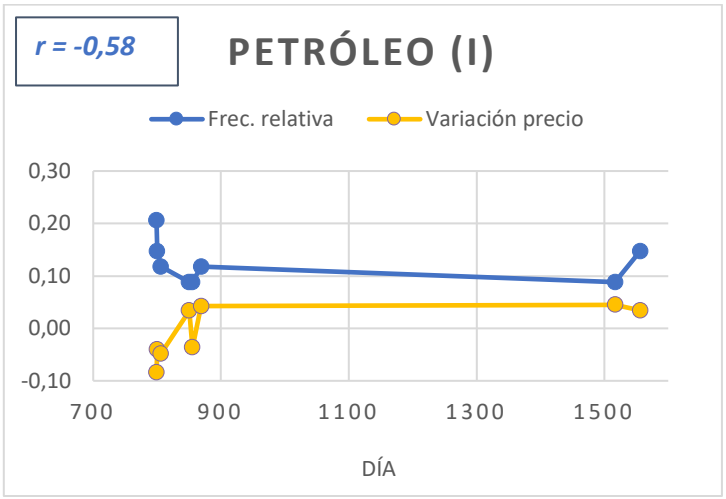

Figura 4.18. Gráfico frecuencia "Petróleo" y variación del Ibex-Energy (elaboración propia, 2023)

Para la palabra "petróleo", la correlación moderada y negativa de -0,58 sugiere que, en general, cuando esta palabra aparece con más frecuencia en las noticias, se tiende a ver una disminución en la variación del precio de las acciones de las empresas energéticas.

El sector energético está estrechamente ligado al precio y a la política del petróleo. Por lo tanto, noticias relacionadas con el petróleo como un cambio en los precios del petróleo, una alteración en la producción de petróleo, o conflictos geopolíticos en regiones productoras de petróleo pueden afectar al precio en Bolsa de las empresas energéticas.

Como puede observarse, la palabra "petróleo" aparece tanto en días de subidas como de bajadas, reflejando su relevancia en el sector tanto en momentos de crecimiento como de descenso. La correlación negativa puede indicar que la frecuencia de los titulares sobre el petróleo aumenta especialmente durante los periodos de crisis o incertidumbre, lo que a su vez lleva a una disminución en el precio de las acciones del Ibex-Energy.

En cuanto a la distribución de los días en los que aparece la palabra, está bastante dispersa a lo largo del tiempo, lo cual nos sugiere que las noticias relacionadas con el petróleo son una constante en el sector energético y que pueden influir en el precio de las acciones en cualquier momento.

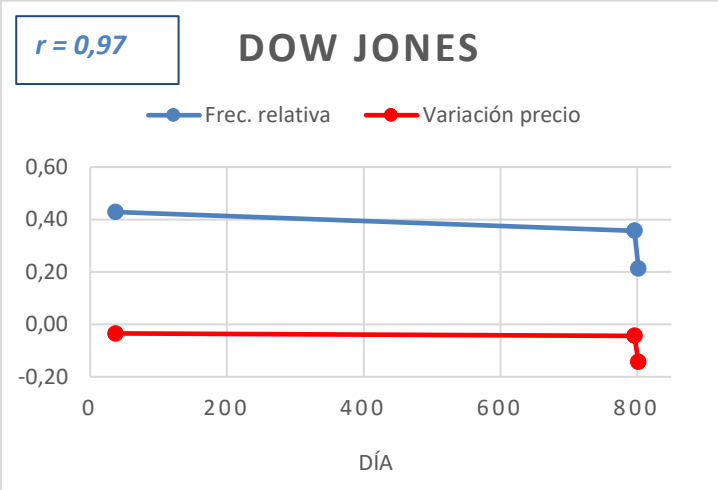

Figura 4.19. Gráfico frecuencia "Dow Jones" y variación del Ibex-Energy (elaboración propia, 2023)

Por último, el análisis de la *Figura 4.14* revela que existe una fuerte correlación positiva entre la frecuencia relativa de la palabra "Dow Jones"[22] y las variaciones en el precio de las acciones del índice Ibex-Energy. Esto podría parecer contradictorio, dado que los días en que la palabra "Dow Jones" aparece con mayor frecuencia corresponden a bajadas en la variación del precio

[22] El Dow Jones Industrial Average (DJIA), a menudo simplemente llamado "Dow Jones", es un índice de bolsa que representa a 30 de las mayores y más influyentes empresas de Estados Unidos.

de las acciones. Sin embargo, el resultado puede indicar que los eventos adversos que afectan al Dow Jones, uno de los índices de referencia más importantes del mundo, también tienden a influir negativamente en el Ibex-Energy.

La explicación lógica es que los mercados financieros de todo el mundo están fuertemente interconectados, y a menudo se mueven en sincronía debido a eventos macroeconómicos globales, flujos de capital transfronterizos y sentimientos del inversor que se difunden a nivel internacional.

4.1.1.5 MODELO PREDICTIVO ENDESA (ELE)

Finalmente, solo se incluyen dos de las palabras utilizadas en el modelo predictivo para el índice ELE:

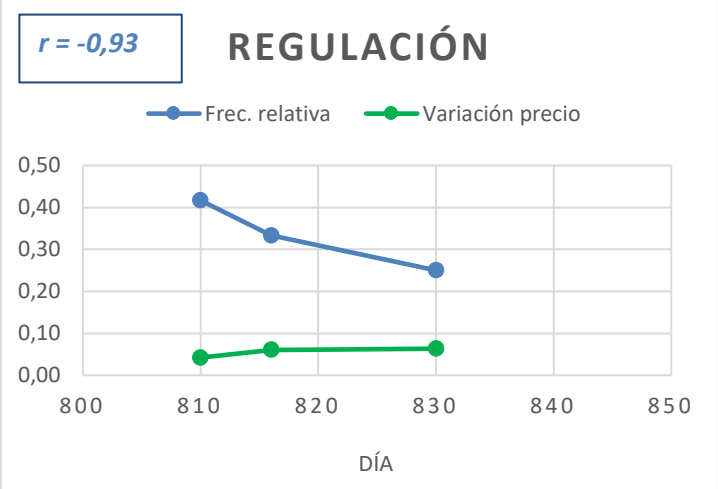

Figura 4.20. Gráfico frecuencia "Regulación" y variación de ELE (elaboración propia, 2023)

Los resultados muestran una correlación negativa significativa de -0,93 entre la frecuencia relativa de la palabra "regulación" y las variaciones en el precio de la acción de Endesa, pese a que esta palabra sólo aparece en días de subidas en el precio de la acción. Esto parece un poco contradictorio a primera vista, pero lo que este coeficiente de correlación nos está mostrando es que a medida que disminuye la frecuencia de la palabra "regulación" en las noticias, es decir, cuanto menos se habla de regulaciones, mayor es la subida en el precio de las acciones de Endesa. De ahí que se trate de una correlación negativa.

Los días en los que aparece el término en cuestión están relativamente concentrados, con un lapso de 20 días entre la primera y la última aparición. Esto puede sugerir que durante este

período hubo un debate intensivo sobre la regulación del sector energético, o que se estaban introduciendo o modificando ciertas normativas.

En general, la introducción de regulaciones más estrictas puede ser vista con escepticismo por los inversores, ya que pueden aumentar los costes de cumplimiento y reducir las ganancias. Sin embargo, en algunos casos, las nuevas regulaciones pueden ser vistas como un positivo, especialmente si benefician a la empresa de alguna manera. En este caso, el hecho de que su aparición coincida solo con subidas puede indicar que los inversores tenían la expectativa de que las nuevas regulaciones favorecerían a Endesa, pero que al irse reduciendo el tema de la regulación, se eliminó un factor que estaba frenando el precio de la acción.

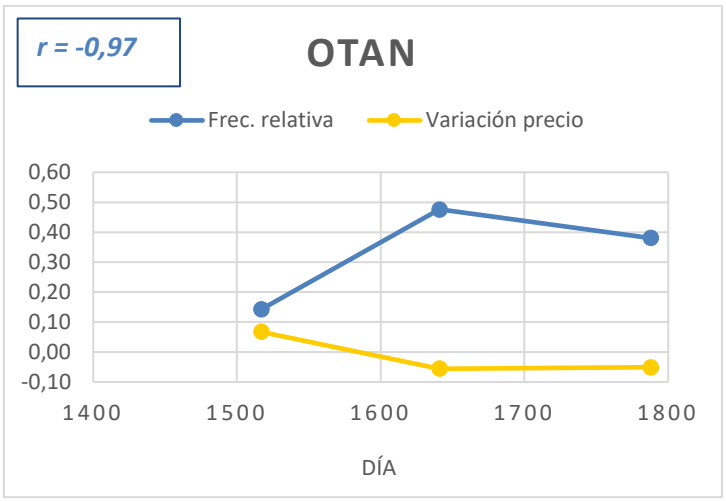

Figura 4.21. Gráfico frecuencia "OTAN" y variación de ELE (elaboración propia, 2023)

El gráfico muestra una fuerte correlación negativa (-0,97) entre la frecuencia relativa de la palabra "OTAN[23]" y las variaciones en el precio de las acciones de Endesa. El hecho de que la palabra aparezca en días que corresponden tanto a subidas como a bajadas del precio de las acciones, podría indicar que las noticias relacionadas con la OTAN tienen un impacto variable para Endesa, probablemente dependiendo del contexto específico de esas noticias.

Los días en que aparece la palabra "OTAN" parecen agruparse en torno a ciertos eventos o periodos, que coinciden con desarrollos significativos en el conflicto entre Rusia y Ucrania. Las tensiones geopolíticas, como un conflicto militar, tienen efectos significativos en los

[23] La Organización del Tratado del Atlántico Norte es una alianza militar intergubernamental entre países de Europa y Norteamérica. Constituye un enlace único entre estos dos continentes, lo que les permite cooperar en el campo de la defensa y la seguridad

mercados globales. En particular, el sector energético puede verse afectado por cuestiones como la seguridad del suministro y los cambios en los precios de la energía.

El análisis gráfico de las palabras clave y sus correlaciones con la variación del precio de las acciones de los índices energéticos estudiados ha revelado patrones intrigantes y potencialmente valiosos. Se han observado correlaciones tanto positivas como negativas entre las frecuencias de palabras y los movimientos del mercado, lo que sugiere que el modelo está captando con éxito la interacción entre las noticias y los mercados financieros.

