UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERIA DE SISTEMAS

MODELO PREDICTIVO APLICANDO ANÁLISIS DE SENTIMIENTOS EN TWITTER PARA DETERMINAR EL COMPORTAMIENTO DE LA CRIPTODIVISA BITCOIN

TESIS

PRESENTADA POR:

ERNESTO ZHILDEER CHURA FLORES

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

PUNO – PERÚ

2022

DEDICATORIA

A Dios y a mi familia.

Atte. Ernesto Zhildeer Chura Flores

AGRADECIMIENTOS

A la Escuela Profesional de Ingeniería de Sistemas ya que en sus aulas viví experiencias maravillosas junto a mis compañeros y docentes, recuerdos inolvidables que me acompañarán por el resto de mi vida.

A mi director de tesis Dr. Henry Iván Condori Alejo por su tiempo, paciencia y consejos brindados para realizar este proyecto de investigación.

A los jurados supervisores de este trabajo, Mg. Carlos Boris Sosa Maidana, M. Sc. Edgar Holguin Holguin y M. Sc. Pablo Cesar Tapia Catacora por sus correcciones y recomendaciones.

A mi familia por su apoyo incondicional y confianza.

Atte. Ernesto Zhildeer Chura Flores

ÍNDICE GENERAL

CAPÍTULO III

MATERIALES Y MÉTODOS

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

Área : Minería de datos

Tema : Inteligencia Artificial y Sistemas Bio-Inspirados

FECHA DE SUSTENTACIÓN: 16 de diciembre de 2022

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ACRÓNIMOS

LSTM : Long Short-Term Memory

BTC : Bitcoin

RMSE : Root Mean Square Error

MAPE : Mean Absolute Percentage Error

MAE : Mean Absolute Error

NADAM : Nesterov-accelerated Adaptive Moment Estimation

VADER : Valence Aware Dictionary for sEntiment Reasoning

CRISP-DM : Cross Industry Standard Process for Data Mining

CSV : Comma-Separated Values

RESUMEN

Hoy en día el Bitcoin es la criptodivisa con mayor capitalización de mercado y una nueva opción de inversión bastante llamativa pero riesgosa debido a la incertidumbre generada por la alta volatilidad de la misma lo que dificulta predecir su comportamiento, es por ello que se implementó un modelo predictivo considerando variables de alcance y aplicando el análisis de sentimientos en Twitter para predecir el comportamiento de esta criptodivisa a corto plazo. Para lograr este objetivo, se realizó la recolección y el preprocesamiento de la data histórica del Bitcoin y de los tweets referentes al Bitcoin y se aplicó el análisis de sentimientos utilizando el clasificador VADER al cual se le agregó un diccionario de léxicos con expresiones comúnmente usadas en la comunidad Bitcoin, después se realizó una selección de las variables más representativas utilizando el índice de correlación de Spearman y posteriormente, se aplicó una red neuronal recurrente (RNN) del tipo LSTM (Long Short-Term Memory) con tres configuraciones diferentes para predecir 1 hora, 6 horas y 12 horas a futuro considerando un lookback de 3 horas utilizando la librería Keras. Para evaluar el performance del modelo se utilizaron las métricas: MAPE y RMSE para obtener valores comparables en términos porcentuales y validación interna del error del modelo respectivamente. Finalmente, se encontró que el modelo predictivo configurado para predecir 1 hr. a futuro fue el que mejores resultados obtuvo con un RMSE de 227.413 y un MAPE de 0.022 lo que demuestra que si es posible predecir el comportamiento del Bitcoin; sin embargo, la desventaja radica en la precisión ya que el resultado no es lo suficientemente bueno con respecto al RMSE por lo que no se recomienda basar decisiones de inversión únicamente en los resultados de este modelo.

Palabras clave: Análisis de sentimientos, Twitter, Bitcoin, predicción, comportamiento, LSTM, Aprendizaje Automático.

ABSTRACT

Today Bitcoin is the cryptocurrency with the largest market capitalization and a new investment option that is quite striking but risky due to the uncertainty generated by its high volatility, which makes it difficult to predict its behavior, which is why a model was implemented. predictive considering range variables and applying sentiment analysis on Twitter to predict the behavior of this cryptocurrency in the short term. To achieve this objective, the collection and preprocessing of the historical data of Bitcoin and of the tweets referring to Bitcoin was carried out, and sentiment analysis was applied using the VADER classifier, to which a dictionary of lexicons with expressions commonly used in the Internet was added. the Bitcoin community, then a selection of the most representative variables was made using the Spearman correlation index and subsequently, a recurrent neural network (RNN) of the LSTM (Long Short-Term Memory) type was applied with three different configurations to predict 1 hour, 6 hours and 12 hours in the future considering a 3 hours lookback using the Keras library. To evaluate the performance of the model, the metrics were used: MAPE and RMSE to obtain comparable values in percentage terms and internal validation of the model error, respectively. Finally, it was found that the predictive model configured to predict 1 hr. In the future, it was the one that obtained the best results with an RMSE of 227.413 and a MAPE of 0.022, which shows that it is possible to predict the behavior of Bitcoin; however, the disadvantage lies in the precision since the result is not good enough with respect to the RMSE, so it is not recommended to base investment decisions solely on the results of this model.

Keywords: Sentiment analysis, Twitter, Bitcoin, prediction, behavior, LSTM, Machine Learning.

CAPÍTULO I

INTRODUCCIÓN

El trabajo de investigación tiene por finalidad implementar un modelo predictivo aplicando análisis de sentimientos en Twitter para determinar el comportamiento de la criptodivisa Bitcoin considerando también variables que muestran expectativas e interacciones de dicha red social, de esta manera se espera llegar a resultados y conclusiones efectivas que sirvan de referencia para trabajos futuros sobre este tema y, sobre todo, que apoyen en la toma de decisiones a los inversionistas en el criptomercado.

El Bitcoin es una de las criptomonedas más populares y con mayor valor de mercado actualmente, esto ha conducido a un incremento sustancial en la cantidad de inversionistas o “traders” de esta criptodivisa en los últimos años. Un inversionista en el criptomercado busca maximizar sus ganancias y reducir al mínimo sus pérdidas, sin embargo, lograrlo es complicado y más aún en un mercado tan volátil y cambiante como lo es el criptomercado, esta necesidad nos lleva a explorar una posible solución mediante los modelos predictivos utilizando redes neuronales. El presente trabajo de investigación expone los siguientes capítulos:

En el Capítulo I, denominado planteamiento del problema, se explican los objetivos e hipótesis del trabajo de investigación, además, en este mismo capítulo se exponen los motivos por los cuales esta investigación es relevante.

En el Capítulo II, denominado marco conceptual, se explican brevemente los términos y conceptos que contribuyen a una mejor compresión de todo el trabajo.

En el Capítulo III, denominado método de la investigación, se expone el tipo y diseño de la investigación, la metodología utilizada para desarrollar este proyecto de

ciencia de datos y la técnica de recolección de datos que se utilizó para este trabajo de investigación.

En el Capítulo IV, denominado resultados y discusión, se explican a detalle los procedimientos realizados para el desarrollo del presente proyecto de investigación y los resultados obtenidos de la investigación. Además, en este capítulo se contrastan los resultados obtenidos en esta investigación con estudios previos mencionados en el capítulo II.

Y por último se exponen las conclusiones, recomendaciones y bibliografía.

1.1. PLANTEAMIENTO DEL PROBLEMA

Hoy en día, el Bitcoin es la criptodivisa más popular a nivel mundial, cuenta con una comunidad bastante amplia de bitcoiners y es también una relativamente nueva opción de inversión bastante llamativa con la promesa de una alta tasa de rentabilidad, pero riesgosa debido a la incertidumbre generada por la alta volatilidad de la misma criptodivisa lo que dificulta predecir su comportamiento. Efectivamente, la criptodivisa Bitcoin ha reportado ganancias de 630% en un año considerando solo el periodo hasta 2017, año en el que durante la primera semana de noviembre el Bitcoin había superado los $ 7,300.00 dólares por unidad, cifra que un año anterior era impensada, cuando apenas bordeaba los $ 1,000.00 dólares y todavía era considerada una moneda relativamente estable. (ESAN, 2017)

Según BBVA (2022), el Bitcoin atrae cada vez más inversionistas por las ventajas que representa frente al dinero convencional ya que permite realizar transacciones inmediatas de manera segura e irreversible lo que protege a los comerciantes contra pérdidas por fraude, pero también cubre a los clientes protegiendo sus datos personales. Además, el inicio de la pandemia en el 2020 ocasionó que más gente se familiarizara en

su interacción diaria con el ciberespacio y las criptomonedas se volvieron un tema recurrente en el que se interesaron y empezaron a seguir, lo que significó una explosión en el criptomercado en estos dos últimos años y según Binance, las razones para su uso van desde "ahorrar y generar ingresos pasivos" en medio de la crisis, hasta "enfrentar los cambios socio-políticos" que presentan los usuarios. Sin embargo, el precio ha ido fluctuando en el llamado "mercado rojo" reduciendo su valor hasta los 29 mil dólares por unidad en el caso del Bitcoin, causando pánico entre los inversionistas del criptomercado que los llevaron a tomar decisiones de inversión precipitadas perdiendo así grandes cantidades de dinero. (RPP, 2021)

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo desarrollar un modelo predictivo aplicando análisis de sentimientos en Twitter para determinar el comportamiento de la criptodivisa Bitcoin?

1.3. JUSTIFICACIÓN DEL PROBLEMA

Actualmente el Bitcoin es la criptomoneda con mayor capitalización de mercado, aproximadamente 1,06 T $ (Investing.com, 2021) y con mayor cantidad de comerciantes activos en el mundo, entre 51,2 y 52,4 M en lo que respecta a criptodivisa (Bucquet et al., 2019), todo esto a partir del denominado "Boom del Bitcoin" del año 2017 y su posterior revalorización en el año 2021, es decir, la comunidad de Bitcoin es extremadamente grande y sigue creciendo día a día. Sin embargo, el problema para los inversionistas radica en la incertidumbre generada por la alta volatilidad del Bitcoin y la dificultad para predecir su comportamiento. El presente estudio de investigación, al utilizar Machine Learning y Natural Language Processing, aporta conocimientos en un tema que es poco valorado, pero de gran importancia para predecir observaciones futuras con una alta precisión considerando el sentimiento del mercado de habla inglesa y la influencia del mismo en una red social como Twitter que, a la vez, sirva de apoyo en la toma de

decisiones financieras por parte de los inversionistas de Bitcoin. Por otro lado, existen ciertos factores que influyen en gran manera al comportamiento del Bitcoin, este proyecto de investigación busca evaluar todos esos factores y desarrollar un modelo predictivo para poder ver a futuro el comportamiento del Bitcoin. Además según Badiola Ramos (2019), debemos considerar que se han realizado distintos trabajos de investigación relacionados con la influencia de los sentimientos en la toma de decisiones por parte de los actores del mercado pero la gran mayoría se enfoca más sobre el cómo influye la información de redes sociales o medios tradicionales a los mercados bursátiles tradicionales y aquellos pocos trabajos que se centran en la influencia de los sentimientos en el criptomercado difieren en sus resultados ya que algunos aseguran que sí existe una relación de influencia entre el sentimiento y la variación del precio de las criptomonedas mientras que otros niegan la existencia de esta relación. La mayoría de los estudios previos relacionados a este tema tuvieron limitaciones en la extracción de datos históricos de Twitter lo que los llevó a realizar el análisis en franjas de tiempo muy pequeñas, lo cual no es óptimo al momento de utilizar modelos predictivos basados en redes neuronales recurrentes (RNN).

Es necesario mencionar también que este trabajo utiliza el clasificador VADER agregándole un diccionario de léxicos o expresiones propias del campo de Bitcoin con el objetivo de obtener clasificaciones más precisas para la presente investigación. Finalmente, este trabajo de investigación propondrá nuevas variables de entrada para el modelo predictivo que anteriores investigaciones no consideraron, dichas variables pueden ser agrupadas bajo un término general “influencia o alcance”.

1.4. OBJETIVOS DE LA INVESTIGACIÓN

1.4.1. Objetivo General

Desarrollar un modelo predictivo para determinar el comportamiento de la criptodivisa Bitcoin aplicando análisis de sentimientos en Twitter.

1.4.2. Objetivos Específicos

- Recopilar y preprocesar la data histórica de Bitcoin y los tweets referidos al Bitcoin.
- Aplicar el análisis de sentimientos y la selección de características operando el índice de Spearman.
- Implementar el modelo de predicción con tres diferentes configuraciones de la red neuronal LSTM.
- Evaluar y comparar el performance de las diferentes configuraciones del modelo predictivo.

1.5. HIPÓTESIS DE LA INVESTIGACIÓN

1.5.1. Hipótesis General

El modelo predictivo predice adecuadamente el comportamiento de la criptodivisa Bitcoin aplicando análisis de sentimientos en Twitter.

CAPITULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES

2.1.1. Antecedentes Internacionales

Badiola Ramos (2019) estudia la posibilidad de predecir el precio del Bitcoin aplicando análisis de sentimientos en 17 737 519 tweets sobre Bitcoin entre agosto del 2017 hasta enero de 2019, la investigación planteaba la hipótesis de que el sentimiento del mercado si tiene relación con los cambios en el precio de Bitcoin y que el volumen de los datos sería irrelevante, de los datos obtenidos se observa que, a pesar de la caída del valor del Bitcoin, el sentimiento medio expresado en los mensajes se mantiene positivo, es decir, el sentimiento no es una buena variable para predecir el valor de Bitcoin. También se observa que la variable “volumen” correspondiente a los mensajes tiene una correlación alta y sólida durante todo el periodo de estudio. Para este trabajo se crearon 3 modelos diferentes que predicen 1h, 4h y 24h en el futuro, la investigación considera dos tipos de variables para crear los modelos de predicción, volumen de tweets y sentimiento medio de los Tweets. Los modelos tienen un error cuadrático medio de 387, 483 y 818 respectivamente, pero algo interesante a resaltar es que después de una gran caída del modelo, este tiende a predecir por encima de los valores reales, esto se debe a que la caída fue un evento único no presente en los datos de entrenamiento del modelo. Sin embargo, al incluir el entrenar el modelo con el evento único los resultados de predicción son mucho mejores con unos errores cuadráticos medios de 244, 320, 600 para los modelos de 4, 24 y 48 horas.

Awoke et al. (2021) compara dos modelos de predicción basados en Deep Learning: long short-term memory (LSTM) y gated recurrent unit (GRU) para manejar la volatilidad del precio del Bitcoin y obtener una precisión alta.

Ferdiansyah et al. (2019) estudian como crear un modelo de predicción del precio de Bitcoin utilizando una red Long-Short Term Memory (LSTM), considerando 500 epochs para el modelo y un dropout de 0 obteniendo, así como mejor resultado un RMSE de 288.59866. Además, se menciona que, según el modelo, el resultado del precio está por encima de $ 12600 para los próximos días considerados en el estudio de investigación.

Jiang (2020) analiza diferentes redes de aprendizaje profundo y métodos para mejorar la precisión, incluida la normalización mínima y máxima, el optimizador Adam y la normalización mínima y máxima de Windows, recopilaron datos sobre el precio del Bitcoin por minuto y los reorganizaron para reflejar el precio del Bitcoin en horas, un total de 56 832 puntos, tomaron 24 horas de datos como entrada y salida del precio del Bitcoin de la siguiente hora, luego compararon los diferentes modelos y descubrieron que la falta de memoria significa que el perceptrón multicapa (MLP) no es adecuado para el caso de predecir el precio según la tendencia actual, Long Short-Term Memory (LSTM) proporciona relativamente la mejor predicción cuando la memoria pasada y la red recurrente cerrada (GRU) se incluyen en el modelo.

Li & Dai (2019) proponen un modelo de red neuronal híbrida basada en una red neuronal convolucional (CNN) y una red neuronal long short-term memory (LSTM). Los datos de las transacciones de Bitcoin en sí, así como la información externa, como las variables macroeconómicas y la atención de los inversores, se toman como datos de entrada. Primero, CNN se utiliza para la extracción de características. Luego, los vectores de características se ingresan en LSTM para entrenar y pronosticar el precio a corto plazo del Bitcoin. El resultado muestra que la red neuronal híbrida CNN-LSTM puede mejorar

efectivamente la precisión de la predicción del valor y la predicción de dirección en comparación con la red neuronal de estructura única.

Pagolu et al. (2017) observan qué tan bien los cambios en los precios de las acciones de una empresa, las subidas y caídas, están correlacionadas con las opiniones públicas expresadas en tweets sobre una empresa. Comprender la opinión del autor a partir de un texto es el objetivo del análisis de sentimientos. Se emplearon dos representaciones textuales diferentes, Word2vec y Ngram, para analizar los sentimientos del público en los tweets. En este trabajo se aplicó el análisis de sentimiento y principios de machine learning supervisado a los tweets extraídos de Twitter y analizaron la correlación entre los movimientos del mercado de valores de una empresa y sentimientos en tweets. De forma elaborada, las noticias y tweets positivos sobre una empresa en las redes sociales definitivamente animan a la gente a invertir en las acciones de esa empresa y como resultado el precio de la acción de esa empresa aumentaría. Al final del artículo, se muestra que existe una fuerte correlación entre los aumentos y caídas de los precios de las acciones con los sentimientos del público en Twitter.

