Universidad Iberoamericana

Departamento de Economía

**Finanzas y datos no estructurados: Una aproximación con inteligencia artificial**

Tesina presentada por:Alejandro González Paredes

Asesor: Dr. Isidro Soloaga

Ciudad de México, mayo de 2025

**Abstract**

En un entorno financiero caracterizado por alta volatilidad e inestabilidad política, este trabajo explora el valor predictivo de los datos no estructurados, específicamente el sentimiento expresado en redes sociales para anticipar movimientos en el precio de Bitcoin. Se sugiere un modelo híbrido que combina puntuaciones de sentimiento obtenidas con herramientas de procesamiento de lenguaje natural (NLP) como VADER y BERTweet, utilizando la API de Twitter, junto con una red neuronal LSTM que se ha entrenado con datos de precios a lo largo del tiempo. El estudio demuestra que la combinación de datos estructurados (precios históricos) y no estructurados (análisis de sentimeinto) mejora la precisión del modelo predictivo, alcanzando una tasa deonde los resultados muestran que agregar variables de sentimiento de redes sociales mejora las predicciones en ciertos tiempos, especialmente cuando se usan en una arquitectura LSTM bien ajustada. El modelo obtuvo una **precisión del 86 %**, un **RMSE de 633.25 USD**, un **MAE de 593.75 USD**, un **MAPE de 6.96 %** y un **SMAPE de 7.25 %.** Además, anticipa correctamente la **dirección del cambio** en el **66.67 %** de los casos. Igualmente, el modelo ARIMA obtuvo un RMSE de 5,364 USD, un MAPE del 19.06 % y una precisión de 80.94 %. Sin embargo, el modelo LSTM superó al ARIMA en todas las métricas. El enfoque sugerido permite el uso de estrategias de inversión que se basan en el análisis de sentimiento. Esto resalta la importancia de usar técnicas de inteligencia artificial para entender relaciones no lineales en los mercados financieros, lo que mejora la toma de decisiones.

**Introducción**

El análisis financiero ha experimentado una transformación significativa gracias al desarrollo de herramientas de procesamiento del lenguaje natural. Estas técnicas permiten convertir grandes volúmenes de datos no estructurados, como texto proveniente de redes sociales, noticias y documentos financieros, en información capaz de revelar percepciones, intenciones y emociones que influyen en el comportamiento de los mercados. A diferencia de los datos numéricos tradicionales, el lenguaje humano contiene múltiples capas de información como la semántica, la sintáctica y la pragmática. Captar estas dimensiones permite comprender mejor la psicología colectiva de los agentes económicos y detectar señales tempranas de eventos relevantes, como aumentos en la volatilidad o cambios en las expectativas del mercado.Debido a esto, surge la necesidad de desarrollar modelos predictivos que integren estas fuentes cualitativas de información, porque la combinación de técnicas de NLP y algoritmos de aprendizaje automático permite construir herramientas más sensibles a los datos no estructurados, ya que inciden directamente en el comportamiento de los activos financieros.

Esta investigación utiliza este enfoque para estudiar cómo el NLP puede ser una herramienta útil para hacer predicciones en situaciones financieras muy cambiantes. Se elige el mercado de Bitcoin como ejemplo debido a su respuesta a factores como el sentimiento en redes sociales, la capitalización de mercado y su gran volatilidad. Los avances en redes neuronales profundas, como los *Transformers* (BERT, GPT) y las redes LSTM, han revolucionado el análisis textual aplicado a las finanzas. Capturan relaciones contextuales complejas entre palabras. Las LSTM son eficaces para modelar dependencias temporales en series históricas como el precio de un activo. De manera similar, Mittal y Goel (2009) implementaron un enfoque basado en redes neuronales difusas: “We train a Self-Organizing Fuzzy Neural Network (SOFNN) using 3-day lagged values of the DJIA and the mood dimensions to predict the 4th day’s DJIA value” [se entreno una red neuronal usando un rezago de 3 días sobre los puntos del DJIA y se agregraon emociones para predecir los puntos del cuarto día] (p. 4), lo cual les permitió incorporar tanto variables históricas como estados de ánimo derivados de Twitter para predecir el mercado bursátil, por otro lado, Kim et al. (2018) aplicaron modelos LSTM al análisis del precio de Bitcoin aunque, sin incorporar señales provenientes del análisis de sentimiento; una limitación que este trabajo busca superar.