Aunque estas palabras conforman una base sólida sobre la que podemos ajustar y afinar, el modelo es sencillo y puede no capturar todas las complejidades de la relación entre las noticias y los precios de las acciones. Sin embargo, proporciona una forma intuitiva y fácil de implementar para hacer predicciones basadas en los datos disponibles. Además, tiene la ventaja de ser fácilmente interpretable, ya que se puede ver directamente cómo cada palabra contribuye a la predicción.

1.4 EVALUACIÓN Y MEJORA DEL MODELO

El ajuste del modelo implicó la asignación de pesos a cada palabra clave en base a su correlación histórica con la variación del precio de las acciones. A través de este método, se estableció una medida cuantitativa de la influencia relativa de cada palabra clave sobre la variación en el precio de las acciones.

El rendimiento del modelo fue mejorado tras modificar algunos de sus parámetros, como el número de palabras clave consideradas. El número de palabras clave se redujo, incluyendo en el modelo solo y exclusivamente aquellas palabras frecuentes que se presentaban en al menos tres días de variaciones significativas. En consecuencia, se excluyeron las palabras que sólo aparecían en uno o dos de estos días, a pesar de que su frecuencia en ese día fuera alta. Este criterio de selección permitió evitar la inclusión de palabras que pudieran haber aparecido en las noticias de días de grandes subidas o bajadas por mera coincidencia, mejorando de esta manera la robustez del modelo.

La fase de evaluación permite refinar el modelo de predicción para el conjunto de datos y el contexto específicos del estudio. Sin embargo, en el *Capítulo 5* se va a evaluar el rendimiento

del modelo aplicado a un conjunto de datos de prueba independiente del conjunto de datos de entrenamiento utilizado para ajustar el modelo. Esto proporciona una medida objetiva de la generalización del modelo, es decir, de cómo podría funcionar en la práctica al predecir la variación de los precios de un índice basándose en nuevas noticias que no se habían visto antes.

El trabajo de campo realizado proporciona una sólida base empírica para respaldar el enfoque propuesto en este estudio. Sin embargo, la teoría y los modelos sólo son tan buenos como su aplicación en escenarios reales. Con esto en mente, avanzamos hacia el siguiente paso lógico en la investigación: la aplicación práctica. La *Sección 4.2* de este estudio se dedicará a poner a prueba el modelo en el mundo real, demostrando su utilidad y eficacia en la predicción de las fluctuaciones de los precios de las acciones basadas en el análisis de las noticias. Como se ha mencionado, esta fase permitirá validar el modelo en un escenario práctico, y proporcionará una prueba concluyente de su capacidad para informar decisiones de inversión más inteligentes.

4.2 APLICACIÓN PRÁCTICA DEL MODELO

Después de un detallado y riguroso recorrido a través de las etapas de desarrollo teórico y construcción del modelo, esta sección marca la transición hacia la comprobación práctica de su efectividad. No es simplemente un paso crucial, sino también la cúspide del trabajo realizado, en el que se verifica si la robustez teórica se traduce en un rendimiento eficaz en situaciones reales.

Para llevar a cabo la fase de aplicación del modelo, se seleccionó el día 8 de marzo de 2023, una fecha aleatoria fuera del conjunto de datos originalmente utilizado para entrenar el modelo.

Para ese día específico, se recolectaron titulares y subtítulos de noticias financieras publicadas en El País, Expansión, El Mundo y El Confidencial. El análisis de estas fuentes de noticias implica un examen detallado de las narrativas que podrían haber influido en el comportamiento del mercado. Tras la recopilación de las noticias, se llevó a cabo su preprocesamiento utilizando el código Python que se había desarrollado anteriormente. A través de este proceso, las noticias se desglosaron en palabras clave, ordenadas por la

frecuencia con la que aparecieron en las noticias del día. Estas palabras más relevantes son las que potencialmente influenciaron el comportamiento del mercado para los tres índices en estudio, IBEX 35, IBEX-Energy y Endesa.

Para cada índice, se hicieron coincidir las palabras clave extraídas del día de estudio con las palabras clave presentes en el modelo, que habían demostrado tener una correlación histórica con la variación del precio de las acciones. A cada palabra coincidente se le asignó un peso, igual al coeficiente de correlación de Pearson *r* obtenido a través del modelo. Por ejemplo, para el IBEX 35, si las palabras 'recuperación', 'precios', 'plan', 'subida' e 'inversión' fueron mencionadas en las noticias del 8 de marzo, se asignó a estas palabras los pesos correspondientes de 0,33, 0,63, -0,9, 0,91, -0,99 respectivamente, según el modelo.

Posteriormente, la predicción para la variación del precio de cada índice se calculó como la suma ponderada de los coeficientes de correlación de las palabras clave, lo que esencialmente es una forma de regresión lineal[24]. Es decir, en nuestro caso los parámetros β del modelo son los coeficientes de Pearson *r*. Recordando la ecuación 3.1 de regresión lineal del *Capítulo 3*, las predicciones tienen la forma:

$$\Delta P = \sum(r_i \alpha_i) + \varepsilon \tag{4.2}$$

Por lo tanto, la variación porcentual estimada para el IBEX 35 en el día de estudio se calculó de la siguiente manera:

$$\Delta P_{Ibex} = 0{,}33 * \alpha_{recuperación} + 0{,}63 * \alpha_{pymes} - 0{,}90 * \alpha_{plan} + 0{,}91 * \alpha_{subida} - 0{,}99 * \alpha_{inversión}$$

El procedimiento fue similar para los índices IBEX Energy y Endesa (ELE):

$$\Delta P_{Ibex-Energy} = 0{,}24 * \alpha_{EEUU} - 0{,}99 * \alpha_{renovable} - 0{,}37 * \alpha_{seguridad} + 0{,}97 * \alpha_{demanda} - 0{,}58 * \alpha_{petróleo}$$

[24] Sin embargo, cabe destacar que la forma en que se determinan los pesos (basándose en las correlaciones históricas) es un poco diferente de cómo se suelen calcular los coeficientes en una regresión lineal clásica, donde se usaría un método como el de mínimos cuadrados ordinarios para estimar los coeficientes que minimizan la suma de los residuos cuadrados

$$\Delta P_{ELE} = 0{,}60 * \alpha_{recuperación} - 0{,}84 * \alpha_{precios} + 0{,}99 * \alpha_{megavatio} + 0{,}43 * \alpha_{renovable} + 0{,}60 * \alpha_{luz} + 0{,}33 * \alpha_{gas} - 0{,}82 * \alpha_{demanda} - 0{,}70 * \alpha_{inversión} - 0{,}37 * \alpha_{petróleo}$$

El procedimiento llevado a cabo en la aplicación práctica del modelo se muestra de forma más visual en la siguiente *Tabla 4.6*. Además, se añaden los resultados de la predicción. Cabe mencionar que no se han incluido todas las palabras clave obtenidas tras el procesamiento de las noticias del 8 de marzo, sino que solo se han incluido aquellas que coinciden con alguno de los tres modelos entrenados, IBEX, IBEX-Energy y ELE.

Fecha	*Palabra*	*Frecuencia*	*IBEX*	*IBEX-Energy*	*Endesa*
08.03.2023	eeuu	6		0,24	
08.03.2023	recuperación	5	0,33		0,6
08.03.2023	precios	5			-0,84
08.03.2023	pymes	5	0,63		
08.03.2023	megavatio	5			0,99
08.03.2023	renovable	4		-0,99	0,43
08.03.2023	plan	4	-0,9		
08.03.2023	luz	4			0,6
08.03.2023	subida	3	0,91		
08.03.2023	seguridad	3		0,37	
08.03.2023	gas	3			0,33
08.03.2023	demanda	3		0,97	-0,82
08.03.2023	inversión	3	-0,99		-0,7
08.03.2023	petróleo	2		-0,58	-0,37
		PREDICCIÓN	**0,96**	**0,34**	**3,56**

Tabla 4.7. Predicción de la variación del precio de acciones, 8 de marzo de 2023 (elaboración propia, 2023)

Con la aplicación práctica del modelo a los datos del 8 de marzo de 2023, hemos concluido el trabajo de campo. En esta fase, hemos podido observar cómo nuestro modelo procesa los datos en tiempo real, adaptándose a las particularidades del día seleccionado y produciendo estimaciones para las variaciones de precios de los índices Ibex, Ibex-Energy y las acciones de Endesa. Este experimento nos permite tener una primera visión de cómo nuestro modelo se comporta fuera del conjunto de datos de entrenamiento, aportando una visión crucial de su potencial de generalización.

Con esto, cerramos el *Capítulo 4* del trabajo. A lo largo de este capítulo, hemos recopilado y preprocesado los datos, construido el modelo, ajustado sus parámetros y finalmente, lo hemos probado en un entorno práctico. Este proceso nos ha permitido entender mejor la relación entre las noticias y las variaciones en los precios de las acciones, así como el papel de las palabras clave en esta relación.

A medida que avanzamos hacia el *Capítulo 5*, nuestro enfoque cambiará de la creación y aplicación del modelo a su evaluación y análisis. Confrontaremos las predicciones con los valores reales para proporcionar una visión cuantitativa y cualitativa de su rendimiento. Asimismo, utilizaremos métricas como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2) para evaluar la precisión de las predicciones y la eficacia del modelo. Así, se podrán comprender mejor sus fortalezas y limitaciones, y plantear mejoras para futuras iteraciones y desarrollos.

CAPÍTULO 5. ANÁLISIS DE RESULTADOS

Tras el desarrollo y la aplicación práctica del modelo predictivo, el siguiente paso crítico es examinar y evaluar los resultados obtenidos. En el presente capítulo, se realiza un examen cuidadoso y exhaustivo de los hallazgos, cuyo propósito no sólo es evaluar la efectividad del modelo, sino también sacar conclusiones valiosas que puedan ayudar a guiar futuros trabajos e investigaciones. Este análisis, por lo tanto, no es simplemente una revisión de los resultados, sino una reflexión que proporciona una nueva perspectiva y una oportunidad para extraer enseñanzas profundas. Más allá de los números, pretende entender en profundidad los hallazgos obtenidos, sus implicaciones y su relevancia en la tarea de predicción del movimiento del mercado de valores.