Stenqvist & Lönnö (2017) analizan 2,27 millones de tweets relacionados con Bitcoin para las fluctuaciones de sentimiento que podrían indicar un cambio de precio en un futuro próximo. Esto se hace mediante un método Naive de atribuir únicamente el aumento o caída en función de la gravedad de cambio de sentimiento de Twitter agregado durante periodos que oscilan entre 5 minutos y 4 horas, y luego adelantando estas predicciones en el tiempo 1, 2, 3 o 4 periodos de tiempo para indicar el tiempo de intervalo BTC correspondiente. La evaluación del modelo de predicción mostró que agregar sentimientos en tweets durante un período de 30 minutos con 4 cambios hacia adelante y un umbral de cambio de sentimiento del 2,2%, arrojó una precisión del 79%.

Abraham et al. (2018) presentan un método para predecir cambios en los precios de Bitcoin y Ethereum utilizando data de Twitter y Google Trends, investigan la relación entre 2 variables interesantes (el volumen y el sentimiento de los tweets que hablan sobre Bitcoin y Ethereum), y el volumen de búsquedas en Google de ambas criptomonedas con la fluctuación del precio registrado en el intervalo de estudio (marzo y mayo de 2018). Durante su trabajo de investigación descartaron la variabilidad del sentimiento porque descubrieron que no tenía una gran correlación con el precio. Según los autores posiblemente el hecho de realizar el estudio en un momento de baja (caída) del mercado es afectó los resultados, mientras que otros estudios que si encontraron relación lo hicieron en un momento de subida del mercado. Aun así, utilizaron un modelo lineal que toma como entrada tweets y data de Google Trends, fueron capaces de predecir acertadamente la dirección de los cambios de precio.

Khedr et al. (2021) explican que las amplias fluctuaciones en los precios de las criptomonedas motivan la necesidad urgente de un modelo preciso para predecir su precio. El trabajo de investigación en este campo utiliza técnicas tradicionales de estadística y aprendizaje automático, como la regresión bayesiana, la regresión logística, la regresión lineal, la máquina de vectores de soporte, la red neuronal artificial, el aprendizaje profundo y el aprendizaje por refuerzo. El artículo que presentan proporciona un resumen completo de los estudios anteriores en el campo de la predicción de precios de criptomonedas de 2010 a 2020. Según los autores, la discusión presentada en este artículo ayudará a los investigadores a llenar el vacío en los estudios existentes y obtener más información futura.

Wołk (2020) propone que el análisis de sentimiento puede ser utilizado como una herramienta computacional para predecir los precios del Bitcoin y otras criptomonedas para diferentes intervalos de tiempo. Una de las principales características

del criptomercado es que la fluctuación de los precios de las criptomonedas depende de las percepciones y opiniones de la gente, no de una regulación institucional de dinero. Además, analizar la relación entre las redes sociales y búsquedas en navegadores es crucial para la predicción del precio de las criptomonedas. Este estudio utiliza Twitter y Google Trends para pronosticar los precios a corto plazo de las criptomonedas primarias, así como estas plataformas de redes sociales son usadas para influenciar decisiones de compra. El estudio adopta e interpola un único enfoque multimodelo para analizar el impacto de las redes sociales en los precios de las criptomonedas. Los resultados prueban que actitudes psicológicas y conductuales tienen un impacto significativo en los precios altamente especulativos de las criptomonedas.

Sabalionis et al. (2021) tienen como objetivo explicar las fluctuaciones del precio de las dos criptomonedas que representan la mayoría de la capitalización del criptomercado, Bitcoin y Ethereum. Se estimado un modelo VAR-GARCH-BEKK para analizar cómo el interés de búsquedas de Google, número de tweets y direcciones activas en el Blockchain impacta en los precios de Bitcoin y Ethereum a lo largo del tiempo. Se encuentra evidencia sólida de que la cantidad de direcciones activas es la variable más significativa entre otras que influencian la fluctuación del precio de Bitcoin y Ethereum. Según los efectos secundarios y los GIRFs, las búsquedas de Google y los Tweets, hasta cierto punto, tienen impactos en los precios de Bitcoin y Ethereum, pero los impactos son más débiles que los de las direcciones activas en términos de magnitud e importancia.

2.1.2. Antecedentes Nacionales

Regal et al. (2019) plantea analizar en qué medida las publicaciones de las redes sociales pueden capturar las expectativas colectivas de los inversores, y afectar el valor futuro de la moneda. Su objetivo es pronosticar el desempeño diario de un mercado en base a dos componentes: aquellos que definen el comportamiento de la criptomoneda en

sí (volumen, valor de apertura, valor de cierre, valor máximo y valor mínimo) y las expectativas e interacciones del entorno, obtenidas de los tweets recolectados. Para ello, proponen el uso de un tipo de red neuronal recurrente, conocida como "Long Short Term Memory" (LSTM). La metodología que emplearon para el preprocesamiento de los datos y la aplicación de esta técnica de pronóstico de series temporales les permite obtener una predicción con un Error Porcentual Absoluto Medio de 34.92%; lo que implica que la representación de la variable de percepción en redes sociales no ha sido la pertinente y, por lo tanto, motiva nuevos trabajos con la finalidad de modelar dichas variables mediante el uso de otras técnicas de NLP.

2.2. MARCO TEÓRICO

2.2.1. Las Criptodivisas o Criptomonedas

Según Sarmiento & Garcés (2016), "las criptodivisas surgen como una alternativa al sistema monetario actual, y tienen como "ventajas" que no se encuentran reguladas por ningún Banco Central, y de cierta manera evita la presencia de intermediarios en el proceso de creación del dinero y las transacciones que con él se realicen, sin el uso de servidores, mediante un esquema P2P".

El European Central Bank (2022) las define como "representaciones digitales de valor no emitidas por ninguna autoridad central bancaria, institución de crédito o emisor de dinero electrónico reconocido que, en ciertas ocasiones, pueden ser utilizadas como medio de pago alternativo al dinero". Existen distintos actores y roles en este tipo de esquemas de moneda virtual, entre los más importantes tenemos:

- **Inventores:** Crean una moneda virtual y desarrollan la parte técnica de su red. En algunos casos, los autores son conocidos y en algunos casos no (por ejemplo,

Bitcoin). Después del lanzamiento algunos continúan involucrados en el mantenimiento y la mejora de las características técnicas de la criptodivisa.

- **Emisores:** Los emisores pueden generar unidades de la moneda virtual. Dependiendo del diseño de la criptodivisa, el volumen total de emisión está predeterminado o depende de la demanda. En el caso de criptodivisas centralizadas, el emisor que usualmente es el administrador, puede establecer reglas para su uso e incluso puede retirar unidades de circulación. Una vez emitidas las unidades, normalmente se entregan a los usuarios, ya sea vendiéndolas o distribuyéndolas gratuitamente. Por otro lado, en el caso de criptodivisas descentralizadas, se pueden crear nuevas unidades automáticamente como resultado de las actividades realizadas por "mineros", quienes reciben algún tipo de recompensa.
- **Mineros:** Los mineros son personas o grupos de personas que, voluntariamente, ponen a disposición el procesamiento informático para validar un conjunto de transacciones (bloques) realizadas con la red de la criptodivisa descentralizada y añadir estos al libro de pagos (cadena de bloques o blockchain). Es importante mencionar que sin los "mineros", la criptodivisa descentralizada no funcionaría correctamente ya que, fácilmente, se podrían introducir unidades falsas o doblemente gastadas. Como recompensa por su trabajo, los mineros normalmente reciben un número específico de unidades de la criptodivisa.
- **Usuarios:** Los usuarios eligen obtener criptodivisas para comprar bienes y servicios reales o virtuales de comerciantes específicos, para realizar (por ejemplo, transfronterizos) o enviar remesas, o con fines de inversión, incluida la especulación. Existen cinco formas de obtener unidades de una criptodivisa: 1) compra: 2) participación en actividades que son recompensadas con una cantidad

determinada de unidades de la criptodivisa (por ejemplo, responder encuestas, participar en actividades promocionales, etc.); 3) realizando actividades como minero (minería de criptodivisas); 4) recibir unidades como pago o, 5) recibir unidades como donación o regalo.

- **Proveedores de monederos:** Los proveedores de monederos ofrecen un monedero digital a los usuarios para almacenar sus claves criptográficas de criptodivisas y códigos de autenticación de transacciones, realizar transacciones y proporciona una descripción general de su historial de transacciones. Existen dos tipos de billeteras que difieren en cuanto a su usabilidad inmediata y su seguridad frente a delitos cibernéticos: monederos en línea (almacenamiento caliente) y monederos fuera de línea (almacenamiento en frío).
- **Plataformas de intercambio:** Las plataformas de intercambio ofrecen servicios comerciales al cotizar los tipos de cambio por los cuales la plataforma comprará o venderá criptodivisas contra las principales monedas principales (dólar estadounidense, yen, euro, etc.) o contra monedas virtuales. Por lo general, aceptan una amplia gama de opciones de pago que incluyen efectivo, transferencias de crédito y pago con otras criptodivisas.
- **Plataformas de trading:** Las plataformas de trading funcionan como mercados, reuniendo a compradores y vendedores de criptodivisas al proporcionarles una plataforma en la que pueden ofrecer y ofertar entre ellos. Sin embargo, a diferencia de las plataformas de intercambio, las plataformas de trading no se involucran en la compra y venta por sí mismas.

2.2.1.1. Principales Criptomonedas

Actualmente, la criptomoneda más conocida es Bitcoin, pero existen muchas más criptomonedas que son denominadas Altcoins, en la Tabla 1 se mostrarán las 5 criptomonedas dominantes según CoinMarketCap (2022).

Tabla 1: Criptodivisas Principales

Nombre	Símbolo	Capitalización de Mercado	Características
Bitcoin	BTC	$403,091,624,747	Criptomoneda que usa distributed-ledger technology (DLT) sin permisos. Tiene un límite de suministro fijado en 21M de BTC
Ethereum	ETH	$147,849,094,454	Ethereum es un sistema blockchain de código abierto descentralizado que incluye su propia criptomoneda, Ether. ETH funciona como plataforma para otras numerosas criptomonedas, así como para la ejecución de contratos inteligentes descentralizados.
Tether	USDT	$66,833,547,878	Tether (USDT) es una moneda digital con un valor destinado a ser un reflejo del dólar estadounidense. Lanzada en 2014, la idea detrás de

(continuación…)

			Tether era crear una criptomoneda estable o "stablecoin" que se pueda usar como dólares digitales. Los Tethers están anclados o "atados" al precio del dólar estadounidense.
USD Coin	USDC	$55,852,470,080	Es una stablecoin que está vinculada al dólar estadounidense en una base de 1:1. Todas las unidades de esta criptomoneda en circulación están respaldadas por $1 que se mantienen en reserva, en una combinación de efectivo y bonos del Tesoro de Estados Unidos a corto plazo.
BNB	BNB	$38,515,520,889	BNB se lanzó a través de una oferta inicial de monedas en 2017, 11 días antes de que el exchange de criptomonedas Binance se pusiera en marcha. Originalmente se emitió como un token ERC-20 que se ejecutaba en la red de Ethereum, con un suministro total limitado a 200 millones de monedas y 100

(continuación…)

	millones de BNB ofrecidos en la ICO. BNB se puede utilizar como método de pago, un token de utilidad para pagar las tarifas en el exchange de Binance y para participar en las ventas de tokens en la plataforma de lanzamiento de Binance.

Fuente: (CoinMarketCap, 2022)

2.2.2. Bitcoin

Noboa & Navas (2022) comentan que el Bitcoin como tal es un sistema digital, una tecnología que tuvo origen en el año 2009 cuyo objetivo es el de implantar un sistema de pagos descentralizado alrededor se crearon 18.5 millones de BTC.

La red Bitcoin fue lanzada en enero de 2009 por un programador informático anónimo o un grupo de programadores bajo el seudónimo de “Satoshi Nakamoto”. La red es un sistema de pago electrónico entre pares que utiliza una criptomoneda llamada bitcoin para transferir valor a través de internet o actuar como una reserva de valor como el oro y la plata. Cada bitcoin se compone de 100 millones de satoshis (las unidades más pequeñas de bitcoin), lo que hace que cada bitcoin sea divisible hasta con ocho decimales. Eso significa que cualquiera puede comprar una fracción de bitcoin con tan solo un dólar estadounidense. (coindesk.com, 2022)

2.2.2.1. Componentes del Bitcoin

Tabla 2: Componentes del Bitcoin

Componente	Concepto
Blockchain (Cadena de bloques)	Según Dolader et al. (2017), la Blockchain (cadena de bloques) es una base de datos que permite almacenar información inmutable y ordenada que puede ser compartida por muchos usuarios. Para el caso de las criptomonedas como el Bitcoin, la información añadida a la blockchain puede ser vista y consultada por cualquier persona ya que es pública y está siempre disponible. Además, solo es posible agregar información a esta blockchain si existe un acuerdo entre la mayoría de las partes, esta agregación de información es realizada por "mineros", denominados así aquellos nodos que participan en el proceso de escritura de datos en la blockchain a cambio de una recompensa (cantidad de satoshis o porciones de Bitcoin).
Proof of Work (PoW)	Según Sheikh et al. (2018), PoW es el algoritmo de consenso original en una red Blockchain, donde los usuarios se envían un token digital entre sí, verifica las transacciones y crea nuevos bloques para la cadena. En este algoritmo, todos los mineros o validadores participan para validar y confirmar cuidadosamente las transacciones en la red para ser recompensados. Todas

(continuación…)

	las transacciones verificadas en la red se recopilan en bloques mediante el libro mayor o principal distribuido y se organizan en consecuencia. Este proceso se llama minería. Proof of Work es un protocolo que previene amenazas cibernéticas como en el ataque de denegación de servicio distribuido (DDoS, por sus siglas en inglés) que tiene la intención de drenar los recursos de la computadora enviando numerosas solicitudes falsas.
Minería	Zocaro (2021) menciona que es importante aclarar que la minería de criptomonedas no se trata de descubrir nuevas criptomonedas, sino que se denomina así a los procesos que los mineros llevan a cabo para validar las transacciones. Y no todos los usuarios de criptomonedas serán mineros, sino sólo aquellos que han decidido invertir en equipos informáticos para llevar a cabo dicha actividad, buscando obtener cierta rentabilidad. Sin profundizar en cuestiones técnicas, el funcionamiento de la minería se puede describir de la siguiente manera: gracias a internet, el minero (persona humana o empresa) conecta determinado tipo de hardware a la red y se descarga el correspondiente software, conformando así un "nodo"; y en forma automática (sin mayor intervención humana) el equipo informático competirá contra otros mineros intentando

(continuación…)

	descifrar ciertos algoritmos (“acertijos” matemáticos). El primero que logre resolver el algoritmo, se le permitirá anexar un nuevo bloque con información a la blockchain y recibir, como “recompensa”, nuevas criptomonedas. Y así es como se “generan” nuevas unidades de estos activos.
Mining Pools	Según Tovanich et al. (2022), el concepto de mining pools se refiere a que los mineros combinan recursos computacionales para obtener un ingreso más estable y predecible lo que es más efectivo para los propios mineros, es por ello que los mineros individuales son muy raros actualmente. Las mining pools compiten entre sí para atraer a más mineros variando la forma en que pagan las recompensas a sus miembros. En general, varían dos factores: las tarifas de grupo, que son tarifas que el grupo mantiene por sus servicios, y los esquemas de pago que determinan cúando y cómo se pagan las recompensas.
Sistema descentralizado	Según Rahouti et al. (2018), en el caso del Bitcoin, su funcionamiento está basado en un sistema de red P2P que contiene un número infinito de nodos interactuando sin ningún intermediario, donde cada nodo cumple la misma función y un protocolo de consenso

(continuación…)

	probabilístico descentralizado donde todos los pagos e intercambios se realizan electrónicamente a través de transacciones entre cliente, además, el poder monetario no se encuentra regulado o controlado por ninguna persona o entidad bancaria. Por otro lado, para Rodriguez Garagorri (2017) es importante mencionar también que, al no haber una persona o entidad que controle el funcionamiento de este sistema, existe un mecanismo de consenso que define las reglas que deben cumplirse para que un bloque del blockchain sea dado como válido, para que no se vulnere la integridad de toda la cadena de este sistema descentralizado.
Monedero Bitcoin	Según (Delgado Mohatar & Ortigosa Juárez, 2019) un monedero de Bitcoin es el espacio virtual en el que se almacenan las claves necesarias para efectuar operaciones con Bitcoin. Lo que se guarda en un monedero Bitcoin son las claves criptográficas formadas por una clave pública y otra privada. La clave pública (bitcoin address) es la que se proporciona a otros usuarios para que transfieran bitcoins a dicho monedero, cualquiera puede verla y no requiere protección ya que únicamente sirve para recibir bitcoins. Por otro lado, tenemos la clave privada, que es

(continuación…)

	la clave necesaria para poder hacer uso de los bitcoins guardados en el monedero. Si se quiere realizar un pago en bitcoins, se debe confirmar la transacción con la clave privada por lo que es fundamental guardarla en un lugar seguro. (Fintech-Observatorio, 2017)

Elaboración propia

2.2.2.2. Características

En el trabajo de D. K. C. Lee et al. (2018), se mencionan cinco características del Bitcoin, las cuales son:

- **Descentralizado:** No existe una institución que controle la red Bitcoin, su suministro se rige por un algoritmo, y cualquiera puede tener acceso a él a través de internet.
- **Flexible:** Los monederos o direcciones de Bitcoin se pueden configurar fácilmente en línea sin tarifas ni regulaciones. Además, las transacciones no son específicas de la ubicación, por lo que los bitcoins se pueden transferir entre diferentes países sin problemas.
- **Transparente:** Cada transacción se transmitirá a toda la red (miles de nodos). Los nodos de minería o los mineros validarán las transacciones, las registrarán en el bloque que están creando y transmitirán el bloque completo a otros nodos (usuarios). Los registros de todas las transacciones se almacenan en la cadena de bloques, que está abierta y distribuida, por lo que cada minero tiene una copia y puede verificarlos.
- **Rápido:** Las transacciones se trasmiten en unos pocos segundos y los mineros tardan unos 10 minutos en verificar la transacción. Por lo tanto, uno puede

transferir bitcoins a cualquier parte del mundo, y las transacciones, generalmente, se completarán minutos después.