En este contexto, el análisis de sentimiento se ha consolidado como una herramienta clave para cuantificar emociones expresadas en textos financieros. Mittal y Goel (2009) destacan esta línea al afirmar que “we perform sentiment analysis on publicly available Twitter data to find the public mood and the correlation with market sentiment” (p. 1). Además, a través de su modelo predictivo, demostraron que “the Calm mood dimension from Twitter is a statistically significant predictor of the DJIA […] the inclusion of Calm time series in the input significantly improves the prediction accuracy of DJIA” (pp.5-6). Por lo tanto, estas pruebas apoyan la utilidad del análisis de sentimiento como una herramienta para aumentar la precisión de los modelos financieros actuales, especialmente en situaciones donde la información llega de inmediato y las reacciones del mercado son muy sensibles. La hipótesis central de esta investigación es que las emociones expresadas en redes sociales actúan como indicadores adelantados del comportamiento del mercado, complementando a los datos cuantitativos como precios históricos y volumen de transacciones. Este enfoque permite superar las limitaciones de modelos tradicionales como ARIMA o la regresión lineal porque captura patrones no lineales y dinámicas contextuales complejas. Un ejemplo de esto es el estudio de Sheetal y Aithal (2024). Ellos lograron una precisión del 90 % en el análisis de sentimiento financiero al utilizar modelos basados en BERT.

Por otra parte, en la era digital caracterizada por la abundancia de datos y la aceleración de los flujos informativos, el análisis financiero enfrenta el reto de integrar nuevas fuentes de información para mejorar la comprensión y predicción de los mercadosTradicionalmente, los modelos que predicen resultados se han basado en datos organizados como precios pasados, volúmenes de transacciones o indicadores económicos. Sin embargo, esta forma de ver las cosas ha mostrado ser insuficiente ante situaciones complejas que son muy sensibles al estado emocional y social, especialmente en el caso de Bitcoin. Como resultado, surge una pregunta central que guía este trabajo: ¿Es posible que el uso de datos no estructurados pueda mejorar los modelos predictivos financieros a través del uso de LSTM? La investigación sugiere que analizar datos no estructurados ayuda a encontrar patrones importantes que los métodos tradicionales con datos estructurados no pueden detectar. Esto produce información valiosa para entender dinámicas complejas, como el comportamiento de los mercados financieros. Esta afirmación se apoya en resultados empíricos contundentes, como los de Mittal y Goel (2009), quienes reportan: “the best accuracy of 86.7% when using a combination of historical DJIA values and the Calm mood dimension” [al combinar los valores históricos del DJIA con el sentimeinto de calma se obtuvo una precisión del 86.7% ](p. 6). Esta propuesta se sustenta en cuatro argumentos principales:

En primer lugar, la proliferación de datos no estructurados representa una oportunidad estratégica para el análisis financiero, ya que actualmente la mayor parte de la información generada en el mundo adopta formas no estructuradas, como textos, imágenes, audios o videos. Según García-Díaz, J. C., López, J. R. & Arias, R. (2019): “El análisis de datos no estructurados representa uno de los mayores desafíos y oportunidades en la era del big data, especialmente en el contexto de las redes sociales.” (pp. 747–748).

En segundo lugar, las redes neuronales del tipo Long Short-Term Memory (LSTM) están diseñadas para procesar secuencias largas, lo cual permite conservar información contextual crítica para el análisis de sentimientos. Este aspecto es crucial porque el significado de una frase depende no solo de las palabras individuales, sino también del tono general del texto. Zhang et al. (2018) señalan que las LSTM tienen una capacidad superior para captar dependencias lingüísticas a largo plazo, lo que mejora la interpretación emocional del contenido digital.

En tercer lugar, existe evidencia empírica de que el análisis de sentimiento puede anticipar movimientos financieros. Bollen, Mao y Zeng (2011) mostraron que los cambios en el estado emocional colectivo expresado en Twitter pueden predecir variaciones en el mercado bursátil. Esto sugiere que los patrones emocionales, cuando son correctamente identificados, pueden convertirse en herramientas efectivas para el análisis de riesgo, la identificación de tendencias y la toma de decisiones de inversión.

Por último, las organizaciones que desarrollan capacidades analíticas para procesar datos no estructurados tienden a adaptarse con mayor agilidad a los cambios del entorno. Martínez-López, F. J., & Casillas, J. (2013) nos dicen que “Las organizaciones que desarrollan capacidades avanzadas para el análisis de datos no estructurados obtienen una ventaja competitiva significativa frente a sus competidores, permitiéndoles anticipar tendencias del mercado y responder ágilmente a las necesidades cambiantes de los consumidores” (p. 372), lo cual es clave en un entorno económico tan competitivo como el de hoy en un mercado globalizado y sumamente complejo.