La confrontación de las predicciones con los valores reales es un componente esencial en la evaluación de cualquier modelo. Así pues, al comparar estas dos cifras, estamos evaluando cómo de bien nuestro modelo puede replicar la complejidad del mundo real. No sólo estamos evaluando la precisión de las cifras, sino también la capacidad del modelo para captar y entender las tendencias y patrones subyacentes en los datos.

Para el Ibex, Ibex-Energy y Endesa, las predicciones obtenidas fueron 0,96%, 0,34% y 3,56% respectivamente. Comparándolas con los valores reales de 0,58%, 0,59% y 4,49%, se puede observar cierta proximidad, lo que a primera vista indica que el modelo logró predecir las variaciones en los precios de las acciones de manera razonable. Estos resultados se muestran en la *Tabla 5.1* a continuación:

Variación 8 marzo 2023	***IBEX***	***IBEX-Energy***	***ELE***
Predicción	0,96%	0,34%	3,56%
Real	0,58%	0,59%	4,49%

Tabla 5.1. Variaciones estimadas y reales de las acciones a 8 de marzo de 2023 (elaboración propia, 2023)

A continuación, se muestran tres gráficos de líneas que representan el precio de cierre de la acción de cada uno de los índices estudiados (en azul), durante los dos meses y medio que han

transcurrido desde el principio de año 2023. Como estamos haciendo predicciones diarias, se deben representar los últimos 30 a 90 días, pues se trata de un marco de tiempo comúnmente utilizado en el análisis de acciones que permite visualizar fácilmente las variaciones diarias. En estos gráficos, se representa con un marcador rojo el precio de la acción estimado para el 8 de marzo de 2023, con el objetivo de poder comparar los resultados con los valores reales de forma visual.

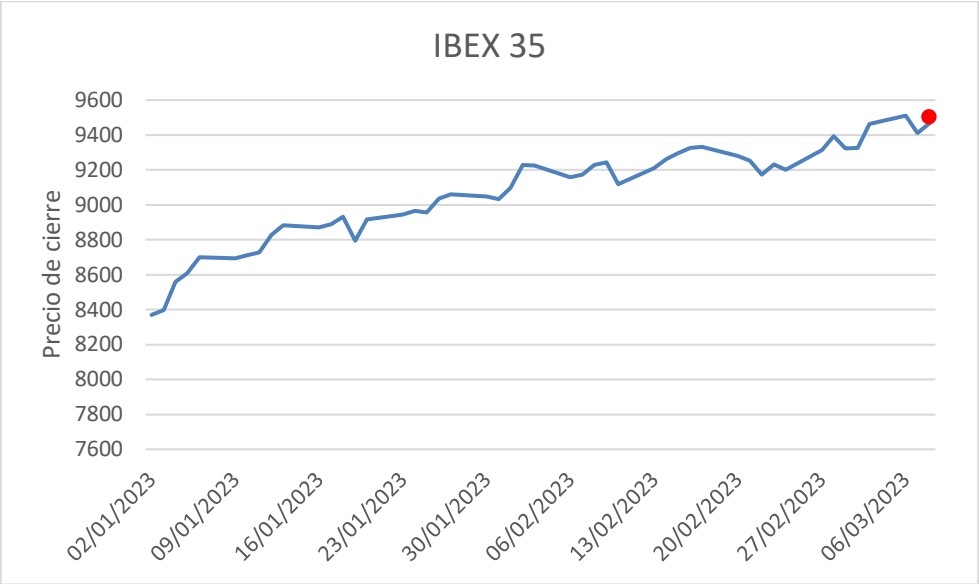

Figura 5.1. Evolución del precio del Ibex en 2023 (elaboración propia, 2023)

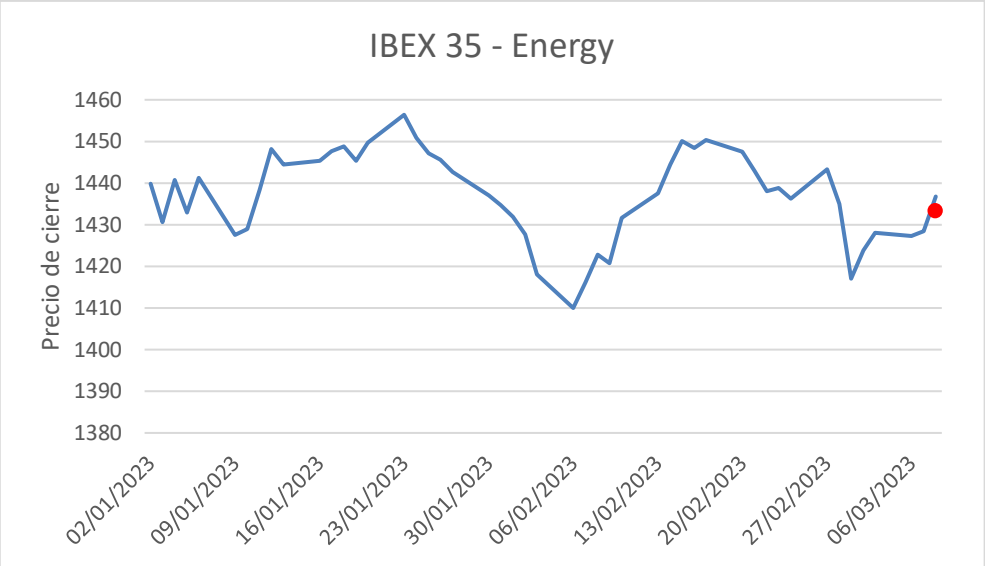

Figura 5.2. Evolución del precio del Ibex-Energy en 2023 (elaboración propia, 2023)

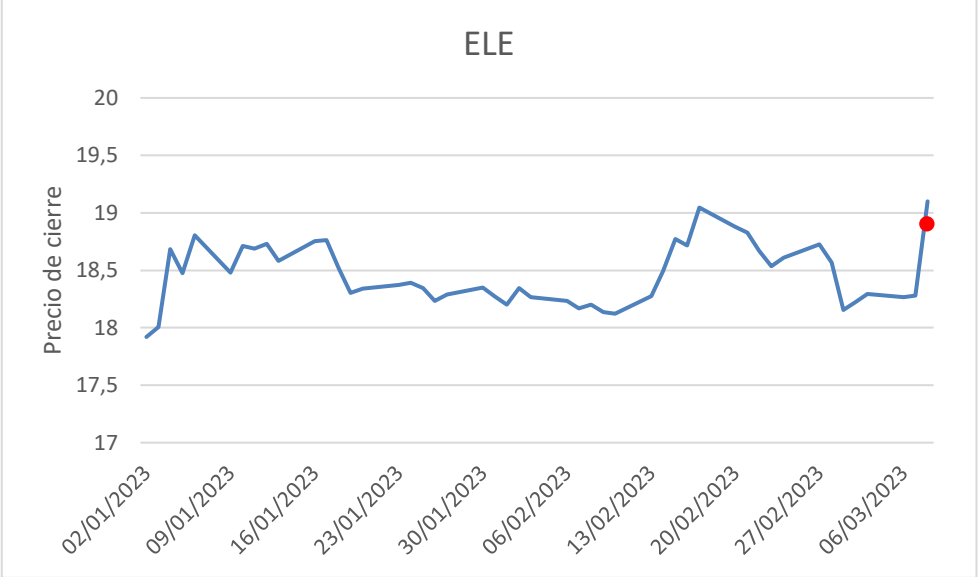

Figura 5.3. Evolución del precio de ELE en 2023 (elaboración propia, 2023)

Antes de evaluar el rendimiento del modelo de forma cuantitativa, es fundamental resaltar una de las fortalezas notables del modelo: su capacidad para discernir correctamente la dirección de las variaciones de precios, es decir, predecir si los precios de las acciones experimentarán un incremento (subida) o una disminución (bajada). Este logro, aunque pueda parecer básico, es en realidad un aspecto significativo de cualquier sistema predictivo en el ámbito financiero. Ser capaz de anticipar de manera fiable la tendencia de los movimientos de los precios es una cualidad de gran valor, ya que el simple conocimiento de la dirección que toma el activo permite a los inversionistas tomar decisiones informadas y gestionar eficazmente el riesgo. Por lo tanto, el hecho de que el modelo haya demostrado esta capacidad sugiere un potencial considerable para su utilidad práctica en el campo financiero.

Para cuantificar el grado de precisión de las predicciones mediante el MSE, mencionado en el *Capítulo 3*, primero se calcularon los errores cuadráticos para cada índice. Recordamos:

$$Error\ cuadr\ = (real - estimado)^2 \tag{5.1}$$

Los resultados, 0,144, 0,063 y 0,865 para el Ibex, Ibex-Energy y ELE respectivamente, proporcionan una medida cuantitativa de la desviación de las predicciones del modelo respecto a los valores reales observados. El promedio de estos errores cuadráticos es el conocido error cuadrático medio (MSE), que resultó ser ***MSE = 0,357***. Este resultado significa que, en promedio, las predicciones del modelo difieren de los valores reales por aproximadamente 0,357 cuando se consideran las diferencias al cuadrado. El MSE puede interpretarse como una

medida de la dispersión de los errores de predicción, por lo que un valor de MSE más cercano a cero sería ideal, ya que indica que las predicciones del modelo están muy cerca de los valores reales.

El valor de MSE debe interpretarse en el contexto del problema y los datos específicos. Un MSE de 0,357 puede ser aceptable, e incluso excelente, para ciertos conjuntos de datos y problemas, mientras que puede considerarse alto para otros [30]. Por ejemplo, en un problema donde las variaciones de precio son generalmente pequeñas, el MSE se puede considerar alto, pero si las variaciones de precio pueden llegar a ser muy grandes, un MSE de 0,357 indica una predicción muy precisa.

En este caso, dada la complejidad inherente de predecir mediante narrativa periodística los precios de las acciones, que pueden ser influenciados por una gran variedad de factores y son notoriamente volátiles, un MSE de 0,357 sugiere que el modelo está proporcionando predicciones con una precisión muy razonable.