- **Tarifas de transacción bajas:** Históricamente, no se requiere una tarifa de transacción para realizar una transferencia, pero el propietario puede optar por pagar más para facilitar una transacción más rápida. Actualmente, la baja prioridad para las transacciones de minería (una función de la edad y el tamaño de entrada) se usa principalmente como indicador de transacciones de spam, y casi todos los mineros esperan que cada transacción incluya una tarifa. Históricamente, los mineros han sido incentivados principalmente por monedas recién creadas, pero eso está cambiando. A medida que la cantidad de bitcoins en circulación se acerque a su límite, las tarifas de transacción eventualmente serán el incentivo para que los mineros lleven a cabo el costoso proceso de verificación.

Adicionalmente, Zarraluqui Matos (2018) menciona dos características interesantes del Bitcoin:

- **Anónimo:** Si bien es cierto las operaciones quedan registradas en la red y cualquier persona puede verlas, lo que no se puede ver son los datos de quien realizó la operación.
- **Volátil:** Los profesionales del mundo financiero, por lo general, no suelen confiar en el Bitcoin justamente por esta característica. El Bitcoin es difícil de anticipar y las inversiones en esta criptomoneda son consideradas de alto riesgo.

2.2.2.3. Funcionamiento

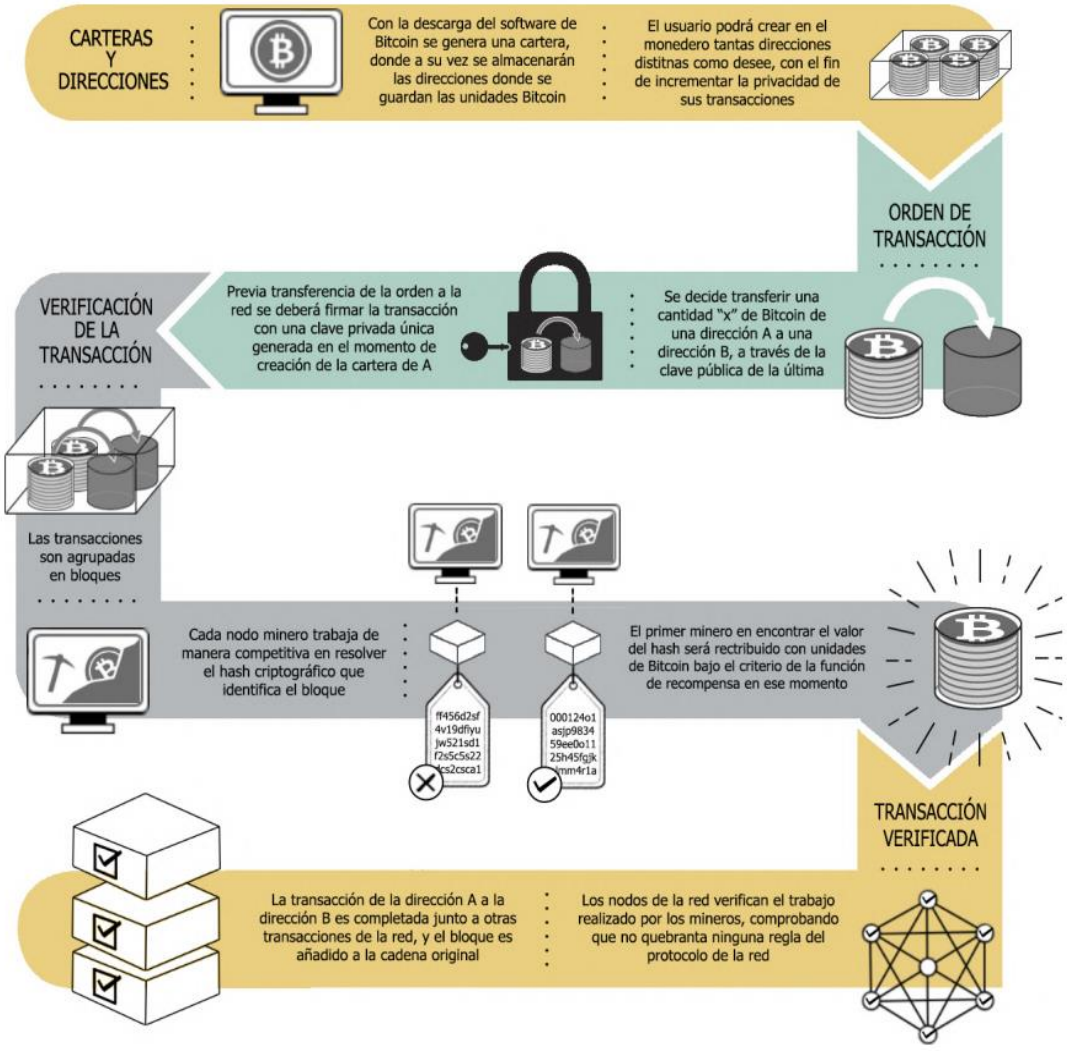
El funcionamiento de la red Bitcoin requiere de distintos procesos y tareas que incluye verificación de las transacciones, minado de los nodos pertenecientes a la red y la blockchain propiamente ya que es la base de esta criptomoneda.

Lo primero para poder empezar con Bitcoin es tener instalado un monedero ya sea en el ordenador o dispositivo móvil, de esta forma se generará la primera dirección Bitcoin para el usuario, es posible crear más direcciones según se necesiten. Esta dirección Bitcoin se puede compartir públicamente con otros usuarios para realizar transacciones. Como mencionamos hace un momento, la cadena de bloques o "blockchain" es una contabilidad pública compartida en la que se basa toda la red Bitcoin y cada transacción confirmada se incluye en esta cadena de bloques, de esta manera los monederos Bitcoin pueden calcular su saldo gastable y las nuevas transacciones pueden ser verificadas para poder asegurar que el cobro se está haciendo al realizar el pago. Aquí entra a tallar un término de suma importancia "criptografía", gracias a esta podemos hacer asegurar la integridad y el orden cronológico de la cadena de bloques.

Otro de los ejes principales del sistema son las transacciones, una transacción es una transferencia de valor entre monederos Bitcoin (usuarios) que se incluye en la cadena de bloques. Cuando un usuario envía una cantidad de bitcoin a otro usuario (mediante los monederos), debe firmar la operación utilizando su clave privada de la cual dispone su monedero, de esta forma se proporciona una prueba matemática de que la transacción está hecha por el propietario del monedero. La firma también evita que la transacción sea alterada por algún tercero asegurando su integridad. Todas las transacciones son difundidas entre los usuarios de la red y, generalmente, empiezan a ser confirmadas en los 10 minutos siguientes a través de un proceso llamado "minería". La minería es un

sistema de consenso distribuido utilizado para confirmar las transacciones pendientes a ser incluidas en la cadena de bloques. Algunas de las tareas que cumple son: Cumplir un orden cronológico en la cadena de bloques, proteger la neutralidad de la red y permitir un acuerdo entre todos los equipos sobre el estado del sistema. Las transacciones son empacadas en bloques cumpliendo normas estrictas de cifrado que serán verificadas por la red. Si un bloque anterior de la cadena de bloques es modificado, automáticamente se invalidan los bloques siguientes y gracias al comportamiento minero competitivo ninguna persona puede controlar lo que está incluido en la cadena de bloques o reemplazar partes de la cadena con propósitos desleales. (bitcoin.org, 2022)

Figura 1: Procesos en una transacción de la red Bitcoin



Fuente: (Gomez Rodriguez, 2020)

2.2.2.4. Regulación Legal del Bitcoin en el Perú

En el Perú es completamente legal que personas naturales o jurídicas realicen transacciones u operaciones con criptomonedas, entre ellas el Bitcoin. No obstante, no existe un marco normativo que reglamente dichas operaciones por lo que constituye un vacío absoluto en el ámbito legal. Si bien es cierto, el año 2021 se presentó al congreso el Proyecto de Ley N° 1042-2021-CR, titulado "Proyecto de Ley Marco de Comercialización de Criptoactivos" en el que se proponen ciertas medidas regulatorias para el criptomercado, no fue aprobado ya que tanto la SBS (Superintendencia de Banca, Seguros y AFP) y el BCRP (Banco Central de Reserva del Perú) emitieron opiniones desfavorables haciendo énfasis en los riesgos de las operaciones con criptomonedas por su naturaleza tan volátil y los riesgos de las criptomonedas en relación con el lavado de activos y otras actividades ilegales. (La Ley, 2022)

2.2.3. Twitter

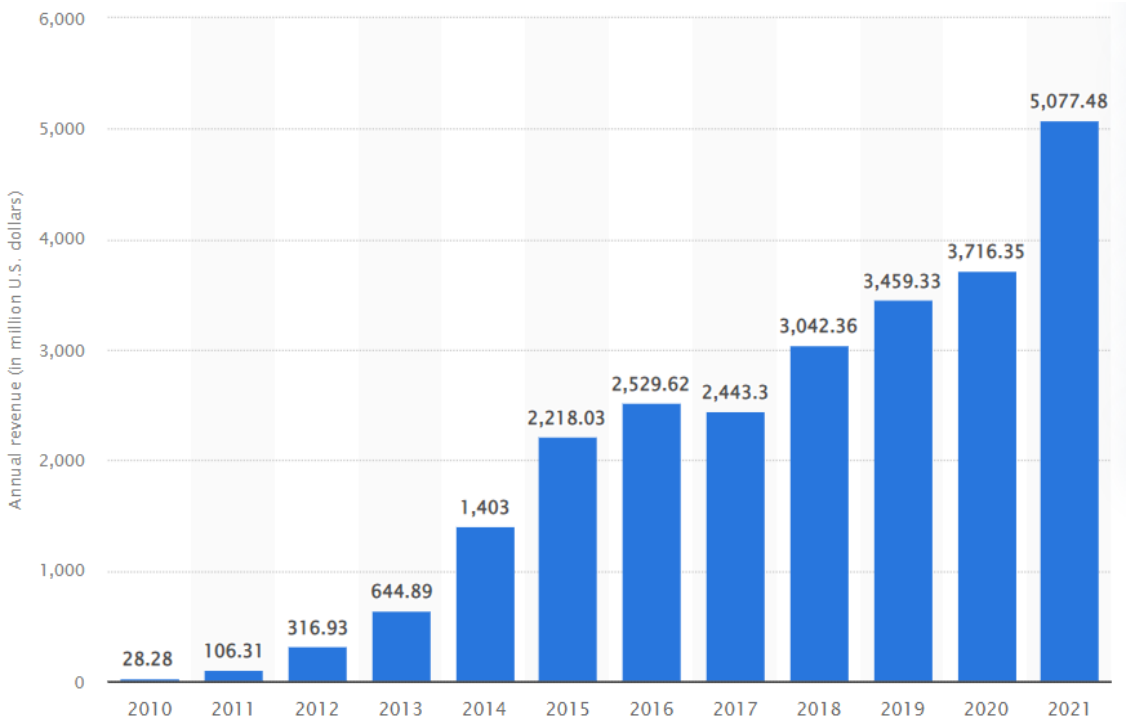
Según Restrepo Betancur et al. (2020), Twitter es una red social de amplio uso a nivel mundial, caracterizada por suministrar información en tiempo real, permitiendo la interacción entre personas en relación con un tema en particular.

Además, Raamkumar et al. (2018) menciona que Twitter puede considerarse como una de las redes sociales en línea contemporáneas y populares. Como sistema de microblogging, es relevante tanto en el ámbito de la comunicación privada como en el público.

Sugimoto et al. (2017) comentan que el éxito de una plataforma de redes sociales como Twitter, radica en su capacidad para atraer a personas de diferentes dominios y ubicaciones geográficas. Los académicos e investigadores de la comunidad científica también están interesados en las redes sociales debido a sus diversos beneficios.

Según Dixon (2022), en el período fiscal informado más recientemente, los ingresos mundiales totales de la red social Twitter ascendieron a más de cinco mil millones de dólares estadounidenses, frente a los 3720 millones de dólares estadounidenses del año anterior. Este es el mayor aumento en los ingresos anuales que la plataforma de microblogging ha visto en los últimos años.

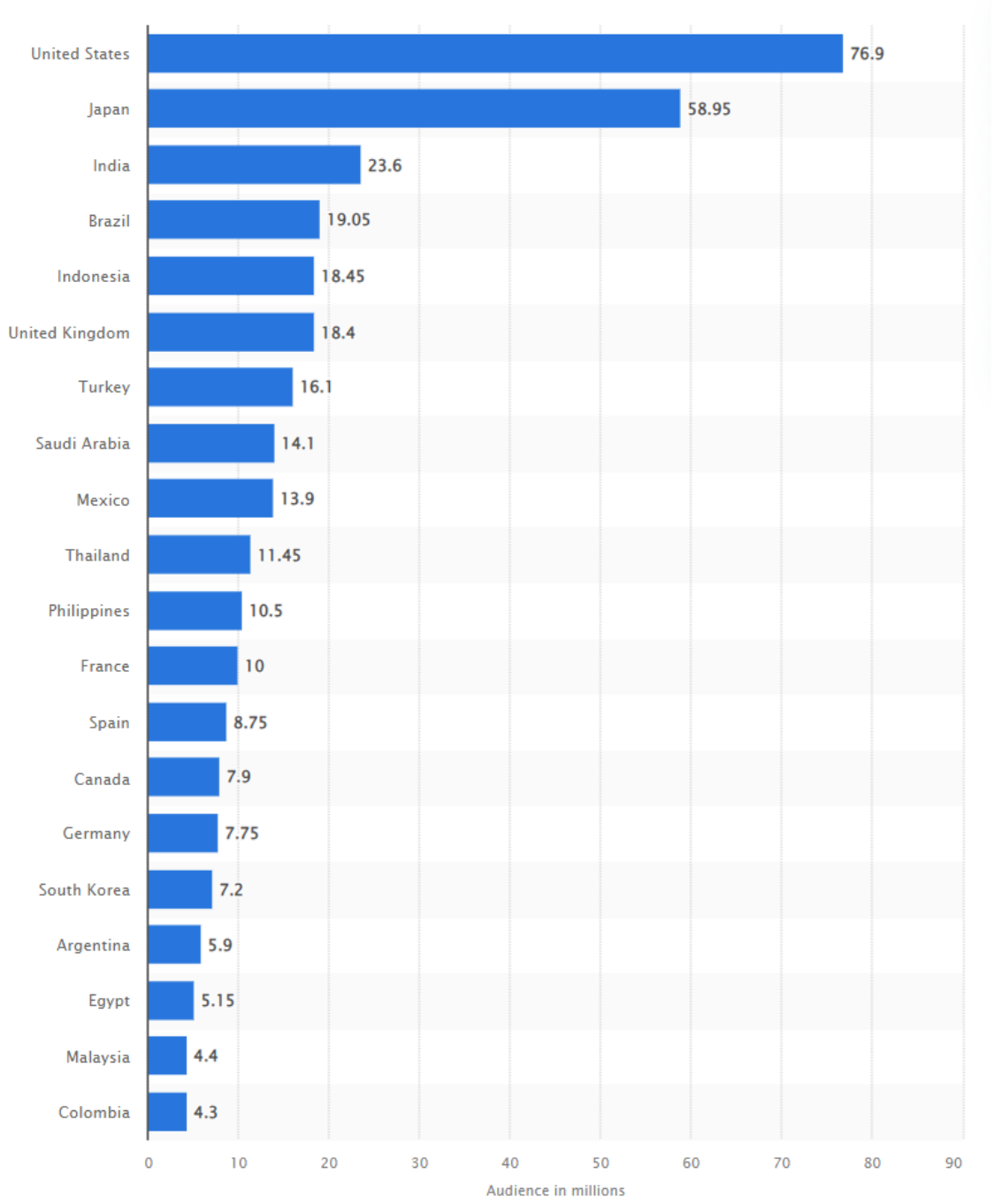
Figura 2: Ingresos mundiales de Twitter de 2010 a 2021 (en millones de dólares estadounidenses).



Fuente: (Dixon, 2022)

La red social Twitter es particularmente popular en los Estados Unidos, donde a partir de enero de 2022, el servicio de microblogging tenía un alcance de audiencia de 76,9 millones de usuarios. Japón e India ocuparon el segundo y tercer lugar con más de 58 y 23,6 millones de usuarios respectivamente. (Dixon, 2022)

Figura 3: Países líderes según el número de usuarios de Twitter a partir de enero de 2022 (en millones)



Fuente: (Dixon, 2022)

2.2.3.1. Terminología Útil de Twitter

Existen distintos términos comúnmente usados por los usuarios de esta red social, algunos de los cuales se detallan a continuación:

Tabla 3: Terminología de Twitter

Término	Definición
Follow	La suscripción a una cuenta de Twitter se denomina "seguir" o “follow” en inglés, de esta forma se pueden ver los nuevos tweets de la cuenta que se está siguiendo apenas se publiquen.
Unfollow	Hace referencia a la acción de dejar de seguir una cuenta de Twitter, de tal forma que ya no se muestren los tweets de la cuenta en mención
Retweet (RT)	Se denomina Retweet a un Tweet que se reenvía a seguidores, es una de las funcionalidades más utilizadas de la plataforma para compartir noticias y demás contenido interesante publicado en Twitter manteniendo la atribución original.
Reply	Se utiliza para responder el Tweet de otra persona, además, es posible ver el conteo de respuestas directas que indica la cantidad total de respuestas que recibió el Tweet.
Mentions	Es posible mencionar a otras cuentas de Twitter incluyendo el signo @ seguido directamente por el nombre de usuario.