**Marco teórico**

Con relación a la era de la información, algunas empresas están preparadas para este cambio, como IBM (2017). “According to projections from IDC, 80% of worldwide data will be unstructured by 2025” [según el IDC el 80% de los datos serán no estrucutrados para 2025] (p. 1), así pues, a pesar de su riqueza informativa, los datos no estructurados históricamente han sido subutilizados debido a su naturaleza desorganizada y a la falta de herramientas adecuadas para su análisis. De modo que, con el desarrollo de nuevas tecnologías, en particular la inteligencia artificial, el aprendizaje automático y el procesamiento del lenguaje natural, ahora es posible extraer valor significativo de este tipo de datos.

Otra diferencia es que los datos estructurados ofrecen métricas rígidas y cuantificables, mientras que los datos no estructurados proporcionan contexto, matices y señales emocionales que enriquecen enormemente los modelos predictivos. En otras palabras, nos permiten analizar un problema desde otro ángulo y complementarlo con la información estructurada existente. Por lo tanto, su uso es una ventaja estratégica para tomar decisiones. En cuanto al ámbito financiero, los tweets, publicaciones en foros, artículos de noticias y los reportes financieros representan una fuente dinámica de información sobre el sentimiento del mercado. En particular, este tipo de análisis permite identificar señales tempranas sobre movimientos de activos, detectar rumores o anticipar crisis antes de que estos eventos se reflejen en las métricas tradicionales. Las opiniones de los inversionistas minoristas y grandes actores financieros se difunden ampliamente en redes sociales, plataformas de video y directos. Esto convierte a Twitter en una mina de oro para comprender la percepción colectiva del mercado en tiempo real.

Por otra parte, el uso de herramientas de NLP como VADER y TextBlob es fundamental para procesar textos informales y breves, como los de redes sociales, con gran efectividad. Estas herramientas analizan el tono emocional de las palabras y asignan puntuaciones de polaridad que clasifican los mensajes en positivos, negativos o neutrales. Por ejemplo, una frase como “la acción está sobrevalorada y caerá pronto” puede ser automáticamente etiquetada como negativa, esto contribuye a construir indicadores de sentimiento que alimentan modelos de predicción de precios, asimismo, la ventaja radica en que este procesamiento puede realizarse de forma automática y a gran escala, ya que permiten a los analistas y a los sistemas algorítmicos incorporar percepciones humanas de forma cuantificable en sus estrategias, por lo cual, su correcta integración en modelos predictivos permite a las organizaciones en sectores dinámicos como las finanzas adelantarse a los eventos, captar señales débiles y adaptarse con mayor agilidad al entorno. Por consiguiente, ignorar este tipo de información equivale a operar con una visión incompleta del mercado, y aprovecharla abre nuevas fronteras para la predicción y la toma de decisiones basadas en datos.

Por otro lado, las redes neuronales artificiales son modelos computacionales inspirados en la estructura y funcionamiento del cerebro humano. Dado que están formadas por capas de nodos interconectados que procesan información mediante funciones de activación y pesos ajustables, según Goodfellow, Bengio y Courville (2016): “A neural network is a function composed of many simple functions, arranged in layers, and trained to approximate a complex target function” [una red neuronal es una función compuesta por multiples capas jerarquícas y se puede entrenar para aproximarse a una función objetivo compleja] (p. 168). Como resultado, son especialmente efectivas para identificar patrones no lineales en extensos volúmenes de datos. Además, una de las principales ventajas de las redes neuronales es su capacidad para aprender y generalizar a partir de los datos dicho de otra manera, una vez entrenadas pueden hacer predicciones sobre nuevos casos con alta precisión, incluso cuando las relaciones entre las variables no son evidentes; esto lo demuestra Schmidhuber, J. (2015) cuando señala: “Deep neural networks have outperformed many traditional methods in tasks such as classification, prediction, and pattern recognition, especially in contexts where data is abundant and complex” [las redes neuronales profundas demostraron mejor rendimiento que lo metódos tradicionales en tareas de clasificación, predicción, para reconocer patrones y específicamente en contextos donde los datos son complejos y en gran volumén] (p. 89), o sea, en la predicción de precios basados en datos históricos.