Además, se quiso complementar el resultado del error cuadrático medio con el coeficiente de determinación (R^2). Este coeficiente, que también se conoce como el cuadrado del coeficiente de correlación de Pearson, proporciona una medida de qué tan bien las predicciones del modelo se ajustan a los valores reales observados [9]. Los valores de R^2 varían entre 0 y 1, donde un valor de 1 indica un ajuste perfecto, es decir, todas las predicciones del modelo coinciden exactamente con los valores reales. Un valor de 0, por otro lado, indica que el modelo no logra explicar ninguna de la variabilidad en los datos.

Para calcular R^2, necesitamos calcular primero dos sumas de cuadrados: la suma total de los cuadrados (SST) y la suma de los cuadrados de los residuos (SSE). SST mide la variabilidad total en los datos, y se calcula como la suma de los cuadrados de las diferencias entre cada valor real y la media de los valores reales:

$$SST = \sum(\Delta P_i - \overline{\Delta P})^2 = 10{,}17 \tag{5.2}$$

SSE mide la variabilidad que el modelo no pudo explicar, y se calcula como la suma de los cuadrados de las diferencias entre cada valor real y su correspondiente valor predicho por el modelo:

$$SSE = \sum(\Delta P_i - \widehat{\Delta P_i})^2 = 1{,}07 \tag{5.3}$$

El coeficiente de determinación se calcula entonces como:

$$\boldsymbol{R^2 = 1 - \frac{SSE}{SST} = 0,895} \quad (5.4)$$

Este valor representa la proporción de la variabilidad total de los precios de las acciones que es explicada por el modelo. Un R^2 de 0,895 implica que el 89,5% de los cambios en la variable de salida (la variación del precio de las acciones) pueden ser explicados por el modelo, por lo que es bastante preciso.

Las métricas del MSE y R^2 proporcionan evidencia de que el modelo ha logrado predecir con cierta precisión las variaciones en los precios de las acciones. Pero estos números, por muy importantes que sean, sólo cuentan una parte de la historia. También es importante entender por qué nuestro modelo ha producido los resultados que ha producido. ¿Qué aspectos del modelo funcionaron bien y cuáles no? ¿Cómo podemos mejorar el modelo para futuras iteraciones?

La respuesta a estas preguntas no sólo nos ayudará a entender mejor el modelo actual, sino que también nos proporcionará información valiosa que podemos utilizar para mejorar y perfeccionar el modelo en el futuro.

Por lo tanto, y a pesar de los resultados alentadores mostrados por el modelo en términos de MSE y R^2, es esencial abordar sus limitaciones y desafíos para proporcionar una visión más completa del rendimiento del modelo y plantear recomendaciones para futuras iteraciones.

En primer lugar, nos encontramos con el problema de la temporalidad y precisión de las predicciones. Aunque el modelo ha demostrado cierta precisión, esta no está garantizada para futuras predicciones. Diversos factores dinámicos y no controlados, como la volatilidad del mercado, factores macroeconómicos, cambios regulatorios inesperados y acontecimientos imprevistos pueden afectar significativamente los precios de las acciones y desafiar la precisión de las predicciones de nuestro modelo. Además, el modelo actual asume que la frecuencia de palabras clave en las noticias matutinas tiene un impacto directo en las variaciones de precios de ese mismo día, pero esta relación temporal podría no ser válida en todas las circunstancias y puede necesitar ajustes y refinamiento.

En segundo lugar, existe una limitación en la selección de palabras clave. La eficacia del modelo se basa en gran medida en la selección de palabras clave relevantes en las noticias, y,

por ejemplo, se selecciona la palabra clave "regulación" debido a su relevancia percibida en el sector energético. Sin embargo, hay un amplio espectro de palabras clave que podrían ser relevantes y que estamos descartando. Por lo tanto, la selección de palabras clave puede limitar la capacidad del modelo para capturar toda la gama de noticias relevantes que pueden influir en las variaciones de precios.

En tercer lugar, nos encontramos con una ausencia de análisis de sentimiento. En su estado actual, el modelo no tiene en cuenta el sentimiento o tono asociado con las noticias, ya sea positivo, negativo o neutral. Esta es una limitación significativa, ya que el sentimiento de las noticias puede influir en cómo se interpretan y, en consecuencia, en cómo afectan los precios de las acciones. Incorporar un análisis de sentimiento en el modelo podría permitirnos captar con mayor precisión el impacto de las noticias en las variaciones de precios.

Por último, la interpretación de las variaciones de los precios de las acciones es realmente un desafío. La variación de los precios de las acciones es el resultado de un complejo conjunto de factores, más allá de la mera frecuencia de palabras clave en las noticias. Estos factores pueden incluir la salud financiera de la empresa, su desempeño en comparación con sus competidores, la situación económica en general, y los desarrollos políticos, entre otros. El modelo actual no captura estos factores, lo que puede limitar su capacidad de predicción

En resumen, los resultados obtenidos en este análisis reflejan la eficacia potencial del modelo para la predicción de las variaciones de precios de los índices IBEX, IBEX-Energy y ELE. Destaca la capacidad del modelo para discernir correctamente la dirección de las variaciones de precios, es decir, si el precio sube o baja. Teniendo en cuenta los resultados obtenidos a lo largo del análisis, se puede afirmar con cierto grado de seguridad que la hipótesis alternativa (H1), que sostiene que determinados sucesos, temáticas o sentimientos reflejados en las noticias tienen el potencial de predecir las dinámicas futuras del mercado bursátil, se ha confirmado. A través de la implementación y evaluación del modelo propuesto, se ha evidenciado que las características derivadas de las noticias desempeñan un papel significativo en la predicción de las variaciones de los precios de las acciones. Esto resulta en la refutación de la hipótesis nula (H0) que sostiene que no existe una relación significativa entre las características derivadas de las noticias y las variaciones en los precios de las acciones.

No obstante, los resultados también subrayan la necesidad de continuar refinando y mejorando el modelo para incrementar su precisión y utilidad. Nuestro viaje en el complejo mundo de la

predicción de precios de acciones apenas está comenzando, pues el verdadero valor del trabajo vendrá a través de iteraciones sucesivas y mejoras, y de la aplicación de lo que hemos aprendido a nuevos conjuntos de datos y contextos. Finalmente, las limitaciones mencionadas en este capítulo proporcionan áreas de investigación futura y de mejora del modelo, que serán revisadas en el *Capítulo 7*.

CAPÍTULO 6. MEMORIA ECONÓMICA

Este Trabajo de Fin de Grado ha centrado sus esfuerzos en el desarrollo de un modelo predictivo basado en el análisis de noticias para prever las fluctuaciones en el mercado de valores. A lo largo del proyecto, se ha llevado a cabo la formulación de un modelo teórico sólido, la implementación práctica de dicho modelo y, finalmente, el análisis exhaustivo de los resultados obtenidos. Ahora, llega el momento de concebir cómo este modelo se podría trasladar al mundo real en forma de un negocio operativo.

La mejor forma de implementar y comercializar este modelo predictivo es mediante la creación de una empresa de tecnología financiera, también conocida como fintech. Esta startup se centraría en el uso de la tecnología para ofrecer servicios financieros innovadores, empleando el análisis de sentimientos y el aprendizaje automático para proporcionar una ventaja informativa en los mercados financieros. Del mismo modo, la startup aspira a proporcionar a sus usuarios pronósticos financieros precisos, basados en un análisis avanzado de las noticias.

Por lo tanto, al considerar la creación de una startup, es fundamental llevar a cabo un estudio económico riguroso. En este capítulo, se analizarán dos aspectos clave: el coste de implementación y operación del proyecto, y la rentabilidad potencial que se puede esperar. El objetivo de este análisis es evaluar si la creación de la fintech sería económicamente viable y rentable, basándose en las suposiciones[25] y cálculos que se describen a continuación [50].

6.1 COSTE DEL PROYECTO

El diseño e implementación de un proyecto innovador como una startup fintech conlleva una serie de costes asociados. Estos costes se dividen generalmente en dos categorías principales: el Coste de Adquisición de Capital (CAPEX) y el Coste Operativo (OPEX). CAPEX hace

[25] Para estimar de manera precisa y realista los costes asociados con la creación y mantenimiento de la startup fintech, el estudio se basa en la infraestructura y modelo de negocio de Yukka Lab (Berlín), una empresa bien establecida y exitosa en el campo de las fintechs. Se ha estudiado su estructura organizativa, sus inversiones en tecnología y la escala de sus operaciones para elaborar un modelo de negocio que refleje de manera realista las demandas financieras que enfrentaría nuestra propia startup.

referencia a los gastos iniciales para poner en marcha el proyecto, mientras que OPEX se refiere a los costes recurrentes necesarios para mantener el proyecto en funcionamiento [45]. En este apartado, se presentará un desglose detallado de estos para proporcionar una visión clara de la inversión necesaria para poner en marcha la empresa de tecnología financiera. Este análisis resulta esencial para determinar la viabilidad económica del proyecto.

6.1.1 INVERSIONES INICIALES (CAPEX)

El desarrollo de este proyecto fintech requiere una serie de inversiones iniciales, también conocidas como gastos de capital (CAPEX)[26]. Estos son los costes que se incurren para crear la capacidad de producción del modelo y son costes no recurrentes.

- **Infraestructura de IT**: se necesitará un conjunto de servidores para alojar la plataforma y realizar los cálculos necesarios para el modelo predictivo. Tomando como referencia los costes medios del mercado, podemos estimar una inversión inicial de unos 20.000€.

- **Desarrollo de Software**: para implementar el modelo y construir la interfaz de usuario y las aplicaciones de software, se necesitará una inversión inicial de desarrollo. Estimamos esta inversión en alrededor de 50.000€, incluyendo la contratación de desarrolladores expertos y las licencias necesarias de software.

- **Instalaciones de la oficina**: supondremos una inversión inicial de 10.000€ para adecuar un espacio de trabajo para el equipo.

Por lo tanto, el CAPEX total será:

$$CAPEX = 20.000 + 50.000 + 10.000 = 80.000€ \quad (6.1)$$

Se incluye en la *Figura 6.1* un gráfico de barras con cada uno de los elementos del CAPEX, con el objetivo de mostrar cuánto contribuye cada elemento al CAPEX total.

[26] Por sus siglas en inglés, Capital Expenditure.