Fuente: (Twitter, 2022)

2.2.4. Análisis de Sentimientos

Rosenbrock et al. (2019) explican que el análisis del sentimiento o la minería de opinión es el estudio computacional de opiniones, sentimientos y emociones expresadas a través de un texto. En general, las opiniones pueden centrarse en un producto, un servicio, un individuo, una organización, un evento o un tema.

Para Paez Guarnizo & Monroy (2020), el procesamiento del lenguaje natural es el que realiza el seguimiento del estado de ánimo de los usuarios frente a un tema o un

producto en particular. El análisis de sentimientos es la construcción de un modelo para recoger y categorizar las diferentes opiniones.

2.2.4.1. Aplicaciones del Análisis de Sentimientos

Según Pauli & Soliani (2019) algunas de las aplicaciones del análisis de sentimientos podrían ser las siguientes:

- **Valoración de opinión de productos y servicios:** Mediante esta técnica es posible que las empresas puedan conocer la opinión de los usuarios acerca de sus productos sin necesidad de llevar a cabo estudios tradicionales como encuestas de satisfacción. De esta forma, las empresas pueden conocer en cualquier momento si sus productos son del agrado de los usuarios y, en caso negativo, poder replantear estrategias en el menor tiempo posible otorgando así ventajas competitivas.
- **Posicionamiento de publicidad on-line:** Los anunciantes de determinados productos podrían requerir que sus anuncios fuesen publicados solo en sitios web en donde se expresen conceptos positivos, huyendo de aquellas páginas en donde los textos expresen sentimientos negativos.
- **Corrección de opinión:** Es habitual que los usuarios expresen su opinión en sitios de compras online indicando, además de una reseña, una puntuación.
- **Mejora de los sistemas de recomendación de productos:** En base a las opiniones de los usuarios, una tienda online podrá priorizar los productos que ofrece en base a dichas opiniones o no recomendar aquellos cuya opinión general sea negativa.

- **Reputación política:** El análisis de sentimientos demuestra un fuerte potencial para conocer la opinión de la gente sobre un determinado partido político o un candidato.
- **Análisis del mercado financiero:** A partir de la información contenida en páginas web, foros y redes sociales sobre una empresa o un tema, es posible prever cuál será su evolución en el mercado financiero a partir del valor agregado de la polaridad a todas las opiniones encontradas.

2.2.4.2. Niveles de Análisis de Sentimientos

En el trabajo de Sobrino Sande (2018) se menciona que el análisis de sentimientos se puede llevar a cabo a tres niveles distintos según la granularidad, profundidad y detalle requeridos:

- **Análisis a nivel de documento:** El análisis de sentimientos demuestra un fuerte potencial para conocer la opinión de la gente sobre un determinado partido político o un candidato.
- **Análisis a nivel de oración:** El análisis de sentimientos demuestra un fuerte potencial para conocer la opinión de la gente sobre un determinado partido político o un candidato.
- **Análisis a nivel de aspecto y entidad:** el análisis de sentimientos demuestra un fuerte potencial para conocer la opinión de la gente sobre un determinado partido político o un candidato.

2.2.4.3. VADER (Valence Aware Dictionary for sEntiment Reasoning)

VADER es un léxico simple y un enfoque basado en reglas que no solo brinda polaridad, sino que también indica qué tan positiva o negativa es una oración. Sin

embargo, en comparación con las técnicas sofisticadas de aprendizaje automático, la simplicidad de VADER conlleva varias ventajas. (Bhagya Laxmi et al., 2020)

En el trabajo de Hutto & Gilbert (2014), se muestra que el léxico VADER funciona excepcionalmente bien en el dominio de las redes sociales. El coeficiente de correlación muestra que VADER (r = 0,881) se desempeña tan bien como los evaluadores humanos individuales (r = 0,888) en la verdad del terreno coincidente (media agregada del grupo de 20 evaluadores humanos para la intensidad del sentimiento de cada tweet). Sorprendentemente, cuando inspeccionamos más a fondo la precisión de la clasificación, vemos que VADER (F1 = 0,96) incluso supera a los evaluadores humanos individuales (F1 = 0,84) al clasificar correctamente el sentimiento de los tweets en clases positivas, neutrales o negativas.

2.2.5. Modelo Predictivo

Es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones (Microsoft, 2022). Podemos decir también que, es un esquema teórico de una realidad compleja con el fin de entender su funcionamiento mediante el uso de técnicas estadísticas para predecir comportamientos futuros que funciona por medio del análisis de datos históricos y actuales. (Gartner, 2022)

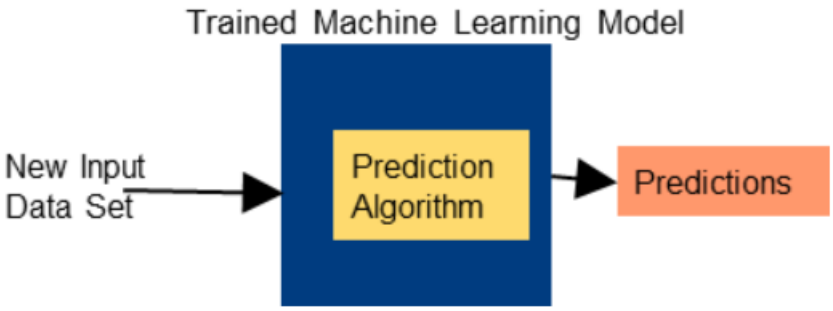
Según Subramanyam (2019), los modelos de predicción de aprendizaje automático se desarrollan en dos pasos. Primero se entrena un modelo de aprendizaje automático; en este paso, los datos de entrada (input), los resultados históricos asociados con el input y un algoritmo de entrenamiento se utilizan para llegar iterativamente al algoritmo de predicción.

Figura 4: Modelo de aprendizaje automático: paso de entrenamiento.

Fuente: (Subramanyam, 2019)

Segundo, el paso de predicción, el modelo ya entrenado utiliza el algoritmo de predicción al que se llegó en el paso anterior (entrenamiento) para transformar nuevas entradas (input) en predicciones.

Figura 5: Modelo de aprendizaje automático entrenado.



Fuente: (Subramanyam, 2019)

2.2.6. Long Short-Term Memory (LSTM)

La red neuronal "LSTM" es un tipo especial de red neuronal recurrente (RNN, por sus siglas en inglés) capaz de aprender dependencias a largo plazo. Las LSTM fueron diseñadas específicamente para evitar el problema de dependencia a largo plazo. Su característica principal es recordar información durante largos periodos de tiempo, esto hace que su aprendizaje sea más fácil. (Marín Vilca & Pineda Torres, 2019)

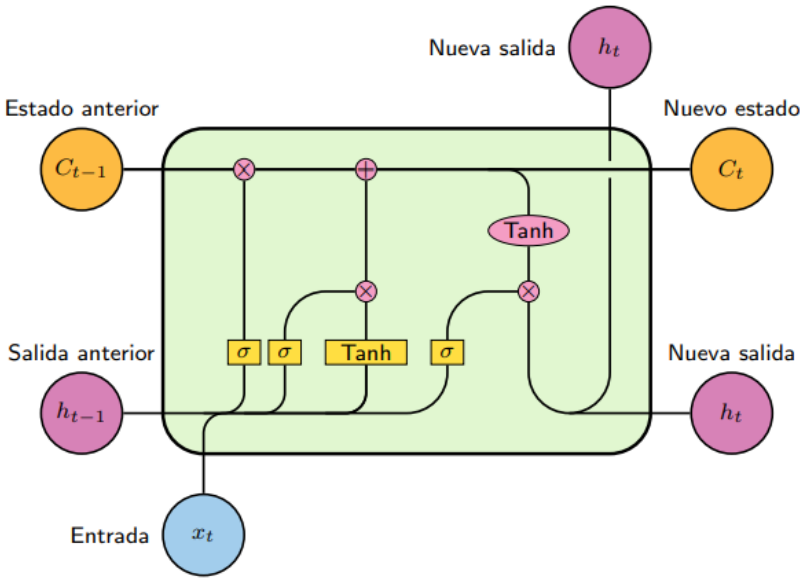
Pérez Sanjuán (2019) menciona que las redes LSTM se componen de una secuencia de unidades de unidades encadenadas, pero que, aunque la estructura de cada

unidad es idéntica, los valores que almacenan en forma de vector son diferentes. Por cada instante de tiempo t, un conjunto de vectores es procesado:

- Una puerta para descartar: $\boldsymbol{f_t = \sigma(W_f \times x_t + U_f \times h_{t-1} + b_f)}$
- Una puerta de entrada: $\boldsymbol{i_t = \sigma(W_i \times x_t + U_i \times h_{t-1} + b_i)}$
- Un vector de nuevos candidatos para el estado de la unidad: $\boldsymbol{C'_t = \tanh(W_C \times x_t + U_C \times h_{t-1} + b_C)}$
- Una puerta de salida: $\boldsymbol{o_t = \sigma(W_o \times x_t + U_o \times h_{t-1} + b_o)}$
- La memoria de la unidad: $\boldsymbol{C_t = i_t \times C'_t + f_t \times C_{t-1}}$
- Una capa de salida oculta: $\boldsymbol{h_t = o_t \times \tanh(C_t)}$

Donde $W_f,\ U_f,\ W_i,\ U_i$, son matrices de pesos y b_f, b_i, b_C, b_0 vectores de sesgos.

Figura 6: Unidad LSTM.



Fuente: (Pérez Sanjuán, 2019)

García López (2018) explica que el bloque de memoria contiene una célula de memoria y tres puertas multiplicativas: una de entrada, una de salida y una de olvido. La entrada al bloque de memoria se multiplica por la activación de la puerta de entrada. La salida es multiplicada por la puerta de salida y el valor previo de la célula de memoria es

multiplicada por la puerta de olvido. Las puertas se ocupan del flujo de información dentro y fuera de la célula de memoria. Las puertas son las que permiten a las redes LSTM modelizar las dependencias a largo plazo. Las tres puertas aprenden sobre qué información deben almacenar, cuanto tiempo lo deben hacer y cuando deben utilizarla. La puerta de entrada aprende a proteger a la célula de entradas irrelevantes. La puerta de olvido decide qué información se va a mantener en la célula y cuánto tiempo. La puerta de salida regula la información que se puede obtener de la célula evitando así la generación de información irrelevante.

2.3. GLOSARIO DE TÉRMINOS BÁSICOS

2.3.1. Bitcoin

Según Izquierdo Cervera (2018), el Bitcoin es una moneda virtual e intangible que fue creada en 2009 por una o varias personas que usan el seudónimo de Satoshi Nakamoto. Esta moneda digital puede intercambiarse en los mercados financieros por euros o dólares como si de una moneda fiat se tratase.

2.3.2. Bitcoiner

Se denomina así a todo defensor de la cadena de bloques (blockchain) de Bitcoin ya sea como validador, minero, desarrollador, validador, inversor o usuario. (PCMag, 2022)

2.3.3. Volatilidad

Puig (2022) menciona que la volatilidad es la variabilidad de la rentabilidad de una acción respecto a su media en un periodo de tiempo determinado.

2.3.4. Aprendizaje Automático

Valdez Alvarado (2019) menciona que el Aprendizaje Automático es un campo en la Inteligencia Artificial, donde las máquinas pueden "aprender" de sí mismas, sin ser explícitamente programadas por los seres humanos. Analizando datos pasados llamados "datos de entrenamiento", el modelo de Aprendizaje Automático forma patrones y usa estos patrones para aprender y hacer predicciones futuras.

2.3.5. Twitter

Según Berrón Ruiz & Régil López (2018), Twitter es una plataforma que permite utilizar 280 caracteres para escribir Tweets (publicaciones) de distinta índole. En esta red social es posible seguir a otros usuarios como también tener seguidores, esto facilita la interacción entre usuarios gracias a la opción de enviar y responder mensajes directos.

2.3.6. Toma de Decisiones

Arévalo Ascanio & Estrada López (2017) explican que la toma de decisiones es uno de los conceptos que se asocian continuamente al contexto empresarial y económico. Hace referencia a las situaciones en las que se debe escoger entre una pluralidad de opciones que llevan a distintos resultados ya sean positivos o negativos.

2.3.7. Análisis de Sentimientos

Para Rosenbrock et al. (2019), el análisis del sentimiento o la minería de opinión es el estudio computacional de opiniones, sentimientos y emociones expresadas a través de un texto.

2.3.8. LSTM

Según D. Lee et al. (2017), Long Short-Term Memory (LSTM) es una arquitectura de red neuronal recurrente (RNN) específica que se diseñó para modelar

secuencias temporales y sus dependencias de largo alcance con mayor precisión que las RNN convencionales.

2.3.9. Modelos Predictivos

Espino Timón (2017) comenta que los modelos predictivos son modelos de la relación entre el rendimiento específico de una unidad en una muestra y uno o más atributos o características conocidos de la unidad, con el fin de evaluar la probabilidad de que una unidad similar en una muestra diferente exhiba un comportamiento específico.

CAPITULO III

MATERIALES Y MÉTODOS

3.1. POBLACIÓN Y MUESTRA DE INVESTIGACIÓN

3.1.1. Población

La población de estudio ha sido definida por la disponibilidad de la información, es por eso que se utilizará el dataset “Bitcoin tweets - 16M tweets” disponible en la plataforma (Kaggle, 2022). Por lo tanto, la población comprende 16,000,000 de tweets desde 01-01-2016 hasta 23-11-2019.

3.1.2. Muestra

El tipo de selección de muestra es no probabilístico, depende del proceso de toma de decisiones del investigador y ha sido definida por las interacciones relacionadas al Bitcoin en la red social Twitter pertenecientes a la comunidad de habla inglesa.

a) Según Mena Roa (2021), Nigeria es el país con mayor cantidad de usuarios de Bitcoin, el 42% de los nigerianos encuestados en línea aseguran poseer o utilizar criptodivisas en 2021, la tasa más alta de los 56 países incluidos en la encuesta. Considerando que el idioma oficial de Nigeria es el inglés, podemos decir que la comunidad de habla inglesa es la más activa en el criptomercado, esto sin considerar a otros países de habla inglesa como Estados Unidos que tienen mucha presencia e influencia en Twitter y el criptomercado.

b) Según Osman (2021), Estados Unidos es el lugar con mayor cantidad de usuarios de Twitter seguido por Japón y el Reino Unido.

c) El presente trabajo utilizará el clasificador de sentimientos VADER el cual fue realizado originalmente para texto en inglés. (Hutto & Gilbert, 2014)

d) Las principales cuentas activas más influyentes de Twitter que hablan sobre el Bitcoin y el criptomercado publican contenido netamente en inglés. (CoinTracking, 2021)

3.2. DISEÑO METODOLÓGICO DE LA INVESTIGACIÓN

3.2.1. Tipo y Diseño de Investigación

Esta investigación de acuerdo a la caracterización del problema es de tipo correlacional (Hernández Sampieri & Mendoza Torres, 2018), ya que se estudia la relación e influencia de las interacciones de Twitter relacionadas al Bitcoin sobre el comportamiento de esta misma criptodivisa.

El diseño de la investigación es no experimental longitudinal de tendencia (Hernández Sampieri & Mendoza Torres, 2018) ya que se analizarán cambios al paso del tiempo del sentimiento promedio de la comunidad crypto en Twitter respecto al Bitcoin y su relación con el comportamiento del Bitcoin (fluctuación de precio).