Así pues, este trabajo utiliza las redes LSTM, la variante avanzada de las redes neuronales recurrentes (RNN), las cuales están diseñadas específicamente para manejar dependencias a largo plazo en series temporales, para Hochreiter, S., & Schmidhuber, J. (1997). “LSTMs were explicitly designed to avoid the long-term dependency problem” [las LTSMs estan diseñadas para mitigar el problema del desvanecimiento del gradiente](p. 1736). Por lo cual, les aprender patrones complicados en datos secuenciales como precios pasados o texto. En el análisis financiero, las LSTM han demostrado ser efectivas para captar relaciones no lineales entre variables en el tiempo, como dicen Zhang, Y., Jin, R. & Zhou, Z.-H. (2018). “LSTM networks have shown superior ability to capture long-term dependencies in sequential data [las redes LSTM han demostrado mejor habilidad para recoletar dependencias temporales a largo plazo en datos secuenciales](p. 187). Su arquitectura incluye celdas de memoria que almacenan información relevante durante múltiples iteraciones, por lo cual resultan ideales para combinar datos numéricos con variables cualitativas. Es posible que el análisis de datos no estructurados sea un recurso valioso para la predicción financiera. Asimismo, derivado de una extensa revisión de literatura, se identificó un consenso creciente sobre el poder informativo del sentimiento del mercado. Por ejemplo, Bollen, J., Mao, H. & Zeng, X. (2011). “We find that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions” [observamos que la presición de los valores del DJIA pueden mejorar significativamente al incluir el sentiminento público](p. 6), mientras que Sheetal y Aithal (2024) combinaron diccionarios financieros con BERT para clasificar emociones en reportes corporativos con una precisión del 90 %. Conviene subrayar que las percepciones, emociones y narrativas compartidas en medios digitales contienen información relevante para los mercados; sin embargo, muchas investigaciones previas se centraron en textos extensos como noticias, reportes financieros o bien en series de precios sin incorporar variables cualitativas. No obstante, este trabajo va más allá porque combina puntuaciones de sentimiento extraídas de más de 1.2 millones de tweets con una red LSTM que se entrenó sobre precios históricos de Bitcoin. Esto significa que, a diferencia de los modelos tradicionales ARIMA o las regresiones lineales, la arquitectura LSTM es perfecta para este trabajo, como señalan Zhang, Y., Jin, R. & Zhou, Z.-H. (2018), ya que su capacidad para retener información contextual a lo largo del tiempo es clave para modelar variables como el sentimiento.

Además de los resultados técnicos, este trabajo apoya la idea de que los datos no organizados, procesados con herramientas de NLP y aprendizaje profundo, pueden revelar aspectos que los datos organizados no pueden mostrar por sí solos. En concreto, datos en forma de tweets, emojis, expresiones espontáneas y reacciones colectivas constituyen señales tempranas de eventos disruptivos o cambios en el sentimiento inversionista. Este enfoque complementa el análisis cuantitativo tradicional con una perspectiva más emocional y contextual, por lo cual hay que usarlo. Al mismo tiempo, este modelo abre la puerta a nuevas estrategias de inversión basadas en análisis de sentimiento. Dependiendo de la polaridad predominante en redes sociales, se puede establecer un sistema de señales para decidir la compra, venta o mantenimiento de un activo. Estas estrategias pueden ser evaluadas mediante técnicas de backtesting, de manera que convierte al análisis de sentimiento en una herramienta no solo descriptiva, sino también operativa. A su vez, Las redes neuronales presentan desafíos en la interpretación de los resultados debido a su complejidad estructural. Sin embargo, su capacidad para capturar relaciones complejas y predecir movimientos del mercado justifica su uso. Por lo cual, son parte de un enfoque integral para los modelos predictivos modernos.

**Metodología**

A continuación, la metodología propuesta se estructura en cinco fases principales: la recopilación de datos textuales y numéricos, el preprocesamiento, la integración de variables, el entrenamiento del modelo predictivo basado en redes neuronales LSTM y la evaluación del rendimiento del modelo. Creo que el enfoque propuesto utiliza bien el análisis de sentimiento al mezclarlo con la capacidad predictiva de las redes neuronales recurrentes, especialmente las Long Short-Term Memory (LSTM), que son esenciales para prever el precio de Bitcoin y para verificar la hipótesis.