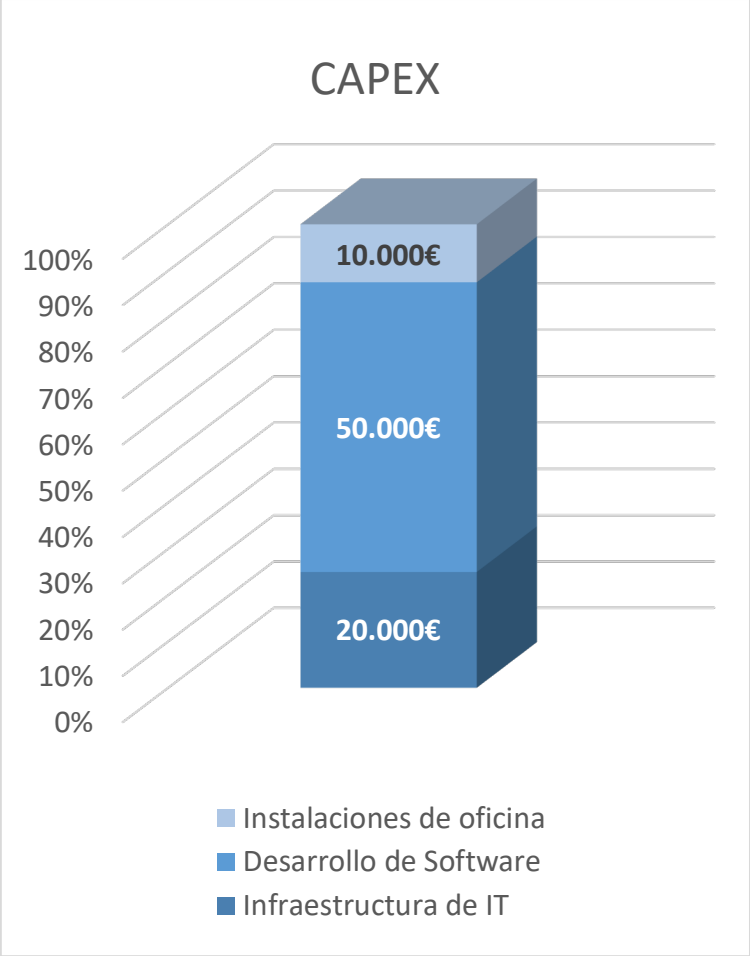

Figura 6.1. Elementos del CAPEX (elaboración propia, 2023)

6.1.2 GASTOS OPERATIVOS (OPEX)

Los gastos operativos (OPEX)[27] son los costes para el funcionamiento diario y el mantenimiento del modelo. Al contrario que el CAPEX, estos son costes recurrentes y se incurren mientras el modelo está en funcionamiento.

- **Personal**: se necesitará un equipo de ingenieros de datos, desarrolladores de software y un pequeño equipo de marketing y ventas. Considerando los salarios promedio en la industria, los costes anuales para un equipo de 8 personas pueden estimarse en 510.000€.

[27] Por sus siglas en inglés, Operating Expenditure.

- **Mantenimiento de IT**: se incluirían los costes de mantenimiento de los servidores y el soporte técnico, y podríamos suponer que estos costes serán del 10% del coste inicial de la infraestructura de IT cada año, lo que sería 2.000€.

- **Costes de oficina**: alquiler de oficinas, suministros y servicios públicos, que podríamos estimar en unos 30.000€ al año.

- **Adquisición de datos**: el proyecto requerirá el acceso continuo a los datos de noticias. Supongamos un coste de 10.000€ al año para esto.

- **Marketing y ventas**: supongamos costes anuales de 20.000€.

Si sumamos todos estos costes, obtenemos un OPEX total en el primer año de:

$$OPEX_{año\,1} = 510.000 + 2.000 + 30.000 + 10.000 + 20.000 = 572.000€ \qquad (6.2)$$

El costo de mantenimiento de IT es un porcentaje (10%) del coste inicial de la infraestructura de IT, que fue una inversión única de 20.000€. En el primer año, esto suma 2.000€ a los gastos operativos. Sin embargo, como esta inversión inicial en infraestructura solo se incurre en el primer año y no es un costo recurrente, no necesitamos considerar este coste de mantenimiento para los años subsiguientes.

Por lo tanto, el OPEX para los años subsiguientes será:

$$OPEX_{año\,n} = 510.000 + 30.000 + 10.000 + 20.000 = 570.000€ \qquad (6.3)$$

Donde n es cualquier año después del primer año.

Igual que para el CAPEX, se incluye en la *Figura 6.2* un gráfico de barras con cada uno de los elementos del OPEX para el año 1.

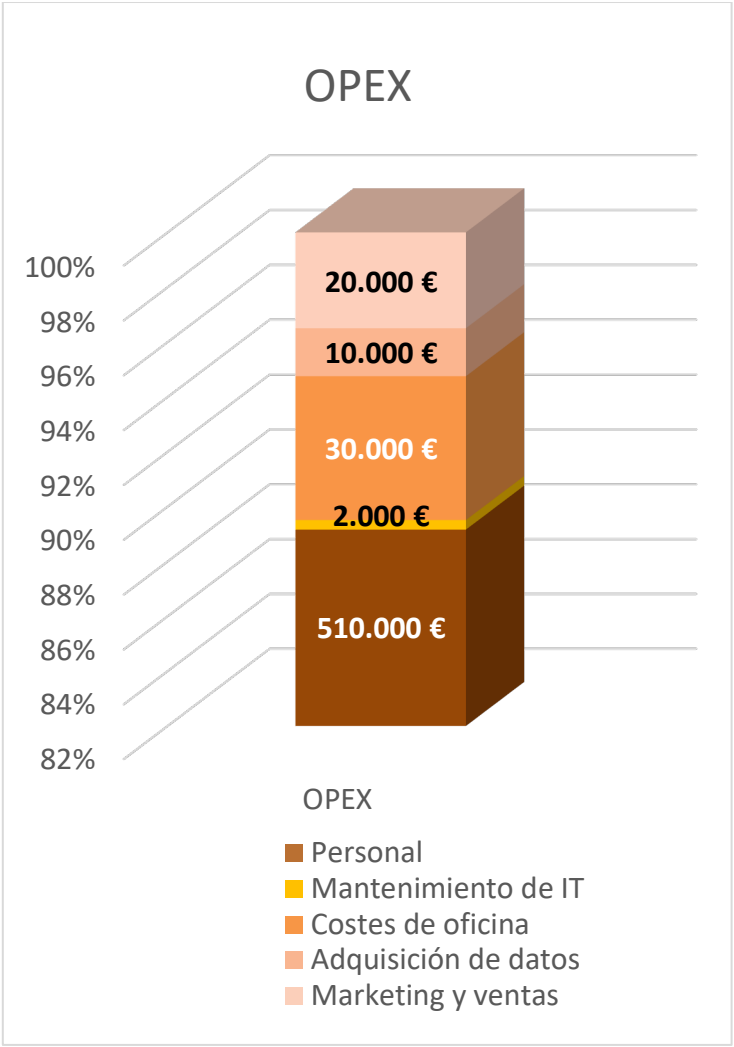

Figura 6.2. Elementos del OPEX del primer año (elaboración propia, 2023)

6.1.3 COSTE NORMALIZADO (LCOX)

Para obtener el Coste Normalizado (LCOX)[28], podemos tomar la suma de los costes CAPEX y OPEX y dividirla por la producción total del modelo [45]. Supongamos que el modelo puede procesar 1000 predicciones por día, lo que da 365.000 predicciones por año.

Así, para el primer año, el LCOX sería:

[28] Del inglés, Levelized Cost of X, siendo X la fintech en este caso

$$LCOX_{año\ 1} = \frac{80.000+572.000}{365.000} = 1{,}79€/predicción \quad (6.4)$$

y para los años siguientes sería:

$$LCOX_{año\ n} = \frac{80.000+570.000}{365.000} = 1{,}78€/predicción \quad (6.5)$$

Basándonos en estos cálculos, la conclusión que podemos obtener es que el proyecto fintech, tal y como se ha diseñado y estimado, parece tener un coste factible y sostenible. Con un LCOX estimado en 1,79€ por predicción en el primer año y 1,78€ en años subsiguientes, estos costes son razonablemente bajos y pueden ser compensados por los ingresos generados por cada predicción, siempre que el precio por predicción esté correctamente establecido.

El análisis de CAPEX y OPEX también indica que los costes de mantener la operación en los años posteriores al inicio son relativamente estables. Esto permite que la startup pueda planificar con anticipación su flujo de caja y presupuesto.

Cabe recordar que estos cálculos son estimaciones y los costes reales pueden variar en función de una serie de factores, pero proporcionan un marco útil para evaluar el coste del proyecto.

6.2 RENTABILIDAD DEL PROYECTO

Una vez determinados los costes asociados con el establecimiento y operación de la startup, el siguiente paso es evaluar su potencial rentabilidad. Este análisis económico es esencial para validar la viabilidad del proyecto y atraer posibles inversores. En este apartado, se calcula el Valor Actual Neto (VAN), un indicador clave de la rentabilidad de un proyecto, que considera los ingresos esperados, los costes incurridos y el valor del dinero en el tiempo [32]. El cálculo del VAN proporcionará una medida cuantitativa del rendimiento esperado de la inversión en la startup, y será un indicador crucial para decidir si el proyecto es económicamente viable y si merece salir adelante.

6.2.1 GENERACIÓN DE INGRESOS

Un modelo de negocio potencial para este proyecto fintech podría basarse en vender suscripciones a la plataforma de predicción de acciones. Además, podría existir un modelo de ingresos por comisión basado en la cantidad de transacciones realizadas a través de la plataforma.

Para realizar una proyección de ingresos, necesitamos algunas suposiciones:

- **Precio de suscripción mensual**: suponemos un precio de suscripción mensual de 100€.
- **Número de suscriptores**: en el primer año, se podrían alcanzar 500 suscriptores, y se espera un crecimiento del 50% cada año durante los primeros cinco años.
- **Comisión por transacción**: suponemos una comisión promedio de 1€ por transacción.
- **Número de transacciones**: vamos a suponer un promedio de 100 transacciones por suscriptor al mes.