3.3. MATERIALES EMPLEADOS

3.3.1. Recursos de Hardware

Tabla 4: Hardware utilizado

HARDWARE UTILIZADO PARA EL DESARROLLO DEL PROYECTO
01 Laptop HP Pavilion Gaming, Intel (R) Core (TM) i5-10300H CPU @ 2.50GHz 2.50 GHz, 8.00 RAM, WINDOWS 11 HOME 21H2
01 SATA Solid State Disk 1TB – Western Digital Blue

Elaboración propia

3.3.2. Recursos de Software

Tabla 5: Software utilizado

SOFTWARE UTILIZADO PARA EL DESARROLLO DEL PROYECTO
Python 3.9.7
Jupyterlab 3.4.2
Google Colab

Elaboración propia

Tabla 6: Librerías utilizadas

LIBRERIAS UTILIZADAS
Pandas
Numpy
Matplotlib
Time
Seaborn
Math
Sklearn
Keras

Elaboración propia

3.3.3. Presupuesto

Tabla 7: Presupuesto del proyecto

Descripción	Unidad de medida	Costo Unitario (S/.)	Cantidad	Costo total (S/.)
A. COSTOS DIRECTOS				**132.57**
Colab Pro	Plan/mensual	40.19	3	120.57

(continuación…)

Papel Bond	Millar	12.00	1	12.00
B. COSTOS				**19.58**
INDIRECTOS				
Gastos Adicionales	15%			19.58
TOTAL				**152.15**

Elaboración propia

3.4. METODOLOGÍA Y PROCEDIMIENTO

El presente trabajo de investigación "Modelo predictivo aplicando análisis de sentimientos en Twitter para determinar el comportamiento de la criptodivisa BITCOIN" por la modalidad corresponde a un proyecto de data science por lo que se aplicará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) ya que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, dicha metodología consiste en seis fases que se detallan a continuación:

3.4.1. Fases de la Metodología

La metodología CRIPS-DM está comprendida por seis fases que se explicarán a continuación según lo explicado por (SNGULAR, 2022):

Fase I. Comprensión del negocio

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Fase II. Comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los

problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Fase III. Preparación de los datos

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Fase IV. Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Fase V. Evaluación

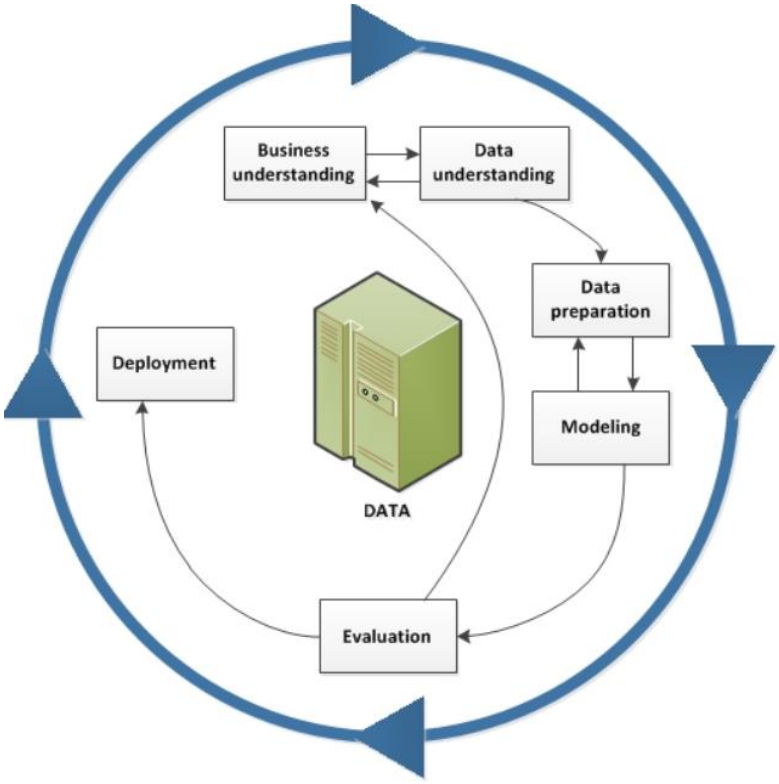
En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.

Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Fase VI. Despliegue

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

Figura 7: Metodología CRISP-DM



Fuente: (IBM, 2021)

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. RESULTADOS

4.1.1. Comprensión del Negocio

Como ya se mencionó, este proyecto busca predecir el comportamiento del Bitcoin aplicando análisis de sentimientos en Twitter considerando variables de influencia o alcance. Para ello, es fundamental cumplir con los objetivos planteados:

- **Recopilar y preprocesar la data histórica de Bitcoin y los tweets referidos al Bitcoin:** En este caso, necesitaremos la data histórica del Bitcoin, es decir, su comportamiento (fluctuación del precio) hora a hora a lo largo de los años (2016 - 2019). También es necesaria la data sobre Bitcoin procedente de Twitter, es decir, los tweets que hablan sobre el Bitcoin en el mismo periodo de tiempo (2016 - 2019). Este rango de fechas (2016 – 2019) es importante ya que en ese rango se encuentran los tweets del dataset perteneciente a Kaggle con el cual se trabajará. Una vez extraída toda la información pertinente, se realizará el preprocesamiento de los datos. En el caso de la data histórica del Bitcoin (comportamiento) obtenida de Gemini (2022), es necesario seleccionar solo algunas columnas como “Date”, “Close” y “Volume”, y considerar solo los registros pertenecientes al intervalo de fechas ya mencionado (2016 - 2019). Por otro lado, para la data sobre Bitcoin extraída de Twitter, se realizará un procedimiento más complejo que se detallará en la fase de “preparación de los datos”.

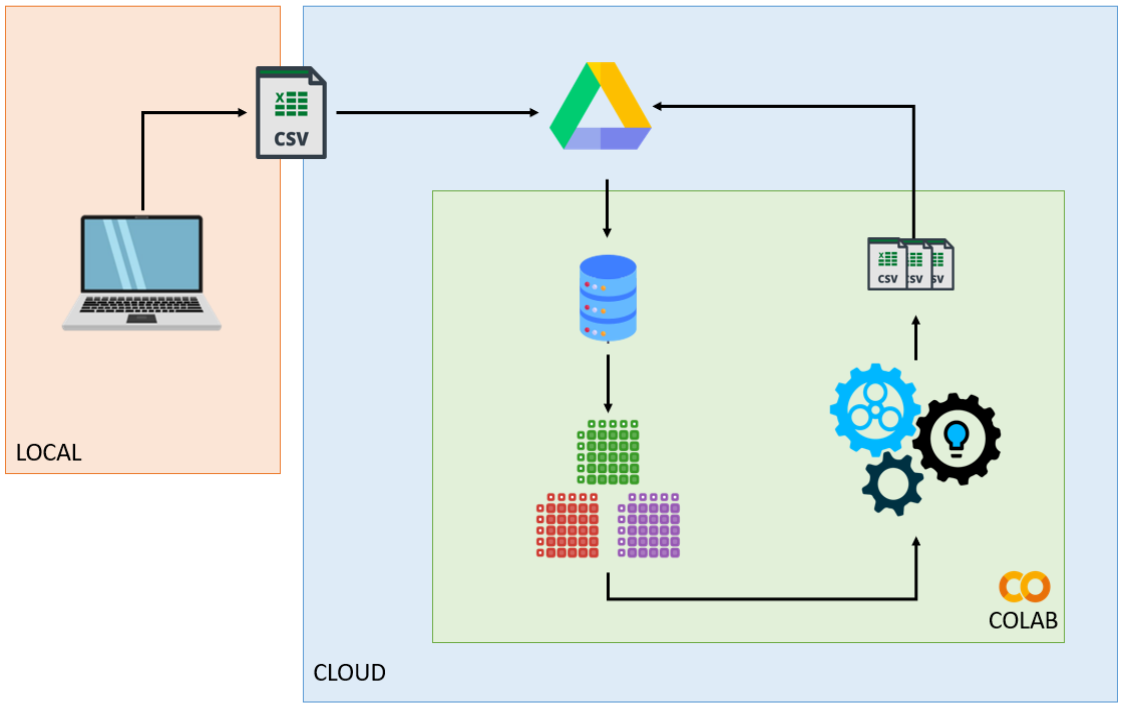
- **Aplicar el análisis de sentimientos y la selección de características:** El resultado del punto anterior es un dataframe mucho más organizado según los requerimientos del estudio. A este dataframe se le aplicará un análisis de sentimientos utilizando VADER al cual se le agregará un diccionario de léxicos con expresiones comúnmente utilizadas por la comunidad de Bitcoin. Para este estudio solamente se utilizará el valor de "compound" que es una de las salidas del clasificador y la más importante en este caso. Además, se realizará la selección de características utilizando la correlación de Spearman, las cuales se utilizarán para implementar el modelo de predicción. También se considerarán las variables de alcance a modo de propuesta para medir la influencia de cada tweet y cuantificar dicha influencia en el criptomercado considerando grupos en rangos de 1 hora.
- **Implementar el modelo de predicción con sus diferentes configuraciones:** Se seleccionarán las entradas para el modelo predictivo y se procederá a convertir la data para que pueda ser utilizada por la red LSTM, esto significa convertirla de una serie de datos de tiempo en una secuencia supervisada, una matriz 3D con variables normalizadas. Se probarán 3 configuraciones del modelo para predecir 1 hr., 6 hrs. y 12hrs. a futuro bajo un enfoque de inversión a corto plazo (scalping) ya que según Sattarov et al. (2020), el resultado más satisfactorio entre los enfoques clásicos de inversión es el trading de especulación (scalping). El modelo considerará 3 horas en el pasado para predecir el valor futuro. La partición del dataset corresponderá a 80 – 20, es decir, 80% del dataset se usará en el entrenamiento del modelo y el 20% restante se usará para testear el modelo. Las características del modelo predictivo se detallarán en la fase de modelado.

- **Evaluar y comparar el performance de las diferentes configuraciones del modelo predictivo:** Para evaluar el performance de cada modelo se calcularán dos medidas de error: RMSE y MAPE para luego comparar las 3 configuraciones del modelo y seleccionar aquella configuración con mejores resultados.

4.1.2. Recopilación y Comprensión de los Datos

En esta etapa se diseñó una arquitectura de procesamiento en Cloud para procesar toda la data según las necesidades del proyecto, que comprende desde la carga de los archivos .csv a Google Drive (almacenamiento en nube) hasta el resultado de su procesamiento respectivo utilizando Google Colab.

Figura 8: Procesamiento en Cloud.



Elaboración propia

4.1.2.1. Data Histórica del Bitcoin

Es importante conocer la información histórica del Bitcoin y como ha ido fluctuando en el tiempo. Como se mencionó en la fase anterior, se extrajo la información horaria de la plataforma Gemini como se muestra a continuación:

Figura 9: Extracto de la data histórica del Bitcoin.

	Unix Timestamp	Date	Symbol	Open	High	Low	Close	Volume
0	1649131200000	2022-04-05 04:00:00	BTCUSD	46666.70	46767.72	46624.50	46701.12	11.205398
1	1649127600000	2022-04-05 03:00:00	BTCUSD	46634.58	46767.72	46593.13	46666.70	14.430378
2	1649124000000	2022-04-05 02:00:00	BTCUSD	46641.93	46748.86	46576.87	46634.58	8.162497
3	1649120400000	2022-04-05 01:00:00	BTCUSD	46519.95	46695.82	46463.62	46641.93	23.637272
4	1649116800000	2022-04-05 00:00:00	BTCUSD	46596.97	46651.12	46431.31	46519.95	23.032921
5	1649113200000	2022-04-04 23:00:00	BTCUSD	46699.39	46886.58	46554.81	46596.97	94.262109
6	1649109600000	2022-04-04 22:00:00	BTCUSD	46477.43	46846.71	46396.00	46699.39	20.572959
7	1649106000000	2022-04-04 21:00:00	BTCUSD	46310.69	46582.27	46193.41	46477.43	69.927108
8	1649102400000	2022-04-04 20:00:00	BTCUSD	45951.49	46385.99	45730.42	46310.69	84.162701
9	1649098800000	2022-04-04 19:00:00	BTCUSD	45525.21	45959.95	45525.21	45951.49	60.694165

Elaboración propia

El dataframe contiene 8 columnas de las cuales la columna “Date” será de vital importancia ya que toda la información de Twitter se agrupará según la fecha y hora correspondiente a esa columna. A continuación, se detallan las columnas presentes en el dataframe:

Tabla 8: Descripción de las columnas de la data histórica del Bitcoin

Nro.	Variable	Descripción
1	Unix Timestamp	Esta es la marca de tiempo de Unix o también conocida como "Epoch Time".
2	Date	Esta marca de tiempo se convierte a la hora estándar de NY EST

(continuación…)

3	Symbol	El símbolo al que se refieren los datos de la serie temporal
4	Open	Este es el precio de apertura del período de tiempo.
5	High	Este es el precio más alto del período de tiempo.
6	Low	Este es el precio más bajo del período de tiempo.
7	Close	Este es el precio de cierre del período de tiempo.
8	Volume	Este es el volumen en el Ccy negociado. Es decir. Para BTC/USD, esto es en cantidad de BTC

Elaboración propia

Para poder ver de una forma más gráfica como ha ido fluctuando el precio del Bitcoin de forma horaria en el transcurso de los años (comportamiento) utilizaremos la biblioteca Matplotlib de Python.

Figura 10: Comportamiento del Bitcoin desde 2016 a 2019

Elaboración propia

Es posible obtener información más detallada del dataset utilizando la función "info()" de Python, lo que nos proporcionará la cantidad total de registros existentes en el dataset y el tipo de dato correspondiente a cada columna antes mencionada.