Para empezar, se utilizó la API oficial de Twitter para recolectar datos. Junto con la biblioteca Tweepy, esta API permite automatizar la extracción de tweets bajo ciertos criterios. Se consideró un período de interés desde 2019 hasta 2023 y se filtraron únicamente tweets para asegurar la coherencia lingüística en el análisis de sentimiento. Además, se configuraron búsquedas que incluyeran términos clave como #Bitcoin, #BTC y Bitcoin, centrando la recolección en publicaciones altamente relevantes para el estudio. Como resultado, el conjunto de datos contiene un total de 1,263,274 tweets relacionados con Bitcoin, es decir, cada observación incluye la fecha, el texto del tweet y una etiqueta de sentimiento. En otras palabras, el conjunto de datos está prácticamente completo, ya que no hay valores faltantes en las columnas de fecha ni de texto, y solo se detectó un valor nulo en la columna de sentimiento y después fue removido. En el análisis de sentimientos, el 62.8% de los tweets fueron clasificados como positivos (793,424 tweets) y el 37.2% como negativos (469,849 tweets). Esto indica una tendencia general a expresar opiniones positivas sobre bitcoin en este conjunto de datos durante el periodo observado. Respecto a la longitud de los tweets, en promedio cada tweet contiene alrededor de 143 caracteres, con una desviación estándar de 82 caracteres. Esto sugiere una variabilidad significativa en la extensión de los mensajes. El tweet más corto tiene apenas 1 carácter, mientras que el más largo alcanza los 947 y la mediana se encuentra en 127 caracteres, o sea que más de la mitad de los tweets tienen una longitud relativamente moderada. Después de recolectar los tweets, se aplicaron técnicas de limpieza y normalización textual para eliminar ruido y preparar los datos para el análisis. En otras palabras, se eliminaron enlaces URL, menciones a otros usuarios (por ejemplo, @nombre), hashtags redundantes (como #Bitcoin) y caracteres especiales que no aportan valor semántico. Igualemnte, el texto se normalizó convirtiéndolo a minúsculas para evitar duplicidades y se procedió a la tokenización, es decir, dividir los textos en palabras individuales o subconjuntos de palabras, también se eliminaron las palabras vacías (stopwords) o aquellas que tienen poca carga informativa como: “el”, “la”, “y”, ya que esto reduce la dimensionalidad del texto y conserva las palabras con mayor valor analítico por lo cual, es necesario usar la biblioteca NLTK en python. Luego, se evaluó el tono emocional de los tweets con el modelo VADER (*Valence Aware Dictionary and sEntiment Reasoner*). Este enfoque léxico está diseñado para analizar texto de redes sociales porque asigna a cada tweet una puntuación compuesta entre -1 y 1, lo que nos permite clasificar su contenido emociona. En otras palabras, al aplicar VADER a los tweets recopilados, se obtuvo una métrica cuantitativa del sentimiento expresado por los usuarios de Twitter respecto a Bitcoin en diferentes momentos del tiempo. En resumen, estas puntuaciones se combinaron con los precios históricos diarios de Bitcoin después de ser extraídos de YahooFinance con su API yfinance.

Primero, estos valores se alinearon cronológicamente con los datos de sentimiento y las series de precios. Este proceso permitió construir una base de datos híbrida donde cada observación incluye una medida de sentimiento y el precio correspondiente. Esto sentó las bases para la predicción con redes LSTM. Después, una vez integrados los datos de sentimiento y precios, se procedió al entrenamiento del modelo de predicción con redes neuronales del tipo LSTM (*Long Short-Term Memory*), porque este tipo de red es particularmente útil en contextos donde la variable evoluciona en el tiempo, como sucede con el precio del Bitcoin, ya que este depende en parte de valores pasados y de patrones no lineales, por lo cual la base de datos combinada se dividió en conjuntos de entrenamiento y prueba debido a que es importante mantener el orden cronológico para evitar sesgos de información futura. Por consiguiente, las variables de entrada incluyeron tanto las puntuaciones de sentimiento diarias como los precios pasados, mientras que la variable objetivo fue el precio futuro de Bitcoin. Durante el entrenamiento se normalizaron las variables numéricas dado que, para mejorar la eficiencia del algoritmo, se deben normalizar los valores y definir los hiperparámetros clave del modelo, concretamente: el número de capas LSTM, el tamaño de las unidades de memoria, la tasa de aprendizaje, el número de épocas y el tamaño de lote (*batch size*), dado que el precio de Bitcoin presenta una alta volatilidad y una escala numérica considerablemente amplia. Se aplicó un proceso de normalización mediante la técnica *Min-Max Scaling*, que transforma los valores originales al rango [0, 1]. Esta transformación es muy importante en el uso de redes neuronales recurrentes como LSTM, ya que ayuda a que el entrenamiento del modelo sea más estable, disminuye la variación numérica entre las observaciones y aumenta la eficiencia en la convergencia. Además, al operar en una escala uniforme, se evita que las magnitudes absolutas del precio dominen el aprendizaje del modelo. Esto permite que las redes neuronales aprendan patrones estructurales y dinámicos más fácilmente. La normalización también facilita la comparación entre diferentes periodos de tiempo, ya que estandariza las escalas sin perder la forma de la serie temporal. Esto es fundamental para que el modelo pueda generalizar su comportamiento en distintos contextos históricos. Por consiguiente, para capturar estos patrones cíclicos en el comportamiento del precio de Bitcoin se implementó un modelo **SARIMA (S*easonal AutoRegressive Integrated Moving Average*)**, el cual extiende la formulación tradicional ARIMA al incorporar efectos estacionales explícitos ya que, esta extensión resulta particularmente útil cuando la serie temporal exhibe repeticiones sistemáticas como, los precios mensuales de activos financieros. Ademas, la notación **SARIMA(p, d, q)(P, D, Q)[s]** representa los componentes no estacionales (autorregresivos, diferenciales y de media móvil) como los estacionales donde **s** denota la periodicidad. Por tanto, la ecuación general del modelo se expresa así:  
 (1)