Por lo tanto, los ingresos del primer año serían:

$$Ing_{año\ 1} = \left(500\ suscriptores * \frac{100€}{suscriptor} * 12\ meses\right) + \left(500\ suscriptores * \frac{100\ transacciones}{suscriptor} * \frac{1€}{transacción} * 12\ meses\right) = 1.200.000€ \quad (6.6)$$

Este cálculo se puede repetir para cada año, teniendo en cuenta el crecimiento esperado de suscriptores.

6.2.2 VALOR ACTUAL NETO (VAN)

El Valor Actual Neto (VAN) es un indicador de la rentabilidad de un proyecto y se calcula descontando los flujos de caja futuros a su valor actual. Por lo tanto, primero necesitamos calcular los flujos de caja para cada año, que se obtienen restando los costes de los ingresos para cada año [38].

El flujo de caja para el primer año sería:

$$FC_{año\ 1} = 1.200.000\ (ingresos) - 572.000\ (OPEX) - 80.000\ (CAPEX) = 548.000€ \quad (6.7)$$

Para los años subsiguientes, como no hay CAPEX, los flujos de caja serían:

$$FC_{año\ n} = 1.200.000\ (ingresos) - 570.000\ (OPEX) = 630.000€ \quad (6.8)$$

Luego, necesitamos descontar estos flujos de caja a su valor presente utilizando una tasa de descuento (r). La tasa de descuento podría ser la tasa de interés que se pagaría por pedir prestado el dinero para financiar el proyecto [38], digamos un 5%. El valor presente (VP) de los flujos de caja se calcula:

$$VP_{año\ n} = FC_{año\ n}/(1+r)^n \quad (6.9)$$

Finalmente, el VAN sería la suma de estos valores presentes de los flujos de caja para cada año.

Sustituyendo los valores y calculando el VAN para un período de 5 años, obtendríamos:

$$VAN = \frac{FC_{año\ 1}}{(1+0,05)} + \frac{FC_{año\ 2}}{(1+0,05)^2} + \frac{FC_{año\ 3}}{(1+0,05)^3} + \frac{FC_{año\ 4}}{(1+0,05)^4} + \frac{FC_{año\ 5}}{(1+0,05)^5} \approx 265M€ \quad (6.10)$$

Se adjunta un gráfico de líneas que muestra cómo el VAN cambia con el tiempo (durante el período de 5 años), lo cual proporciona una representación visual del rendimiento esperado del proyecto:

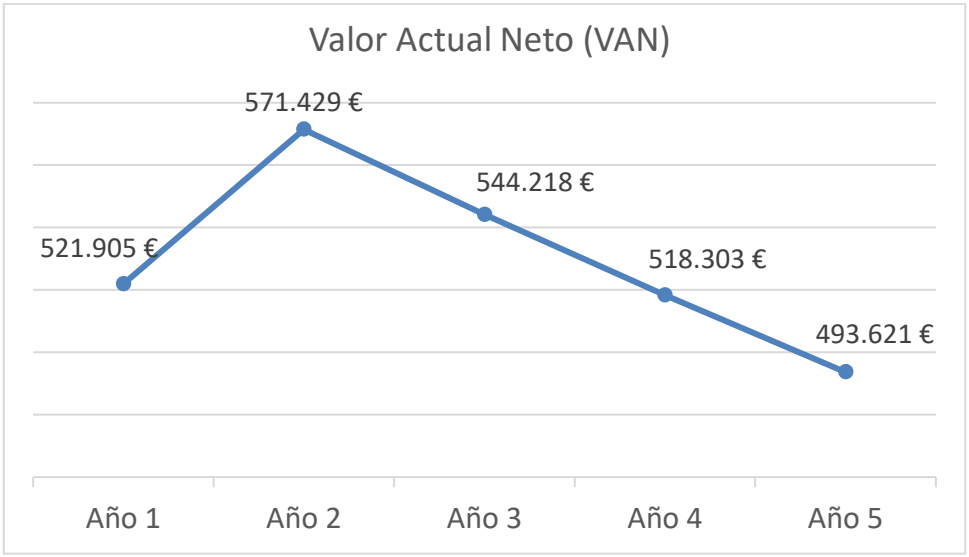

Figura 6.3. Evolución del VAN en 5 años (elaboración propia, 2023)

Basándonos en estos cálculos, la conclusión que podemos obtener es que el proyecto fintech, tal y como se ha diseñado y estimado, tiene un potencial significativo de rentabilidad.

El Valor Actual Neto (VAN) es positivo y considerablemente alto, estimado en alrededor de 265M€ para un período de 5 años. Este valor positivo indica que se espera que el proyecto genere más ingresos que costes a lo largo de su vida, incluso después de ajustar los flujos de caja futuros a su valor presente utilizando una tasa de descuento. Por este motivo, se trata de un indicador fuerte de la rentabilidad y viabilidad del proyecto.

Además, los flujos de caja proyectados son positivos para cada año, lo que indica que los ingresos superarán a los costes cada año. Esto sugiere que el proyecto no solo será rentable en términos globales, sino que también generará suficiente efectivo cada año para cubrir sus costes operativos y potencialmente proporcionar un rendimiento a los inversores.

CAPÍTULO 7. CONCLUSIONES Y TRABAJOS FUTUROS

El *Capítulo 7* representa la culminación de este Trabajo de Fin de Grado, un momento de reflexión y síntesis donde se integran los hallazgos y las lecciones aprendidas durante todo el proceso de investigación. Aquí se presentan las conclusiones, subrayando los logros más significativos y destacando la contribución única y valiosa de este estudio al campo de la predicción de precios de acciones.

Esta sección también se enfoca hacia el futuro, delineando las posibles direcciones para el desarrollo y la mejora continuos del modelo presentado. El campo de la predicción del mercado bursátil es dinámico y en constante evolución, y este estudio no es más que un paso en el viaje de descubrimiento. En definitiva, el capítulo pretende tanto mirar hacia atrás, reconociendo y valorando lo que se ha logrado, como mirar hacia adelante, identificando oportunidades y retos que pueden orientar los próximos pasos en este emocionante campo de estudio.

7.1 CONCLUSIONES

El presente Trabajo de Fin de Grado ha abordado el desafío de adentrarse en la confluencia entre el procesamiento del lenguaje natural, el aprendizaje automático y las finanzas cuantitativas. Este reto no solo ha permitido ampliar la comprensión de cada una de estas disciplinas individualmente, sino que también ha revelado una sinergia en ciernes con el potencial de transformar la forma en que se desarrolla el análisis financiero y se toman las decisiones de inversión.

Se han logrado una serie de hitos significativos que destacan nuestra contribución a la predicción de los precios de las acciones basándonos en el análisis de noticias financieras:

- ✓ Se ha construido un sólido marco conceptual para un modelo de predicción, diseñado específicamente para anticipar las fluctuaciones diarias de los precios de las acciones basándose en el análisis de noticias, mediante dos enfoques distintos: regresión lineal y redes neuronales.

- ✓ Se ha demostrado con éxito la aplicabilidad del marco teórico en un entorno real, dando lugar a predicciones palpables y precisas de los movimientos de precios. La eficacia del modelo ha sido validada a través de una evaluación exhaustiva, mediante el uso de métricas como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2).
- ✓ Se ha detallado un estudio económico del proyecto, demostrando que nuestro modelo no solo es efectivo desde el punto de vista técnico, sino que también tiene un potencial económico significativo.

La realización de este Trabajo de Fin de Grado (TFG) ha permitido ir más allá de las fronteras del conocimiento establecido, al plantear y poner en práctica un enfoque de predicción de precios de acciones que integra el análisis de noticias periodísticas con técnicas avanzadas de aprendizaje automático.

El principal logro de la investigación es el desarrollo exitoso de un modelo capaz de predecir las fluctuaciones diarias en los precios de las acciones. Se trata de una hazaña relevante y significativa, que coloca este trabajo en la vanguardia de la intersección entre las finanzas cuantitativas y la inteligencia artificial.

Además, se ha demostrado a través de la implementación práctica del modelo y la posterior evaluación de los resultados, que el enfoque es no solo teóricamente sólido, sino también aplicable y valioso en un contexto real. A través de este trabajo, se ha aportado un ejemplo concreto y tangible de cómo las técnicas de aprendizaje automático y procesamiento del lenguaje natural pueden ser aplicadas en el campo de las finanzas.

Se puede afirmar que este estudio es pionero al introducir un enfoque integrado y matizado en el análisis de las finanzas utilizando el aprendizaje automático y el procesamiento del lenguaje natural. Esta contribución supone un avance considerable en el campo de las finanzas cuantitativas, al abrir nuevos caminos que hasta ahora no habían sido explorados en profundidad.

En la literatura existente sobre predicción de precios de acciones, los estudios anteriores se han centrado principalmente en la utilización de datos históricos de precios para predecir las fluctuaciones futuras. Aunque este enfoque proporciona resultados útiles, deja de lado factores cualitativos potencialmente relevantes, como el sentimiento del mercado reflejado en las noticias financieras.

Por otro lado, los pocos estudios que sí consideran el análisis de sentimientos basado en noticias, por lo general, se enfocan en las empresas más grandes y conocidas de Estados Unidos, como Apple, Microsoft, o Amazon [33], o se centran en índices como el S&P 500 o el Dow Jones [26][40]. Estos estudios, a pesar de su valiosa contribución, han dejado un vacío en términos de aplicación a mercados y sectores menos explorados, como el sector energético español.

Además, los pocos trabajos que incorporan noticias periodísticas como variables de entrada en sus modelos tienden a combinarlas con una serie de factores financieros y de precios históricos [52], lo que dificulta la identificación de la influencia directa de las noticias en la predicción. A menudo, las noticias seleccionadas para estos estudios son muy específicas y estrechamente relacionadas con la empresa o el sector bajo análisis, lo que limita su aplicabilidad y alcance.

Este TFG, por tanto, ofrece una valiosa aportación al incorporar el análisis de noticias generales en un modelo predictivo para el mercado bursátil español, un ámbito que ha sido poco explorado en la literatura previa. No solo eso, sino que este estudio ha priorizado el análisis de las noticias como el factor principal de la predicción, una estrategia distinta que proporciona una visión más amplia de la influencia del clima informativo en los movimientos del mercado. En definitiva, este trabajo amplía los límites del conocimiento existente en el campo de las finanzas cuantitativas y proporciona una nueva perspectiva que puede ser de gran utilidad para futuras investigaciones.