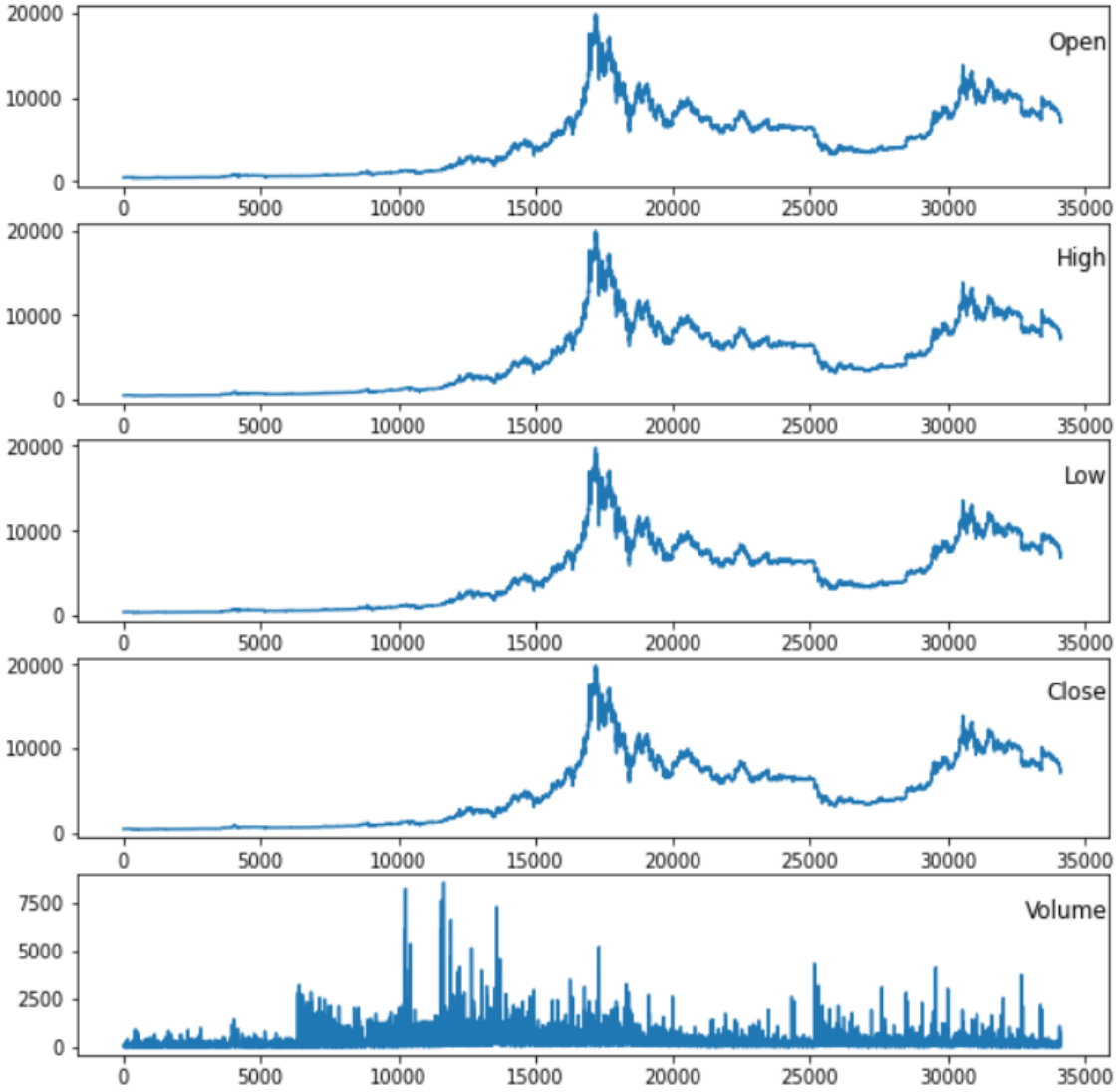
Figura 11: Información adicional del dataset correspondiente a la data histórica del Bitcoin.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56893 entries, 0 to 56892
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unix Timestamp  56893 non-null  int64
 1   Date            56893 non-null  datetime64[ns]
 2   Symbol          56893 non-null  object
 3   Open            56893 non-null  float64
 4   High            56893 non-null  float64
 5   Low             56893 non-null  float64
 6   Close           56893 non-null  float64
 7   Volume          56893 non-null  float64
dtypes: datetime64[ns](1), float64(5), int64(1), object(1)
memory usage: 3.5+ MB
```

Elaboración propia

También se realizó una representación gráfica de 5 de las variables o columnas presentes en este dataset ("Open", "High", "Low", "Close" y "Volume") para ver su comportamiento.

Figura 12: Representación gráfica de las variables "Open", "High", "Low", "Close" y "Volume".



Elaboración propia

4.1.2.2. Data Extraída de Twitter

Por lo general la extracción de datos de Twitter es complicada ya que las API's existentes solo permiten extraer tweets de un periodo de tiempo bastante corto y una cantidad limitada de los mismos, por lo que realizar este proyecto habría sido bastante complicado. Felizmente existe un dataset en la plataforma Kaggle disponible para cualquier usuario que dispone de 16,889,765 de tweets que comprende desde el año 2016 hasta el año 2019.

Figura 13: Extracto de la data extraída de Twitter.

	Id	User	Fullname	Url	Timestamp	Replies	Likes	Retweets	Text
0	1.132977e+18	KamdemAbdiel	Abdiel kamdem	NaN	2019-05-27 11:49:14+00	0	0	0	È appena uscito un nuovo video! LES CRYPTOMONN...
1	1.132977e+18	bitcointe	Bitcointe	NaN	2019-05-27 11:49:18+00	0	0	0	Cardano: Digitize Currencies; EOS https://t.co...
2	1.132977e+18	3eyedbran	Bran - 3 Eyed Raven	NaN	2019-05-27 11:49:06+00	0	2	1	Another Test tweet that wasn't caught in the s...
3	1.132977e+18	DetroitCrypto	J. Scardina	NaN	2019-05-27 11:49:22+00	0	0	0	Current Crypto Prices! \n\nBTC: $8721.99 USD\n...
4	1.132977e+18	mmursaleen72	Muhammad Mursaleen	NaN	2019-05-27 11:49:23+00	0	0	0	Spiv (Nosar Baz): BITCOIN Is An Asset & NO...
5	1.132977e+18	0nurTOKA	🐻🔫✴️🐂🚀	NaN	2019-05-27 11:49:25+00	0	0	0	#btc inceldiği yerden kopsun bakalım 17:00 ye ...
6	1.132977e+18	evilrobotted	evilrobotted	NaN	2019-05-27 11:49:25+00	0	0	0	@nwoodfine We have been building on the real #...
7	1.132977e+18	jabur_guilherme	Guilherme Jabur	NaN	2019-05-27 11:49:27+00	0	0	0	@pedronauck como investidor, vc é um ótimo dev...
8	1.132977e+18	INTBICON	億り人彼氏	NaN	2019-05-27 11:49:32+00	0	0	0	ブラジルはまぁ置いといてもドイツは存在感出してくるのかな。ロシアもマイニングなどで元気になる...
9	1.132977e+18	MLWright15	ML Wright	NaN	2019-05-27 11:49:32+00	0	0	0	CHANGE IS COMING...GET READY!!! Boom, Another ...

Elaboración propia

Como se ve en la Figura 13, el dataframe está compuesto por 9 columnas de las cuales las más importantes son: timestamp, replies, likes, retweets y text. A continuación, en la Tabla 9 se detallan las columnas presentes en el dataframe:

Tabla 9: Descripción de las columnas de la data extraída de Twitter.

Nro.	Variable	Descripción
1	Id	Corresponde al identificador del usuario que publicó el tweet.
2	User	Es el nombre de usuario o alias bajo el cual el usuario publica tweets.
3	Fullname	Es el nombre completo registrado del usuario que utiliza la red social.
4	Url	Es un campo opcional por si existe una dirección url en el tweet.
5	Timestamp	Corresponde a la fecha y hora de la publicación del tweet.
6	Replies	Es la cantidad de respuestas del tweet recibidas hasta el momento de la extracción de la data.

(continuación…)

7	Likes	Es la cantidad de "me gusta" correspondiente a la publicación o tweet recibidos hasta el momento de la extracción de la data.
8	Retweets	Es la cantidad de veces que otros usuarios volvieron a publicar un tweet hasta el momento de la extracción de la data.
9	Text	Corresponde al tweet en sí, es el texto que escribió algún usuario para expresar alguna idea u opinión que son relevantes para el proceso de análisis de sentimientos que se explicará a detalle más adelante.

Elaboración propia

Para conocer un poco más de este dataset, usamos la función "info()" en Python que nos mostrará algunas características adicionales como la cantidad exacta de registros y el tipo de dato correspondiente a cada columna.

Figura 14: Información adicional del dataset correspondiente a Twitter.

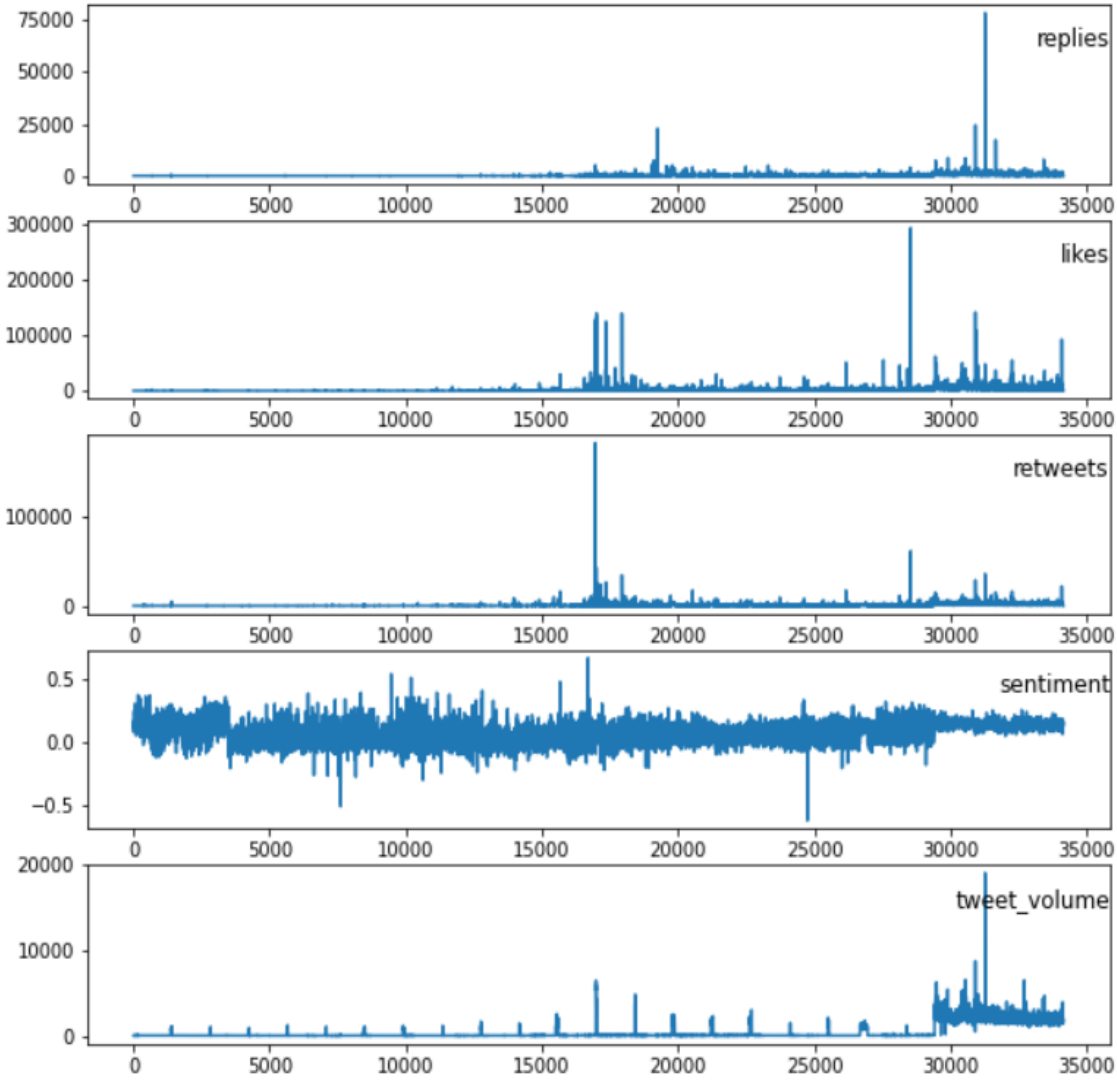
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16889765 entries, 0 to 16889764
Data columns (total 9 columns):
 #   Column     Dtype
---  ------     -----
 0   Id         float64
 1   User       object
 2   Fullname   object
 3   Url        object
 4   Timestamp  object
 5   Replies    int64
 6   Likes      int64
 7   Retweets   int64
 8   Text       object
dtypes: float64(1), int64(3), object(5)
memory usage: 1.1+ GB
```

Elaboración propia

Para la representación gráfica se seleccionaron solo 3 variables del dataset ("Replies", "Likes" y "Retweets"); sin embargo, se añadieron 2 variables extras

utilizando el mismo dataset (“Sentiment” y “Tweet_volume”) las que son de vital importancia para el estudio.

Figura 15: Representación gráfica de las variables "Replies", "Likes", "Retweets", "Sentiment" y "Tweet_volume".



Elaboración propia

4.1.3. Preparación de los Datos

4.1.3.1. Preprocesamiento de la Data Histórica del Bitcoin

Lo primero que se realizó fue la selección de los datos pertenecientes al intervalo de tiempo definido anteriormente que comprende desde el “2016-01-01 00:00:00” hasta “2019-12-31 23:00:00”. Para ello se definieron las variables “startdate” y “enddate” que

corresponden a las fechas de inicio y fin respectivamente. Estas variables se utilizaron como parámetros para crear la máscara de filtrado que se aplicó al dataset. Con el dataset resultante, fue necesario sobrescribir el índice con la fecha correspondiente a cada registro para posteriormente crear una nueva columna llamada “Date_merge” que se utilizó posteriormente para realizar la unión de los datasets.

Figura 16: Extracto del dataset final de la data histórica del Bitcoin.

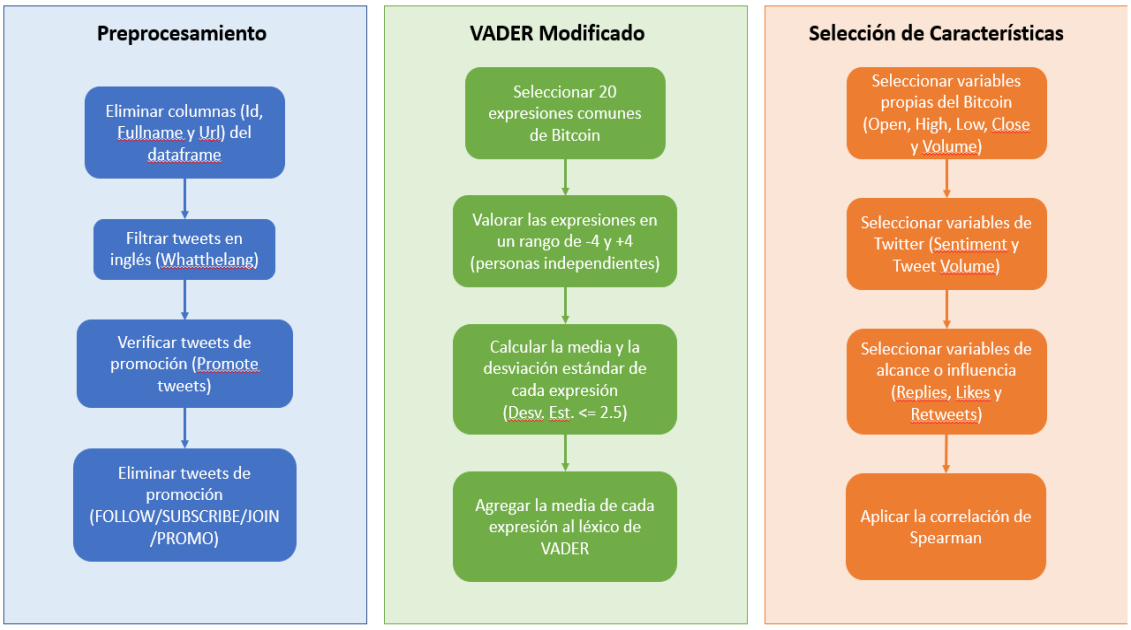
Date	Unix Timestamp	Date	Symbol	Open	High	Low	Close	Volume	Date_merge
2019-12-31 23:00:00	1577833200000	2019-12-31 23:00:00	BTCUSD	7173.53	7180.00	7156.01	7165.90	49.524838	2019-12-31 23:00:00
2019-12-31 22:00:00	1577829600000	2019-12-31 22:00:00	BTCUSD	7154.49	7180.58	7146.35	7173.53	93.290183	2019-12-31 22:00:00
2019-12-31 21:00:00	1577826000000	2019-12-31 21:00:00	BTCUSD	7146.92	7161.22	7146.92	7154.49	23.662080	2019-12-31 21:00:00
2019-12-31 20:00:00	1577822400000	2019-12-31 20:00:00	BTCUSD	7150.70	7160.00	7141.00	7146.92	61.640907	2019-12-31 20:00:00
2019-12-31 19:00:00	1577818800000	2019-12-31 19:00:00	BTCUSD	7141.39	7156.87	7133.69	7150.70	36.288595	2019-12-31 19:00:00

Elaboración propia

4.1.3.2. Preprocesamiento de la Data Extraída de Twitter

Para este caso de los tweets referidos a Bitcoin, el preprocesamiento viene acompañado de 2 bloques adicionales que comprenden la modificación del clasificador VADER y la posterior selección de características o variables. En la Figura 17 se muestra un diagrama de bloques de resumen correspondiente a este caso en específico.

Figura 17: Diagrama de bloques correspondiente a la data de Twitter.



Elaboración propia

Como se observó en la fase anterior, este dataset contiene 9 columnas (Id, User, Fullname, Url, Timestamp, Replies, Likes, Retweets y Text) de las cuales se eliminarán 3 (Id, Fullname y Url) ya que no aportan conocimiento importante para el estudio.

En este caso el preprocesamiento de la data fue mucho más complejo y lento ya que al tener tantos datos se requieren más recursos de computación para realizar las distintas tareas necesarias. Es por eso que se procedió a dividir la data en 4 grupos de 4 000 000 de registros cada uno a excepción del último bloque que tiene algunos registros adicionales, a estos bloques se les denominó “splits”.

Figura 18: Código para dividir la data en splits.

```
split_1 = tweet_data.iloc[:4000000]
split_2 = tweet_data.iloc[4000000:8000000]
split_3 = tweet_data.iloc[8000000:12000000]
split_4 = tweet_data.iloc[12000000:]
```

Elaboración propia

Cada split se guardó como un archivo “csv” sin considerar el índex ya que no es necesario por ahora, dichos archivos serán necesarios para continuar su procesamiento.

Figura 19: Splits guardados en archivos .csv

```
split_1.to_csv('../content/drive/MyDrive/BITCOIN_PREDICTION/tweet_data_split_1.csv', index=False)
split_2.to_csv('../content/drive/MyDrive/BITCOIN_PREDICTION/tweet_data_split_2.csv', index=False)
split_3.to_csv('../content/drive/MyDrive/BITCOIN_PREDICTION/tweet_data_split_3.csv', index=False)
split_4.to_csv('../content/drive/MyDrive/BITCOIN_PREDICTION/tweet_data_split_4.csv', index=False)
```

Elaboración propia

Anteriormente se explicó que para este estudio solo se considerarían los tweets en inglés por los motivos explicados en el Capítulo III: Materiales y métodos. Por lo tanto, se utilizó la librería "whatthelang" para reconocer el idioma de cada tweet y poder filtrar los tweets y mantener solo los tweets en inglés. Cabe mencionar que esta librería soporta hasta 176 lenguajes dentro de los cuales se encuentra el inglés. (Sangeeth & Manoj, 2017)

Figura 20: Código para reconocer el idioma de cada tweet utilizando whatthelang.

```
from whatthelang import WhatTheLang
wtl = WhatTheLang()
result = [wtl.predict_lang(row) for row in tweet_data_split_1['text']]
tweet_data_split_1['lang'] = result
```

Elaboración propia

Después de reconocer el idioma de cada tweet, resultó un dataframe con una columna adicional "lang" que hace referencia al idioma del tweet.

Figura 21: Extracto del dataset con el reconocimiento de idioma.

	user	timestamp	replies	likes	retweets	text	lang
0	KamdemAbdiel	2019-05-27 11:49:14+00	0	0	0	È appena uscito un nuovo video! LES CRYPTOMONN...	it
1	bitcointe	2019-05-27 11:49:18+00	0	0	0	Cardano: Digitize Currencies; EOS https://t.co...	en
2	3eyedbran	2019-05-27 11:49:06+00	0	2	1	Another Test tweet that wasn't caught in the s...	en
3	DetroitCrypto	2019-05-27 11:49:22+00	0	0	0	Current Crypto Prices! \n\nBTC: $8721.99 USD\n...	en
4	mmursaleen72	2019-05-27 11:49:23+00	0	0	0	Spiv (Nosar Baz): BITCOIN Is An Asset & NO...	en
5	0nurTOKA	2019-05-27 11:49:25+00	0	0	0	#btc inceldiği yerden kopsun bakalım 17:00 ye ...	tr
6	evilrobotted	2019-05-27 11:49:25+00	0	0	0	@nwoodfine We have been building on the real #...	en
7	jabur_guilherme	2019-05-27 11:49:27+00	0	0	0	@pedronauck como investidor, vc é um ótimo dev...	pt
8	INTBICON	2019-05-27 11:49:32+00	0	0	0	ブラジルはまぁ置いといてもドイツは存在感出してくるのかな。ロシアもマイニングなどで元気になる...	ja
9	MLWright15	2019-05-27 11:49:32+00	0	0	0	CHANGE IS COMING...GET READY!!! Boom, Another ...	en

Elaboración propia

Para seleccionar solamente los tweets en inglés, se realizó un código bastante práctico que cumple con lo que se necesita y seleccionamos 20 registros al azar como ejemplo para verificar el correcto funcionamiento de la selección.

Figura 22: Código para seleccionar solo los tweets en inglés.

```
en_tweet_data_split_1 = tweet_data_split_1[tweet_data_split_1["lang"] == 'en']
en_tweet_data_split_1.sample(20, random_state = 5)
```

Elaboración propia

Con todo lo anterior, queda un dataframe conteniendo solo los tweets en inglés que fueron guardados en archivos ".csv" para continuar con el procesamiento y limpieza correspondiente.