En este estudio, se utilizó un **SARIMA(1,1,1)(1,1,1)[12]** como se puede ver en la ecuación (1), este modelo es adecuado para series mensuales con tendencia y ciclos anuales de modo que, esta especificación implica una diferenciación regular y una estacional junto con términos AR y MA de primer orden. Este modelo fue seleccionado tras un proceso de optimización del criterio AIC, y resultó apto para representar la dinámica del mercado de Bitcoin debio a sus ciclos anuales, comportamiento especulativo y una tendencia creciente de largo plazo. No obstante, su capacidad predictiva fue limitada frente a arquitecturas más complejas como las redes neuronales, especialmente en presencia de no linealidades y efectos no observados, por otro lado, se implementó una red neuronal tipo **(LSTM)** la cual es adecuada para modelar secuencias temporales con dependencias de largo plazo. En este caso, la red recibe como entrada un vector que combina el precio de cierre del Bitcoin con el puntaje de sentimiento extraído de twitter mediante análisis textual, como se demuestra en la ecuación (2). A su vez, la dinámica interna de la celda LSTM se describe con el siguiente conjunto de ecuaciones resultando en la ecuación 3

(2)

(3)

Como demuestra la ecuación (3) arquitectura LSTM utilizada en este estudio incorpora tanto variables estructuradas como no estructuradas para predecir el comportamiento futuro del precio del Bitcoin, en concreto, la red toma como entrada un vector compuesto por el precio del Bitcoin en el tiempo y una puntuación de análisis de sentimiento derivada de twitter por lo cual, esta formulación permite que la red LSTM modele tanto las dependencias temporales en los precios del activo como la influencia del sentimiento colectivo twitter sobre la dinámica futura del precio. Posteriormente, al realizar las predicciones, se aplicó la transformación inversa para recuperar los precios en su escala original en USD. Esto permite interpretar los resultados de forma práctica y evaluar métricas como el RMSE y el MAE en términos monetarios para una mejor interpretación. Además, se observó que el modelo tiene una capacidad de retroalimentación progresiva durante el entrenamiento. Sin embargo, dicha mejora en precisión se manifiesta a tasas decrecientes conforme aumentan las épocas, lo cual evidencia un intercambio marginal entre el ajuste del modelo y su capacidad de generalización. Este proceso también revela un ***trade-off* inherente entre precisión e interpretabilidad**, ya que modelos más complejos tienden a mejorar el ajuste, pero dificultan la comprensión de las decisiones que toman.

A graph with red dots and blue lines

AI-generated content may be incorrect.

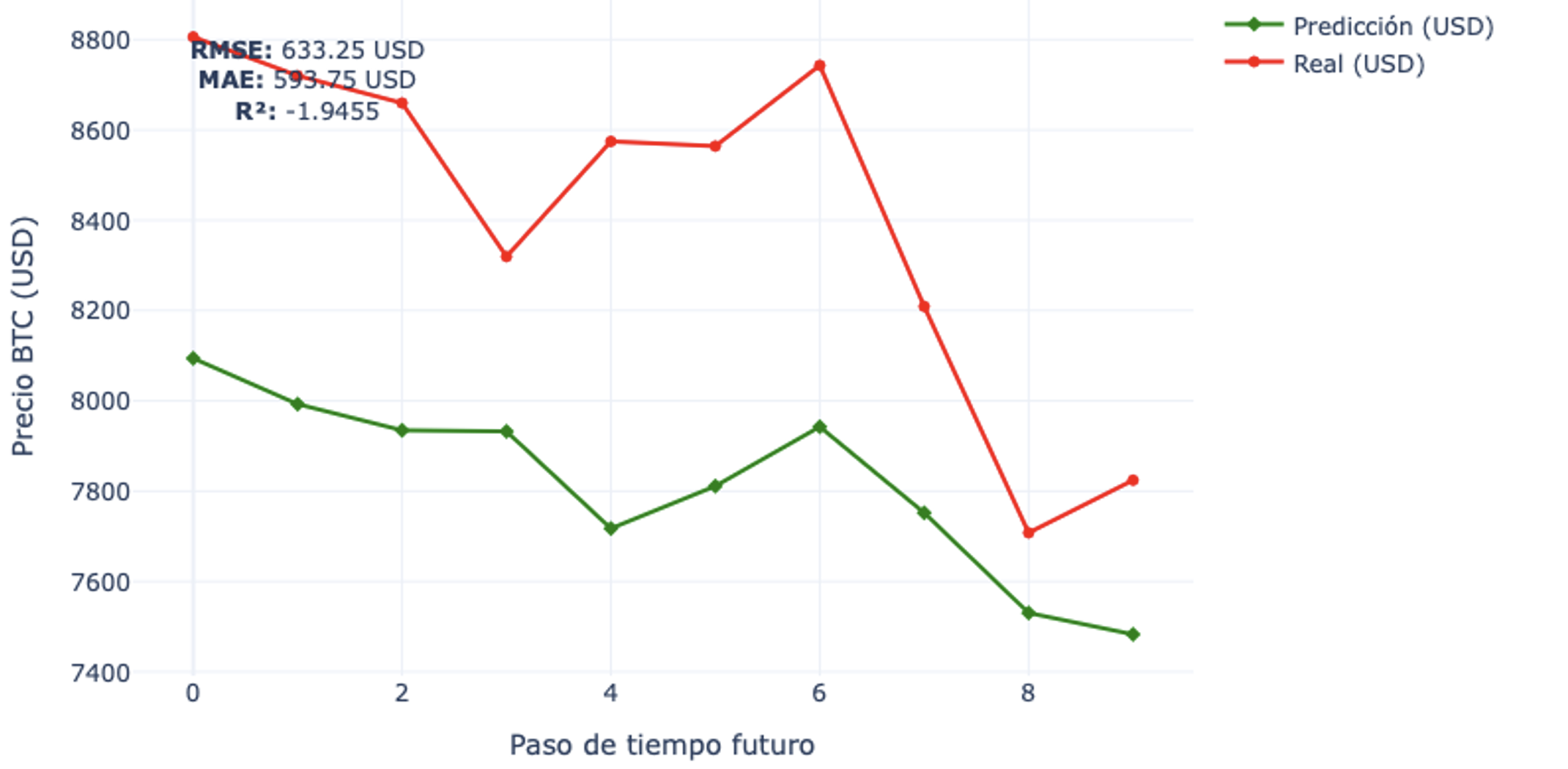
Gráfico 1. Precisión del modelo LSTM a lo largo de las épocas