Más allá de su solidez técnica y precisión en la predicción de precios de acciones, el estudio económico del *Capítulo 6* reveló que el proyecto fintech propuesto posee una viabilidad económica significativa. Los cálculos detallados de CAPEX, OPEX y LCOX subrayaron la sostenibilidad financiera del proyecto, mientras que las proyecciones de ingresos y el cálculo del Valor Actual Neto (VAN) destacaron su potencial de rentabilidad atractivo. El resultado fue un VAN positivo y considerable, indicando que se espera que el proyecto genere más ingresos que costes a lo largo de su vida, reafirmando así la fortaleza del modelo propuesto no solo desde una perspectiva técnica, sino también económica.

En conclusión, este trabajo ha logrado su objetivo de desarrollar y poner en práctica un modelo de predicción de precios de acciones basado en el análisis de noticias, y ha demostrado su valor y eficacia a través de su implementación y evaluación. Este logro, combinado con la contribución a la literatura existente y la creación de un nuevo camino para futuras

investigaciones, marca este TFG como un valioso aporte al campo de las finanzas cuantitativas y la inteligencia artificial.

7.2 TRABAJOS FUTUROS

El desarrollo de este proyecto ha sentado una base sólida sobre la que se pueden construir trabajos futuros. Aunque el estudio ha obtenido resultados significativos y ha aportado contribuciones valiosas a la comprensión de la influencia de las noticias financieras en las variaciones de los precios de las acciones, existen numerosas vías de investigación abiertas que podrían explorarse para ampliar y profundizar los hallazgos actuales. En la *Figura 7.1* se muestra un esquema que indica las tres direcciones propuestas que podrían tomar futuros trabajos:

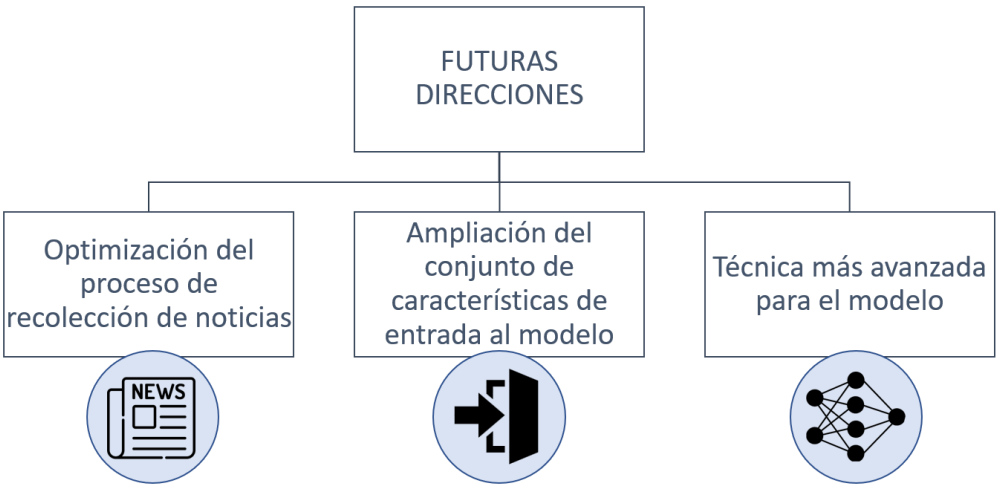

Figura 7.1. Recomendaciones de mejora para futuros trabajos (elaboración propia, 2023)

En primer lugar, la optimización del proceso de recolección de noticias se presenta como un avance prometedor para incrementar la eficacia del modelo. En este trabajo, se ha empleado un conjunto particular de noticias, recopiladas manualmente de los principales medios de comunicación españoles. Este enfoque, aunque efectivo, podría ser significativamente mejorado mediante la automatización del proceso de recolección de noticias. Se podría diseñar un sistema, por ejemplo, a través de un script de Python, que utilice una API para recopilar automáticamente las noticias relevantes de un día específico de diversas fuentes de noticias

[50], tanto nacionales como internacionales. Esta mejora permitiría el acceso a un volumen y diversidad de noticias mucho mayor, lo que resultaría en una representación más rica y matizada de los factores que inciden en las fluctuaciones del mercado. Así mismo, podría considerarse la utilización de una base de datos de noticias ampliamente extensa. De esta forma, el modelo tendría a su disposición una mayor cantidad de información y, en consecuencia, tendría un alcance mayor para identificar y aprender los patrones que influyen en los precios de las acciones.

Un segundo aspecto a considerar en futuros trabajos es la ampliación del conjunto de características de entrada del modelo. En la presente investigación, se ha hecho énfasis en el análisis de la frecuencia de aparición de las palabras en las noticias. Si bien este enfoque ha demostrado su utilidad, la inclusión de características adicionales podría proporcionar una representación más profunda del contenido de las noticias y, por lo tanto, mejorar la capacidad predictiva del modelo. Por ejemplo, una de estas características podría ser el análisis de sentimiento [13]. Esta técnica implica la evaluación de las connotaciones emocionales de las palabras utilizadas en las noticias, lo que podría proporcionar información valiosa sobre cómo los sentimientos expresados en las noticias pueden afectar las fluctuaciones del mercado.

Otras características del campo del procesamiento del lenguaje natural [36] podrían incluir la detección de entidades nombradas, que identificaría y clasificaría entidades específicas mencionadas en el texto (como personas, organizaciones o lugares), y el análisis de la semántica latente, que podría descubrir temas ocultos o subyacentes en el conjunto de noticias. También se podría explorar la inclusión de metadatos de las noticias, como la fuente de la noticia, la hora de publicación o el autor, que podrían tener un impacto en cómo las noticias son recibidas y, por lo tanto, en las reacciones del mercado.

Por último, el modelo de aprendizaje automático empleado podría ser objeto de revisión y mejora. Aunque el modelo basado en el procesamiento del lenguaje natural utilizado en este estudio ha demostrado ser eficaz, existen nuevas técnicas y enfoques emergentes que podrían proporcionar un mayor grado de precisión y eficacia.

En este sentido, una línea de trabajo futuro podría ser la exploración y adopción de técnicas más avanzadas o la combinación de varios enfoques de modelado. Por ejemplo, las técnicas de aprendizaje profundo, que son un subcampo del aprendizaje automático, se están convirtiendo cada vez más en la piedra angular de numerosas aplicaciones de inteligencia artificial. Estas

técnicas se basan en redes neuronales artificiales con múltiples capas ocultas, que son capaces de aprender de manera automática y generalizar características de los datos de entrada.

Dentro del aprendizaje profundo, las redes neuronales recurrentes (RNN), que han demostrado ser especialmente eficaces en el análisis de secuencias de datos, podrían ser de particular interés para este estudio. Las RNN son capaces de procesar secuencias de longitud variable [48], lo que las hace ideales para tareas relacionadas con el texto, como la interpretación de noticias. Incorporar un modelo de RNN en la arquitectura existente podría ofrecer una mejor capacidad para captar los matices y las estructuras subyacentes en los textos de las noticias.

Este Trabajo de Fin de Grado ha inaugurado un sendero innovador en el escenario de la predicción del mercado bursátil, proporcionando una nueva lente a través de la cual examinar y entender las fluctuaciones del mercado. En el entramado de variables que mueven las palancas del mercado financiero, el análisis de las noticias, hasta ahora un terreno menos explorado, ha demostrado ser una fuente de conocimiento y percepción valiosa.

Los resultados alentadores obtenidos en este estudio no sólo ratifican la viabilidad de combinar el procesamiento del lenguaje natural y el aprendizaje automático para predecir el mercado, sino que también plantean una serie de preguntas intrigantes y apuntan a direcciones prometedoras para investigaciones futuras.

Sin embargo, no debemos olvidar que este es sólo el comienzo. La visión que hemos desvelado y las metas que hemos alcanzado en este trabajo representan un solo peldaño en una escalera que se extiende hacia una multitud de posibilidades aún por explorar. Por tanto, si bien debemos celebrar los logros de este estudio, también debemos verlos como un estímulo, una señal de partida para la exploración de nuevos horizontes.

Este proyecto es mucho más que una conclusión: es una invitación al asombro, a la curiosidad y al deseo incansable de aprender y mejorar. Es un recordatorio de que, a pesar de lo lejos que hemos llegado, aún queda un vasto mar de conocimiento por descubrir. En la confluencia de la tecnología y la economía, espero que este trabajo sirva como un instrumento de orientación, señalando el camino hacia el futuro de la predicción del mercado bursátil.

REFERENCIAS

[1] Afkhami, M., Cormack, L. y Ghoddusi, H. (2017). *Google search keywords that best predict energy price volatility*, Energy Economics, vol. 67, pp.17-27.

[2] Aicart Pauner, M. (2022). Red neuronal para la elección de los parámetros de comprensión óptimos.

[3] Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation* (pp. 106-112). IEEE.

[4] Banerjee, R. y Gupta, K. (2019). Does OPEC news sentiment influence stock returns of energy firms in the United States?, Energy Economics, vol. 77, pp 34-45.

[5] Burgess, M., Javed, F., Okpara, N., & Robinson, C. (2022). Stock Forecasts with LSTM and Web Sentiment. *SMU Data Science Review*, *6*(2), 10.

[6] Butler, K. C., & Malaikah, S. J. (1992). Efficiency and inefficiency in thinly traded stock markets: Kuwait and Saudi Arabia. *Journal of Banking & Finance*, *16*(1), 197-210.

[7] Chandrasekaran, A., Elliot, B., & Rigon, G. (2022). Cool Vendors in Natural Language Technology for Processing Enormous Volumes of Unstructured Data. Gartner. https://www.gartner.com

[8] Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert Systems with Applications*, *80*, 340-355.

[9] Cheng, C. H., Chen, T. L., & Wei, L. Y. (2010). A hybrid model based on rough sets theory and genetic algorithms for stock price forecasting. *Information Sciences*, *180*(9), 1610-1629.

[10] Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science & Technology*, *9*(2), 14-14.

[11] Constantino, M., Naranpanawa, A., Paes Herrera, G., Su, J. (2022). *Renewable energy stocks forecast using Twitter investor sentiment and deep learning*, Energy Economics, vol. 114.