Figura 23: Extracto del dataset conteniendo solo tweets en inglés.

	user	timestamp	replies	likes	retweets	text	lang
2072403	BitcoinSpreads	2016-03-06 13:00:07+00	0	0	0	1 #BTC (#Bitcoin) quotes:\n$406.37/$408.00 #Bi...	en
510447	ProofofResearch	2019-04-14 20:26:55+00	1	3	2	2/ Since Bitcoin is UTXO-based, explorers have...	en
2142527	coinok	2016-08-01 00:30:04+00	0	1	4	1 BTC Price: BTC-e 613.986 USD Bitstamp 617.00...	en
400296	MercuryByHLE	2019-05-14 10:12:59+00	0	0	0	Want to capitalize on $8000 BTC and beyond? ht...	en
894933	CoinCapsAi	2019-05-20 15:10:03+00	0	0	0	The #big #slump: #full-blown #us-china #trade ...	en
1662196	airdroplite	2014-05-02 15:09:02+00	0	0	0	BTCTurk 935.01 TL Koinim 950.02 TL CampBx 445....	en
315338	BlockWatcher	2019-05-13 15:58:52+00	0	0	0	Mon May 13 19:58:12 2019 (1:15)\nUSD : 7793.41...	en
1475512	mtgoxtrades	2013-07-21 21:05:14+00	0	0	0	Sun Jul 21 23:00:27 2013 - Traded 0.02 BTC at ...	en
2240554	BTCticker	2017-06-15 18:00:07+00	0	1	0	One Bitcoin now worth $2333.94@bitstamp. High ...	en
1264878	Parijatsen3	2019-05-25 09:44:20+00	0	0	0	@cryptokanoon I am waiting for big bull run\nI...	en
1786221	ProjectCoin	2014-10-19 20:24:08+00	0	0	1	LIVE: Profit = $351.83 (9.34 %). BUY B9.76 @ $...	en
1033380	cryptoicobull	2019-05-22 09:05:47+00	0	0	1	How to Buy Pizza With Bitcoin Cash. #cryptocur...	en
414486	slavank85	2019-05-14 12:35:07+00	0	0	0	@XtLyman Looks like BTC is alive again . Sm...	en
3831488	ecassidy47	2017-12-08 16:50:51+00	0	2	1	Is there a petition to get #segwit activated o...	en
180542	neoprice_	2019-05-12 08:00:02+00	0	0	0	$NEO is now worth $9.65 (-0.70%) and 0.0013198...	en
1490547	mtgoxtrades	2013-08-21 22:15:07+00	0	0	0	Thu Aug 22 00:13:32 2013 - Traded 0.0111427 BT...	en
1148017	amruthasuri	2019-05-23 19:24:03+00	0	0	0	Lightweight, fast, and ready to mine! #CryptoC...	en

Elaboración propia

Es importante mencionar que todo el procedimiento anterior corresponde al split 1 pero se realizaron los mismos pasos para los otros 3 splits de la misma manera.

Como siguiente tarea se eliminarán aquellos tweets que promuevan "FOLLOW/SUBSCRIBE/JOIN/PROMO" ya que no expresan una opinión como tal. Para ello se considerará solamente la idea general de la regla 4 planteada por Badiola Ramos (2019), ya que este dataset tiene características distintas. Primero se identificará la cantidad de tweets que contienen y no contienen los términos antes mencionados.

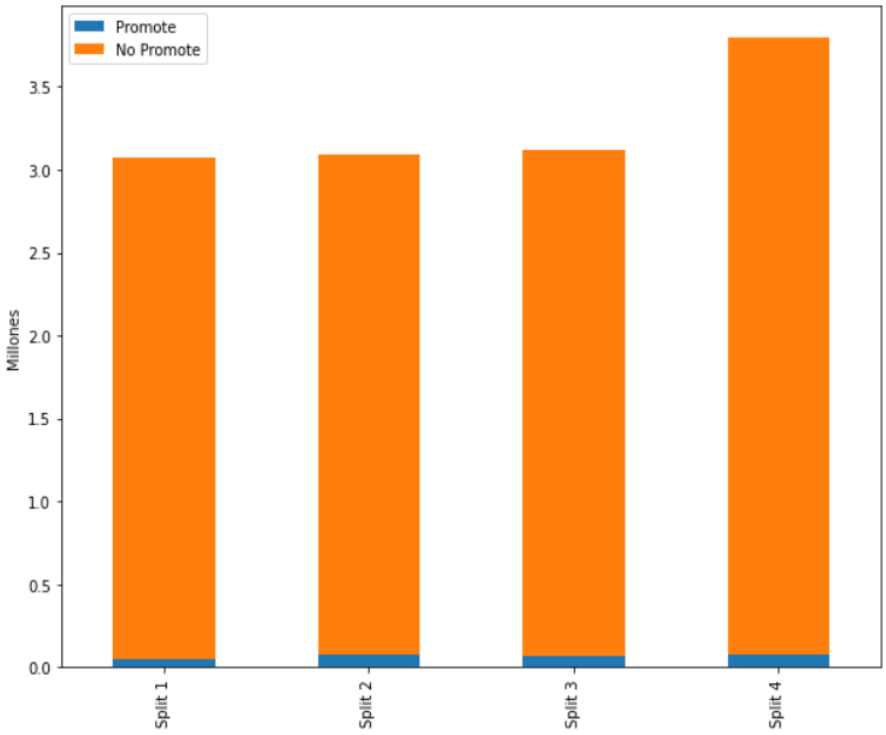
Tabla 10: Cantidad de tweets promote y no promote.

	SPLIT 1	SPLIT 2	SPLIT 3	SPLIT 4
Promote	48,294	73,086	66,446	72,581
No Promote	3,023,371	3,021,460	3,055,960	3,725,250
TOTAL	3,071,665	3,094,546	3,122,406	3,797,831

Elaboración propia

Para ver gráficamente la diferencia de cantidades entre los tweets considerados "Promote" y los "No Promote" utilizamos la librería Matplotlib que nos muestra el siguiente gráfico.

Figura 24: Relación de cantidad de tweets "Promote" y "No Promote".



Elaboración propia

Una vez identificados los “Promote” tweets, se procedió a extraerlos ya que no son necesarios para este estudio y podrían afectar el rendimiento del modelo predictivo más adelante. Para realizar la extracción y posterior eliminación de los “Promote” tweets, primero se creó una nueva columna en el dataset llamada “promote”. Si el tweet contiene alguno de los términos (FOLLOW/SUBSCRIBE/JOIN/PROMO) entonces en la nueva columna se registrarán los términos encontrados en la misma cantidad, por ejemplo, si el tweet contiene 2 veces la palabra “follow”, entonces en la nueva columna se registrará la palabra “follow” 2 veces. Por otro lado, si el tweet no contiene ninguno de los términos “Promote” entonces en la nueva columna simplemente se registrará un guion (-).

Figura 25: Ejemplos de "No Promote" tweets.

	user	timestamp	replies	likes	retweets	text	lang	promote
329407	dandolfa	2019-05-14 14:02:05+00	3	21	7	Interesting paper by Eric Budish on the Econom...	en	-
2242293	BTC_PRICE	2017-08-13 20:30:03+00	0	0	0	bitstamp: $4112.26\nbtce: $2546.78\nkraken: $4...	en	-
1673089	BitcoinSpreads	2016-03-19 21:00:10+00	0	0	0	1 #BTC (#Bitcoin) quotes:\n$405.79/$406.58 #Bi...	en	-
1105006	HoldYourColors	2019-05-27 07:13:07+00	0	0	0	@TecraCoin have created an Internet platform f...	en	-
1900905	btctickerbot	2018-01-14 02:00:01+00	0	0	0	BTC Price: 14212.11$, \nBTC Today High : 14449...	en	-
1480949	ThePriceOfBTC	2015-01-25 20:00:03+00	0	0	0	$249.98 #bitstamp; \n$244.00 #btce; \n\nInstan...	en	-
1739298	btcusd	2016-10-12 21:45:18+00	0	0	0	$634.24 at 23:45 UTC [24h Range: $628.76 - $64...	en	-
3034953	Fund_Media	2018-06-04 22:20:04+00	0	0	0	Learn more about SKS Media offerings > http://...	en	-
1019667	Blacktradelines	2019-05-25 18:06:07+00	0	0	0	Bitcoin community\n#BlackCommunity \nJoin this...	en	-
2285882	sangye	2017-06-15 02:52:35+00	2	68	23	When people "Flippen" to Ethereum only to real...	en	-

Elaboración propia

Figura 26: Ejemplos de "Promote" tweets.

	user	timestamp	replies	likes	retweets	text	lang	promote
0	bitcointe	2019-05-27 11:49:18+00	0	0	0	Cardano: Digitize Currencies; EOS https://t.co...	en	follow follow
25	cryptowhitewalk	2019-05-25 10:14:41+00	8	94	76	ohh and just incase anyone is interested in so...	en	follow
52	jasonclarktwit	2019-05-27 11:50:09+00	0	0	0	this channel called $DGB before it made 6x. th...	en	join
154	r_gadangholic	2019-05-27 11:51:45+00	0	0	0	2nd round 1000T free BTC hashrate giveaway,fol...	en	follow follow
252	Bullyena	2019-05-27 06:09:02+00	0	0	0	$ETH\ncould be following BTC as well with firs...	en	follow
...	...	...	...	...	...	...	...	...
3071507	ICryptoDesk	2018-06-04 02:30:46+00	0	1	0	Another Cryptocurrency Experiences 51% Attack,...	en	follow
3071515	BitCoin_Invest_	2018-06-04 02:30:17+00	1	44	8	Retweet and Like if you want to gain more foll...	en	follow

Elaboración propia

Para efectos de este estudio se seleccionaron solo los “No Promote” tweets y se procedió a guardar los nuevos dataframes en archivos “.csv”.

4.1.3.3. Análisis de Sentimientos

Como se mencionó anteriormente, este proyecto utilizó el clasificador de sentimientos VADER que es muy efectivo para las redes sociales como Twitter ya que los tweets tienen un límite de caracteres por lo que los usuarios utilizan abreviaciones de palabras, emoticonos y mucho argot (jerga).

4.1.3.3.1. VADER Modificado

A pesar de todas las bondades que ofrece el clasificador de sentimientos VADER, muchos de los términos o expresiones propias de la comunidad de Bitcoin no se encuentran consideradas dentro del gran léxico de VADER por lo que es imprescindible implementar un diccionario específico del dominio. Por este motivo es que esta investigación añadió expresiones comúnmente utilizadas en la comunidad Bitcoin para representar de mejor manera el sentimiento del criptomercado respecto al Bitcoin.

Para implementar el diccionario específico respecto al Bitcoin, se siguió el mismo método que los autores de VADER realizaron, se seleccionaron las 20 expresiones o términos más utilizados por la comunidad Bitcoin mediante el análisis de foros y la revisión de diccionarios Bitcoin existentes en la web. Dichas expresiones fueron valoradas por 10 personas independientes familiarizadas con el Bitcoin en un rango de -4 y +4, siendo -4 extremadamente negativo y el +4 extremadamente positivo. Con las valoraciones respectivas de cada expresión nueva, se realizó el cálculo de la media y la desviación estándar de cada una, teniendo en cuenta que la desviación estándar no exceda el valor de 2.5 para mantener la consistencia de todo el léxico. La media será el valor final que se incluirá en el léxico de VADER.

Siguiendo el procedimiento ya mencionado se añadieron 20 nuevas expresiones al léxico de VADER, como se observa en la siguiente tabla.

Tabla 11: Extracto de léxicos añadidos a VADER.

Expresión	Media	Desviación Estándar
Bull	3.4	0.7
Bear	-2.8	0.9
SEC	0.4	0.5
Hodl	0.4	0.5
Bubble	-2.4	0.7
Dust	-1.8	0.9
DYOR	0.6	0.8
Rekt	-3.2	0.7
Bearwhale	-2.4	0.9
Ashdraked	-3.8	0.3

Elaboración propia

Con las nuevas expresiones añadidas al clasificador de sentimientos VADER, se realizó el análisis de sentimientos de cada tweet. Para ello, primero se importaron los 4 splits de la etapa anterior y se concatenaron en un solo dataframe “tweet_data”.

Figura 27: Importación de data y concatenación de los 4 splits.

```
tweet_data_path_01 = 'no_promote_twitter_data/no_promote_bitcoin_data_1.csv'
tweet_data_01 = pd.read_csv(tweet_data_path_01, skiprows=0, lineterminator='\n')

tweet_data_path_02 = 'no_promote_twitter_data/no_promote_bitcoin_data_2.csv'
tweet_data_02 = pd.read_csv(tweet_data_path_02, skiprows=0, lineterminator='\n')

tweet_data_path_03 = 'no_promote_twitter_data/no_promote_bitcoin_data_3.csv'
tweet_data_03 = pd.read_csv(tweet_data_path_03, skiprows=0, lineterminator='\n')

tweet_data_path_04 = 'no_promote_twitter_data/no_promote_bitcoin_data_4.csv'
tweet_data_04 = pd.read_csv(tweet_data_path_04, skiprows=0, lineterminator='\n')


tweet_data = pd.concat([tweet_data_01, tweet_data_02, tweet_data_03, tweet_data_04])
```

Elaboración propia

Paso seguido se agregó el sentimiento de cada tweet en una nueva columna llamada “sentiment”. Cabe resaltar que para esta investigación solo se consideró el “Compound Score” que se calcula sumando los puntajes de valencia de cada palabra en el léxico, se ajusta de acuerdo con las reglas y luego se normaliza entre -1 (extremo negativo) y +1 (extremo positivo), esta métrica es la más útil ya que nos brinda una única medida unidimensional de sentimiento para una oración determinada.

Figura 28: Código para agregar el sentimiento a cada tweet y extracto del dataset principal.

```
new_tweet_data_2['sentiment'] = new_tweet_data_2['text'].apply(lambda x: sid.polarity_scores(x)['compound'])
```

```
new_tweet_data_2.head(20)
```

	user	timestamp	replies	likes	retweets	text	lang	promote	sentiment
0	3eyedbran	2019-05-27 11:49:06+00	0	2	1	Another Test tweet that wasn't caught in the s...	en	-	0.0000
1	DetroitCrypto	2019-05-27 11:49:22+00	0	0	0	Current Crypto Prices! \n\nBTC: $8721.99 USD\n...	en	-	0.0000
2	mmursaleen72	2019-05-27 11:49:23+00	0	0	0	Spiv (Nosar Baz): BITCOIN Is An Asset & NO...	en	-	0.3612
3	evilrobotted	2019-05-27 11:49:25+00	0	0	0	@nwoodfine We have been building on the real #...	en	-	-0.4767
4	MLWright15	2019-05-27 11:49:32+00	0	0	0	CHANGE IS COMING...GET READY!!! Boom, Another ...	en	-	0.7422
5	ltonews	2019-05-27 11:49:19+00	0	14	2	One of the useful articles of Stefan; here is ...	en	-	0.4404
6	Cybintelligence	2019-05-22 12:42:16+00	3	2	7	BestMixer has been seized by the Dutch Police ...	en	-	0.0000
7	optbus_hw45	2019-05-27 11:49:30+00	1	1	1	Invested my Life Savings into Bitcoin and Ethe...	en	-	0.0000
8	CCNMarkets	2019-05-27 08:13:06+00	5	167	68	Bitcoin Price Hits $8,939 in New 2019 High: Wh...	en	-	0.0000

Elaboración propia

Para finalizar esta parte, se guardó el nuevo dataframe en un archivo “.csv”.

4.1.3.4. Unión de los Datasets Preprocesados

Antes de unir la data histórica del Bitcoin con la data extraída de Twitter se realizaron algunos pasos adicionales. El primer paso fue quitar la información de la zona horaria (tzinfo) de la columna “timestamp” para tener el formato “YYYY-MM-DD HH:MM:SS”.

Figura 29: Código para deshabilitar la información de zona horaria y extracto del dataframe.

```
tweet_data['timestamp'] = tweet_data['timestamp'].apply(lambda d: d.replace(tzinfo=None))
```

```
tweet_data.head(10)
```

	user	timestamp	replies	likes	retweets	text	lang	promote	sentiment
0	3eyedbran	2019-05-27 11:49:06	0	2	1	Another Test tweet that wasn't caught in the s...	en	-	0.0000
1	DetroitCrypto	2019-05-27 11:49:22	0	0	0	Current Crypto Prices! \n\nBTC: $8721.99 USD\n...	en	-	0.0000
2	mmursaleen72	2019-05-27 11:49:23	0	0	0	Spiv (Nosar Baz): BITCOIN Is An Asset & NO...	en	-	0.3612
3	evilrobotted	2019-05-27 11:49:25	0	0	0	@nwoodfine We have been building on the real #...	en	-	-0.4767
4	MLWright15	2019-05-27 11:49:32	0	0	0	CHANGE IS COMING...GET READY!!! Boom, Another ...	en	-	0.7422
5	ltonews	2019-05-27 11:49:19	0	14	2	One of the useful articles of Stefan; here is ...	en	-	0.4404
6	Cybintelligence	2019-05-22 12:42:16	3	2	7	BestMixer has been seized by the Dutch Police ...	en	-	0.0000
7	optbus_hw45	2019-05-27 11:49:30	1	1	1	Invested my Life Savings into Bitcoin and Ethe...	en	-	0.0000
8	CCNMarkets	2019-05-27 08:13:06	5	167	68	Bitcoin Price Hits $8,939 in New 2019 High: Wh...	en	-	0.0000
9	malfouh	2019-05-27 11:49:17	1	1	1	#Countdown #ComingSoon The $QLC airdrop, #Q_Ga...	en	-	0.0000

Elaboración propia

El segundo paso consistió en agrupar la data extraída de Twitter por hora, esto con el fin de coincidir con el timestamp de la data histórica del Bitcoin para poder realizar el merge entre los dos dataframes. Adicionalmente, se creó una nueva columna llamada "Date" que nos servirá de punto de ancla para realizar el merge.