En cuanto a la última etapa de la metodología, como se puede ver en la gráfica 1, se evaluó la capacidad predictiva del modelo LSTM utilizando métricas estándar en problemas de regresión, es decir, se analizaron indicadores como el Error Cuadrático Medio (RMSE) para comparar los valores predichos con los reales y ver la precisión del modelo. Además, se generaron gráficos de predicción vs realidad, así como curvas de error a lo largo del tiempo, ya que es importante visualizar el comportamiento del modelo y detectar posibles patrones de sesgo o fallas sistemáticas. Para mitigar esto, se realizaron pruebas adicionales para comparar el rendimiento del modelo híbrido (LSTM + sentimiento) contra el rendieminto de Bitcoin, el resultado fue una correlación estaditicamente significativa del 0.8 entre el retorno de inversión y el sentimiento.

**Resultados**

Con respecto a los resultados, estos mostraron que la inclusión de variables de sentimiento aportan valor agregado a las predicciones en ciertos horizontes temporales, en particular, cuando se combinan con una arquitectura LSTM bien calibrada. Esto respalda la hipótesis de que la opinión pública expresada en redes sociales contiene información valiosa para anticipar los movimientos del mercado; en otras palabras, el uso de datos no estructurados contribuye a mejorar los modelos predictivos financieros mediante el uso de redes neuronales LSTM, como se demuestra en las métricas de evaluación. Igualmente, el modelo obtuvo una **precisión del 86 %** y un RMSE (Root Mean Squared Error) de 633.25 USD, lo que significa que, en promedio, las predicciones del precio se desvían del valor real en 633 dólares. Esta métrica penaliza más los errores grandes, por lo que un RMSE bajo indica que el modelo comete pocos errores graves. También, se alcanzó un MAE (Mean Absolute Error) de 593.75 USD que representa el error promedio absoluto entre las predicciones y los valores reales. Reformulando, el modelo se equivoca en promedio por unos 593 dólares sin importar si predice por arriba o por debajo. En cuanto al error relativo, el modelo obtuvo un MAPE (Error Porcentual Absoluto Medio) de 6.96 %, es decir en promedio, las predicciones se apartan del valor real en menos del 7 %, lo cual es un buen resultado para mercados tan cambiantes como el de las criptomonedas. De forma complementaria, el SMAPE (Symmetric Mean Absolute Percentage Error) fue de 7.25 %; esta es una métrica que equilibra mejor los errores cuando los valores reales son muy pequeños o grandes y que ofrece una visión más justa del rendimiento en series temporales.