[12] Drachal, K. (2021). Forecasting selected energy commodities prices with Bayesian dynamic finite mixtures, Energy Economics, vol. 99.

[13] Dridi, A., Atzeni, M., & Reforgiato Recupero, D. (2019). FineNews: fine-grained semantic sentiment analysis on financial microblogs and news. *International Journal of Machine Learning and Cybernetics*, *10*, 2199-2207.

[14] Elshendy, M., Colladon, A. F., Battistoni, E., y Gloor, P. A. (2018). *Using four different online media sources to forecast the crude oil price*, Journal of Information Science, 44(3), 408–421.

[15] Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, *38*(1), 34-105.

[16] Ferreira, F. G., Gandomi, A. H., & Cardoso, R. T. (2021). Artificial intelligence applied to stock market trading: a review. *IEEE Access*, *9*, 30898-30917.

[17] Ferreira S., Karali, B. y Liu, H. (2021). *Hurricanes as news? Assessing the impact of hurricanes on the stock market returns of energy companies*, International Journal of Disaster Risk Reduction, vol. 66.

[18] Freedman, D. A. (2009). *Statistical models: theory and practice*, Cambridge University Press.

[19] Gallagher, L. A., & Taylor, M. P. (2002). Permanent and temporary components of stock prices: Evidence from assessing macroeconomic shocks. *Southern Economic Journal*, *69*(2), 345-362.

[20] García-Quintero, E., Muñoz, N. y Villada, F. (2016). *Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro*, Información tecnológica, 27(5), 143-150.

[21] Gaure, S. (2013). OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis*, *66*, 8-18.

[22] Hug, R., Hübner, W., & Arens, M. (2020, April). Introducing probabilistic bézier curves for n-step sequence prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 06, pp. 10162-10169).

[23] Karsoliya, S. (2012). Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *International Journal of Engineering Trends and Technology*, *3*(6), 714-717.

[24] Kavussanos, M. G., & Dockery, E. (2001). A multivariate test for stock market efficiency: the case of ASE. *Applied Financial Economics*, *11*(5), 573-579.

[25] Kim, E.H. y Youm, Y.N. (2017). How Do Social Media Affect Analyst Stock Recommendations? Evidence from S&P 500 Electric Power Companies' Twitter Accounts, Strat. Mgmt. J, 38: 2599-2622.

[26] Koenigstein, N. (2022). Dynamic and context-dependent stock price prediction using attention modules and news sentiment. *arXiv preprint arXiv:2205.01639*.

[27] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

[28] Linares Llamas, P. (2009). Eficiencia energética y medio ambiente.

[29] Liu, C. (2021). *COVID-19 and the energy stock market: evidence from China*, Energy RESEARCH LETTERS, vol. 2, issue 3.

[30] Marsland, S. (2015). *Machine learning: an algorithmic perspective*. CRC press.

[31] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

[32] Moro-Visconti, R., Cruz Rambaud, S., & López Pascual, J. (2020). Sustainability in FinTechs: An explanation through business model scalability and market valuation. *Sustainability*, *12*(24), 10316.

[33] Muthivhi, M., & van Zyl, T. L. (2022, July). Fusion of sentiment and asset price predictions for portfolio optimization. In *2022 25th International Conference on Information Fusion (FUSION)* (pp. 1-8). IEEE.

[34] Muthukrishnan, R., & Rohini, R. (2016, October). LASSO: A feature selection technique in predictive modeling for machine learning. In *2016 IEEE international conference on advances in computer applications (ICACA)* (pp. 18-20). IEEE.

[35] Naciones Unidas (2015), *La Agenda 2030 y los Objetivos de Desarrollo Sostenible.*

[36] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551.

[37] Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25, pp. 15-24). San Francisco, CA, USA: Determination press.

[38] Núñez-Godínez, L. R. (2020). Estudio de la viabilidad financiera para una empresa Fintech.

[39] Nusrat, I., & Jang, S. B. (2018). A comparison of regularization techniques in deep neural networks. *Symmetry*, *10*(11), 648.

[40] Ortiz Rubio, A. (2019). Analisis del efecto de la información financiera y no financiera en el precio de cotizacion de las empresas del Dow Jones, en el primer cuatrimestre de 2018.

[41] Panday, H., Vijayarajan, V., Mahendran, A., Krishnamoorthy, A., & Prasath, V. B. (2020). Stock prediction using sentiment analysis and long short term memory. *European Journal of Molecular & Clinical Medicine*, *7*(2), 5060-5069.

[42] Pant, D. R., Neupane, P., Poudel, A., Pokhrel, A. K., & Lama, B. K. (2018, October). Recurrent neural network based bitcoin price prediction by twitter sentiment analysis. In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)* (pp. 128-132). IEEE.

[43] Ranjit, S., Shrestha, S., Subedi, S., & Shakya, S. (2018, October). Foreign rate exchange prediction using neural network and sentiment analysis. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 1173-1177). IEEE.

[44] Teoh, E. J., Tan, K. C., & Xiang, C. (2006). Estimating the number of hidden neurons in a feedforward network using the singular value decomposition. *IEEE Transactions on Neural Networks*, *17*(6), 1623-1629.

[45] Termes, R. (1997). Inversión y coste de capital: manual de finanzas.

[46] Thakur, N., & Han, C. Y. (2021). A study of fall detection in assisted living: Identifying and improving the optimal machine learning method. *Journal of sensor and actuator networks*, *10*(3), 39.

[47] Ticknor J. L. (2013). A Bayesian regularized artificial neural network for stock market forecasting, Expert Systems with Applications, vol. 40, issue 14, pp 5501-5506.

[48] Wang, J., & Wang, J. (2016). Forecasting energy market indices with recurrent neural networks: Case study of crude oil price fluctuations. *Energy*, *102*, 365-374

[49] Yilmaz, E.S., Ozpolat, A. y Destek, M.A. (2022). Do Twitter sentiments really effective on energy stocks? Evidence from the intercompany dependency, Environ Sci Pollut Res 29, 78757–78767.

[50] Yukka Lab. (n.d.). Inicio. Recuperado el 1 de julio de 2023, de https://www.yukkalab.com/

[51] Zaman S., Yaqub U. y Saleem T. (2022). *Analysis of Bitcoin's price spike in context of Elon Musk's Twitter activity*, Global Knowledge, Memory and Communication.

[52] Zhao, B., He, Y., Yuan, C., & Huang, Y. (2016, July). Stock market prediction exploiting microblog sentiment analysis. In *2016 International Joint Conference on Neural Networks (IJCNN)* (pp. 4482-4488). IEEE.

[53] Zhou, X. Y., Lu, G., Xu, Z., Yan, X., Khu, S. T., Yang, J., & Zhao, J. (2023). Influence of Russia-Ukraine war on the global energy and food security. *Resources, Conservation and Recycling*, *188*, 106657.

ANEXO I

Se adjunta el script de Python que fue desarrollado y utilizado para llevar a cabo el preprocesamiento de las noticias:

```
import pandas as pd
import nltk
from collections import Counter
from nltk.corpus import stopwords
from itertools import combinations

# Descargar los recursos necesarios de NLTK
nltk.download('punkt')
nltk.download('stopwords')

# Cargar datos del archivo Excel en un DataFrame de pandas
df = pd.read_excel('Noticias Endesa bajadas.xlsx')

# Fecha específica que deseas analizar (en formato DD.MM.AA)
fecha_especifica = '23.11.2022'

# Filtrar el DataFrame para obtener solo la fecha específica
df_fecha = df[df['FECHA'] == fecha_especifica]

# Limpiar y normalizar el texto
stop_words = set(stopwords.words('spanish'))

def clean_text(text):
    if pd.isnull(text):  # Verificar si el valor es nulo
        return []
    tokens = nltk.word_tokenize(str(text).lower())  # Convertir a cadena y aplicar
lower()
    tokens = [token for token in tokens if token.isalpha() and token not in
stop_words]
    return tokens

# Procesar el texto de las noticias
df_fecha['texto_procesado'] = df_fecha['TÍTULO'].apply(clean_text) +
df_fecha['SUBTÍTULO'].apply(clean_text)

# Contar la frecuencia de aparición de las palabras
word_counts = Counter([word for sublist in df_fecha['texto_procesado'] for word in
sublist])

# Ordenar las palabras por frecuencia de aparición
top_words = word_counts.most_common()

# Filtrar palabras y seleccionar las más relevantes
relevant_words = [word for word, count in word_counts.items() if count > 5 and
len(word) > 1]

# Realizar análisis de co-ocurrencia de palabras
coocurrence_counts = Counter()
```

```
for tokens in df_fecha['texto_procesado']:
    for r in range(2, min(len(tokens) + 1, 4)):  # Considerar co-ocurrencias de 3
palabras
        for combination in combinations(tokens, r):
            if all(word in relevant_words for word in combination):
                coocurrence_counts[combination] += 1

# Ordenar las co-ocurrencias por frecuencia
top_coocurrences = coocurrence_counts.most_common()

# Imprimir las co-ocurrencias más frecuentes
for combination, count in top_coocurrences:
    print(f"{combination}: {count} veces")
```

ANEXO II

Se adjuntan todos los archivos Excel utilizados en la construcción y aplicación del modelo:

DATOS HISTÓRICOS DE PRECIOS:

..\EXCEL Datos Históricos\Datos históricos IBEX 35.xlsx

..\EXCEL Datos Históricos\Datos históricos IBEX 35 Energy.xlsx

..\EXCEL Datos Históricos\Datos históricos ELE (Endesa).xlsx

NOTICIAS RECOPILADAS:

..\EXCEL Datos Históricos\Noticias IBEX 35 subidas.xlsx

..\EXCEL Datos Históricos\Noticias IBEX 35 bajadas.xlsx

..\EXCEL Datos Históricos\Noticias IBEX-Energy subidas.xlsx

..\EXCEL Datos Históricos\Noticias IBEX-Energy bajadas.xlsx

..\EXCEL Datos Históricos\Noticias Endesa subidas.xlsx

..\EXCEL Datos Históricos\Noticias Endesa bajadas.xlsx

RESULTADOS:

..\EXCEL Datos Históricos\RESULTADOS (por fecha).xlsx

..\EXCEL Datos Históricos\Resultados (gráficos).xlsx