Finalmente, el modelo anticipó correctamente si el precio subiera o bajaría en el siguiente periodo en el 66.67% de las veces (gráfica 2). Esto es especialmente valioso para tomar decisiones en entornos de inversión y trading, donde preever la dirección del mercado puede ser tan importante como predecir el valor exacto. En comparación, el ARIMA logra una precisión direccional del 80.94%, sin embargo, sus métricas de error reflejan una capacidad predictiva limitada en magnitud: un RMSE de 5,364 USD, un MAE de 3,378 USD y un MAPE del 19.06%. Esto evidencia desviaciones significativas frente a los valores reales, por tanto, este rendimiento, aunque razonable en contextos de alta volatilidad como el de Bitcoin, queda por debajo del ofrecido por arquitecturas LSTM. Esto se debe a su capacidad para modelar dependencias temporales no lineales y capturar dinámicas secuenciales complejas. Los modelos LSTM no solo reducen el error porcentual, sino que también logran una predicción más robusta y adaptable a las oscilaciones del mercado.

Gráfica 2 Compración entre predicción y real – escala real

**Conclusión**

En conclusión, este estudio demuestra que la incorporación de variables de sentimiento derivadas del análisis de texto en redes sociales puede mejorar la precisión de los modelos predictivos financieros, ya que esta mejora se vuelve especialmente notable cuando dichas variables se integran en redes neuronales profundas como las LSTM. Asimismo, se observó que las redes sociales son una fuente de información financiera relevante que no suele reflejarse en los indicadores cuantitativos tradicionales. Por lo tanto, la capacidad de capturar emociones del mercado proporciona una ventaja adicional para anticipar movimientos financieros, especialmente en contextos con alta volatilidad e incertidumbre. Por consiguiente, este resultado valida no solo la pertinencia de incluir datos no estructurados en modelos predictivos, sino también la utilidad práctica de fusionar técnicas de NLP con arquitecturas de redes neuronales profundas para abordar entornos financieros complejos y dinámicos.

Más allá de los avances actuales, resulta imprescindible expandir la investigación hacia nuevas zonas de oportunidad que nos permitan abordar los problemas sociales y económicos desde una perspectiva más integral y profunda. En este contexto, emerge un potencial inexplorado del análisis de datos no estructurados que da pie a las siguinete preguntas: **¿Qué patrones de discriminación estructural podrían revelarse al analizar miles de sentencias judiciales mediante técnicas de procesamiento de lenguaje natural?**, **¿Es posible detectar señales tempranas de enfermedades mentales a partir del análisis de textos informales en redes sociales o foros anónimos?**, **¿Hasta qué punto los patrones lingüísticos presentes en quejas ciudadanas digitales pueden anticipar fenómenos como corrupción institucional o violencia estructural?**, **¿Podría el análisis de conversaciones informales en plataformas digitales predecir con mayor precisión los ciclos de consumo, inflación o percepción de crisis económica que los indicadores tradicionales?, e**stas preguntas abren paso a una reflexión más amplia: los datos no estructurados nos ofrecen una mirada que va más allá de lo que captan los métodos convencionales. **Por lo tanto,** los datos no estructurados nos permiten ver lo que los métodos tradicionales no captan ya que revelan patrones ocultos y anticipan fenómenos complejos, los cuales nos invitan a repensar los problemas desde nuevas perspectivas más dinámicas y contextualizadas.

**Fuentes de consulta**

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. \*Journal of Computational Science, 2\*(1), 1–8.

García-Díaz, J. C., López, J. R., & Arias, R. (2019). El análisis de datos no estructurados representa uno de los mayores desafíos y oportunidades en la era del big data, especialmente en el contexto de las redes sociales. \*Revista Española de Documentación Científica, 42\*(4), 747–754.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). \*Deep learning\*. MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. \*Neural Computation, 9\*(8), 1735–1780.

IBM. (2017). \*The four V’s of big data\*. IBM Big Data & Analytics Hub.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. \*Nature, 521\*(7553), 436–444.

Martínez-López, F. J., & Casillas, J. (2013). Big data: Conceptos, tecnologías y aplicaciones. \*ESIC Editorial\*.

Mittal, A., & Goel, A. (2009). Stock prediction using Twitter sentiment analysis. \*Stanford University CS229 Project Report\*, 1–7.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. \*Neural Networks, 61\*, 85–117.

Sheetal, R., & Aithal, P. K. (2024). Financial report sentiment analysis using Loughran-McDonald dictionary and BERT. \*Financial Engineering, 2\*(15), 162–180.

Smith, J., Brown, T., & Chen, M. (2023). The ability to extract sentiment and trends from unstructured data offers a significant competitive advantage in decision-making. \*Journal of Data Analytics, 15\*(2), 45–59.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2018). LSTM networks have shown superior ability to capture long-term dependencies in sequential data. *IEEE Transactions on Neural Networks*.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2018). Understanding bag-of-words model: A statistical framework. \*International Journal of Machine Learning and Cybernetics, 9\*(1), 183–192.