Figura 30: Código para agrupar horariamente la data extraída de Twitter y extracto del dataframe.

```
tweet_data_hourly = tweet_data.groupby(pd.Grouper(key='timestamp', freq='1H'))
                    .agg(replies=('replies','sum'), likes=('likes','sum'), retweets=('retweets','sum'),
                         sentiment=('sentiment','mean'), tweet_volume=('timestamp','count'))

tweet_data_hourly['Date'] = tweet_data_hourly.index

tweet_data_hourly.head()
```

timestamp	replies	likes	retweets	sentiment	tweet_volume	Date
2016-01-01 00:00:00	0	0	5	0.241254	13	2016-01-01 00:00:00
2016-01-01 01:00:00	0	5	14	0.188743	21	2016-01-01 01:00:00
2016-01-01 02:00:00	0	0	7	0.139074	19	2016-01-01 02:00:00
2016-01-01 03:00:00	0	3	8	0.131791	23	2016-01-01 03:00:00
2016-01-01 04:00:00	0	0	10	0.153862	21	2016-01-01 04:00:00

Elaboración propia

Para el caso de la data histórica del Bitcoin, solo se agregó una columna nueva llamada “Date_merge” con la que se realizó la unión de los 2 dataframes.

Figura 31: Creación de la columna "Date_merge" y extracto del dataframe.

```
bitcoin_hourly['Date_merge'] = bitcoin_hourly.index
```

```
bitcoin_hourly.head()
```

Date	Unix Timestamp	Date	Symbol	Open	High	Low	Close	Volume	Date_merge
2019-12-31 23:00:00	1577833200000	2019-12-31 23:00:00	BTCUSD	7173.53	7180.00	7156.01	7165.90	49.524838	2019-12-31 23:00:00
2019-12-31 22:00:00	1577829600000	2019-12-31 22:00:00	BTCUSD	7154.49	7180.58	7146.35	7173.53	93.290183	2019-12-31 22:00:00
2019-12-31 21:00:00	1577826000000	2019-12-31 21:00:00	BTCUSD	7146.92	7161.22	7146.92	7154.49	23.662080	2019-12-31 21:00:00
2019-12-31 20:00:00	1577822400000	2019-12-31 20:00:00	BTCUSD	7150.70	7160.00	7141.00	7146.92	61.640907	2019-12-31 20:00:00
2019-12-31 19:00:00	1577818800000	2019-12-31 19:00:00	BTCUSD	7141.39	7156.87	7133.69	7150.70	36.288595	2019-12-31 19:00:00

Elaboración propia

Finalmente, para unir los 2 dataframes se consideraron las columnas “Date” del dataframe que contiene la data extraída de Twitter y la columna “Date_merge” del dataframe que contiene la data histórica del Bitcoin. Además, se actualizó el index del dataframe general.

Figura 32: Merge de los 2 dataframes y extracto del dataframe final.

```
final_hourly_data = pd.merge(tweet_data_hourly, bitcoin_hourly, left_on='Date', right_on='Date_merge')
```

```
final_hourly_data_index = final_hourly_data.set_index(pd.DatetimeIndex(final_hourly_data['Date_merge']))
```

```
final_hourly_data_index.head()
```

Date_merge	replies	likes	retweets	sentiment	tweet_volume	Date_x	Unix Timestamp	Date_y	Symbol	Open	High	Low	Close	Volume	Date_merge
2016-01-01 00:00:00	0	0	5	0.241254	13	2016-01-01 00:00:00	1451606400	2016-01-01 00:00:00	BTCUSD	429.95	429.95	429.95	429.95	0.000000	2016-01-01 00:00:00
2016-01-01 01:00:00	0	5	14	0.188743	21	2016-01-01 01:00:00	1451610000	2016-01-01 01:00:00	BTCUSD	429.95	432.68	429.95	432.68	0.229500	2016-01-01 01:00:00
2016-01-01 02:00:00	0	0	7	0.139074	19	2016-01-01 02:00:00	1451613600	2016-01-01 02:00:00	BTCUSD	432.68	432.68	432.68	432.68	0.000000	2016-01-01 02:00:00
2016-01-01 03:00:00	0	3	8	0.131791	23	2016-01-01 03:00:00	1451617200	2016-01-01 03:00:00	BTCUSD	432.68	432.68	432.68	432.68	0.000000	2016-01-01 03:00:00
2016-01-01 04:00:00	0	0	10	0.153862	21	2016-01-01 04:00:00	1451620800	2016-01-01 04:00:00	BTCUSD	432.68	436.53	432.68	436.53	63.503759	2016-01-01 04:00:00

Elaboración propia

El dataframe resultante fue guardado en un archivo “.csv” para su posterior uso en el siguiente paso “Modelado”.

4.1.3.5. Selección de Características

Para este estudio se consideraron 3 variables adicionales al sentimiento promedio de cada tweet, estas variables son “Replies”, “Likes” y “Retweets”. Dichas variables nos permitirán cuantificar el alcance o influencia de un tweet, ya que como explica BAdiola Ramos (2019) es importante decir que no todos los tweets tienen la misma influencia o alcance, la mayoría de las publicaciones llega a pocos usuarios, y algunas otras son distribuidas a grandes redes de seguidores.

Además, con el fin de mejorar el comportamiento y precisión de las configuraciones del modelo, se realizó una medición de las variables presentes en el dataframe final (a excepción de las columnas “Date_x”, “Unix Timestamp”, “Date_y”, “Symbol” y “Date_merge”) obtenido del punto anterior para evaluar la relación que tienen entre sí mediante un análisis correlativo.

El análisis correlativo se realizó utilizando la Correlación de Spearman para medir la relación estadística entre las variables.

Figura 33: Heatmap de los indicadores de correlación de Spearman

Elaboración propia

Como se puede observar en la Figura 29, el resultado puede variar entre -1 y +1, esto nos ayudó a identificar la existencia de una relación positiva (cerca de +1), negativa (cerca de -1) o neutra (0), esta última significaría que no existe correlación por lo que la variable sería de poca relevancia para el estudio. Además, podemos observar que la variable que más resalta es “Likes”, que corresponde a la cantidad de likes que tiene un tweet con un coeficiente de correlación Spearman de 0.72 (72%) respecto a la variable objetivo “Close”. A continuación, se muestra un resumen de los coeficientes de correlación Spearman obtenidos respecto a la variable “Close”.

Tabla 12: Análisis correlativo de Spearman (Close).

Característica A	Característica B	Coeficiente de correlación
Close	Close	1.000000
Close	Likes	0.716701
Close	Replies	0.686800
Close	Tweet_Volume	0.651952
Close	Retweets	0.554935
Close	Volume	0.298148
Close	Sentiment	0.215071

Elaboración propia

Por otro lado, ya que en este trabajo de investigación se considera el análisis de sentimientos (compound) como una variable para predecir el comportamiento, es importante e interesante analizar la correlación existente entre esta variable y las demás variables ya vistas anteriormente.

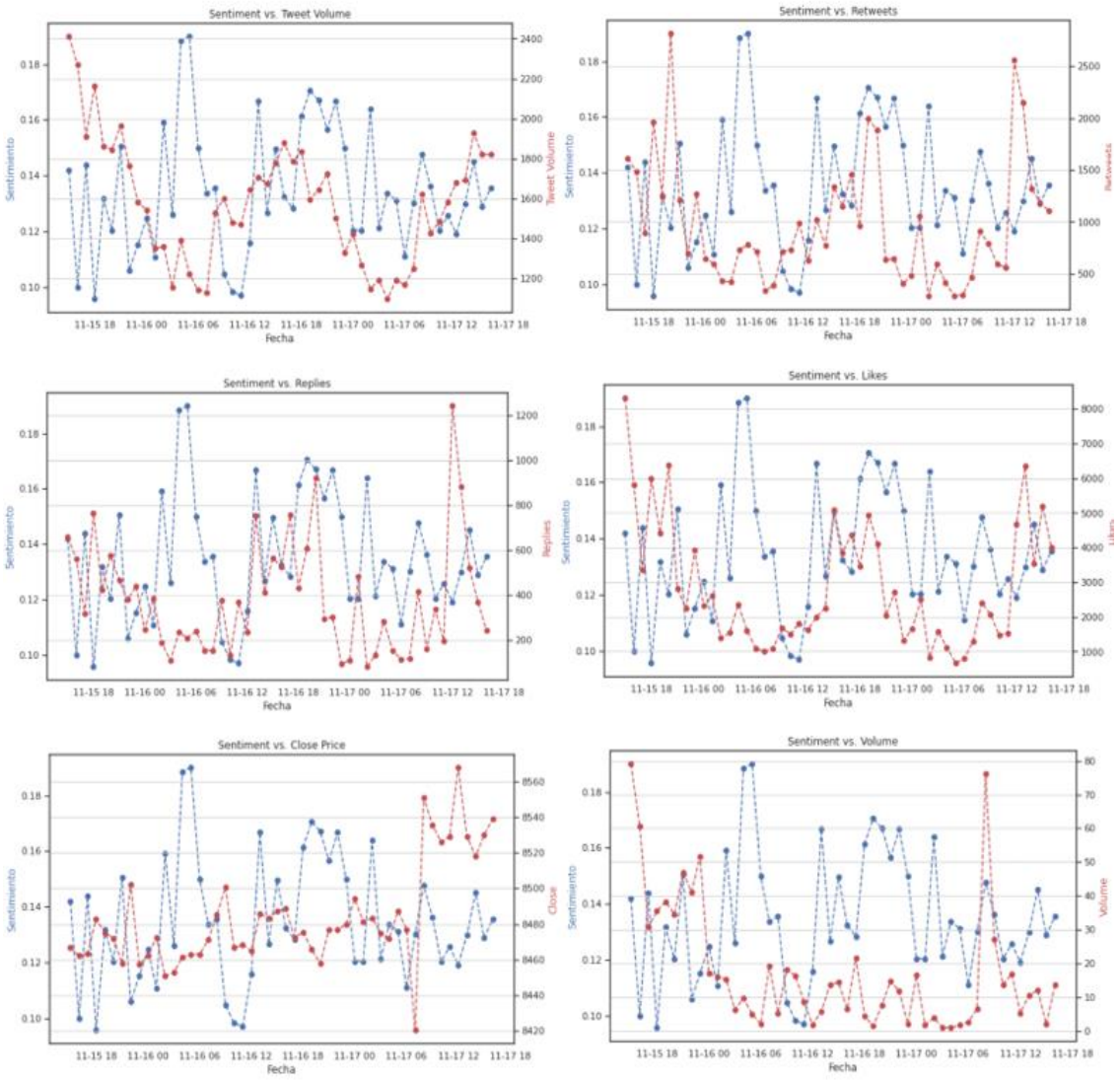
Tabla 13: Análisis correlativo de Spearman (Sentiment).

Característica A	Característica B	Coeficiente de correlación
Sentiment	Sentiment	1.000000
Sentiment	Tweet_Volume	0.516635
Sentiment	Retweets	0.482869
Sentiment	Replies	0.418635
Sentiment	Likes	0.418091
Sentiment	Close	0.215071
Sentiment	Volume	-0.135009

Elaboración propia

De la Tabla 13, podemos apreciar que la variable “Tweet_Volume” es la más significativa respecto a la variable “Sentiment”, es decir, es la variable que más relación guarda con el sentimiento promedio de los tweets (directamente proporcional), con un coeficiente de correlación de 0.516635 seguida de la variable “Retweets”. Las variables “Replies” y “Likes” tienen coeficientes de correlación bastante parecidos entre sí respecto al sentimiento de los tweets y su valor es lo suficientemente alto como para ser consideradas como variables significativas para el modelo predictivo.

Figura 34: Relación entre las variables de estudio y el sentimiento promedio de los Tweets.



Elaboración propia

Sin embargo, las variables “Close” y “Volume” tienen coeficientes de correlación muy bajos lo cual quiere decir que no existe una relación lo suficientemente fuerte entre estas variables y el sentimiento promedio por lo que serían consideradas poco significativas. Sin embargo, para este trabajo de investigación se considerará la variable “Close” (precio de cierre del Bitcoin) como variable objetivo ya que lo que se busca es predecir el comportamiento del Bitcoin (fluctuación de precio) por lo que nos interesan aquellas variables que guarden una relación significativa, ya sea directamente proporcional o inversamente proporcional, con la variable “Close” por lo que se tomará la información expuesta en la Tabla 12.

Con esto fue posible descartar aquellas variables donde la relación es poco significativa lo que nos deja un total de 7 variables de estudio (“Replies”, “Likes”, “Retweets”, “Sentiment”, “Tweet_Volume”, “Close” y “Volume”) que son más apropiadas para la implementación del modelo en sus 3 configuraciones. Cabe resaltar que la variable “Sentiment” tiene el coeficiente de correlación más bajo de todas las variables consideradas para la implementación del modelo lo que indica que no es una buena variable para predecir el comportamiento del Bitcoin (Precio de cierre - Close). Aun así, se consideró dicha variable para la implementación del modelo ya que constituye una parte fundamental del trabajo de investigación, aunque definitivamente reducirá la precisión del modelo en sus 3 configuraciones de predicción a futuro.

Figura 35: Variables finales de estudio.

Date_merge	Replies	Likes	Retweets	Sentiment	Tweet_Volume	Close	Volume
2016-01-01 00:00:00	0	0	5	0.241254	13	429.95	0.000000
2016-01-01 01:00:00	0	5	14	0.188743	21	432.68	0.229500
2016-01-01 02:00:00	0	0	7	0.139074	19	432.68	0.000000
2016-01-01 03:00:00	0	3	8	0.131791	23	432.68	0.000000
2016-01-01 04:00:00	0	0	10	0.153862	21	436.53	63.503759

Elaboración propia

4.1.4. Modelado

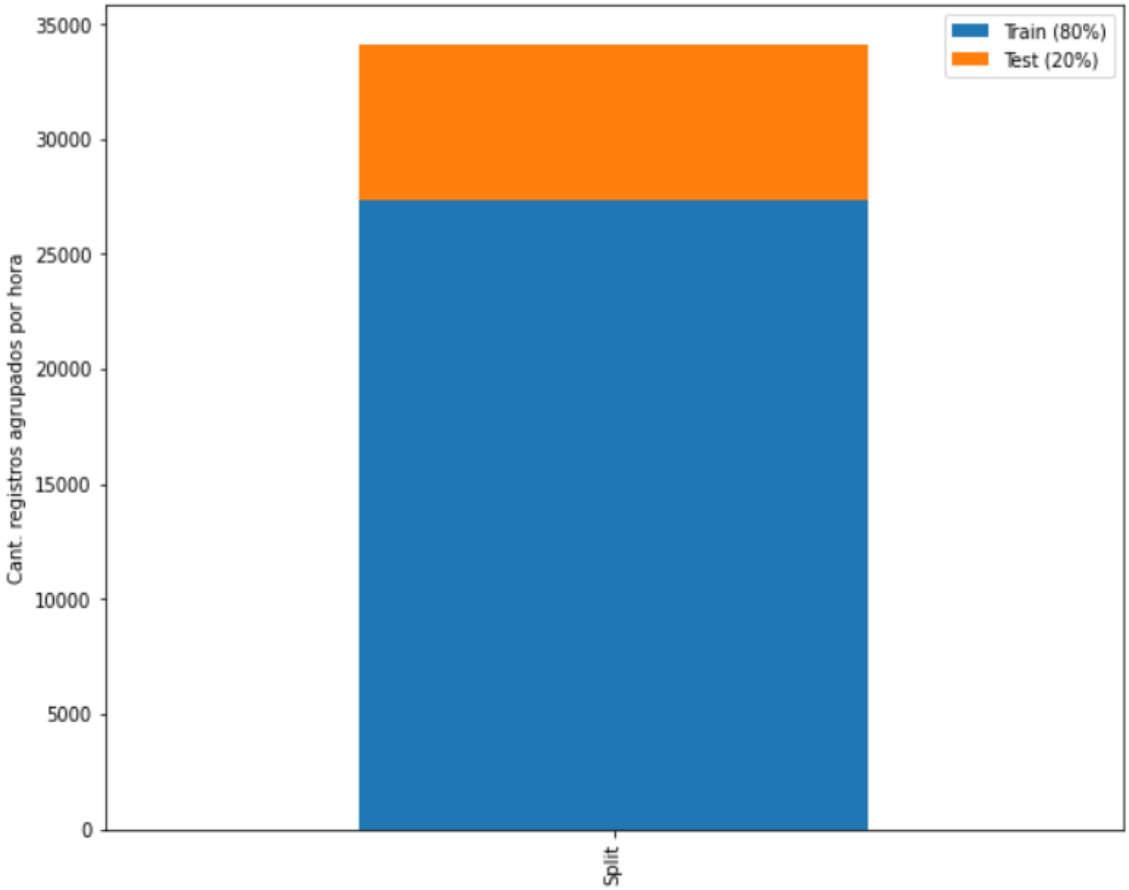
En esta sección se describirá todo el proceso concerniente a la construcción del modelo de predicción y sus 3 configuraciones respectivas.

Para este estudio la mejor solución consiste en utilizar redes neuronales recurrentes (Recurrent Neural Network) ya que se utilizan datos secuenciales que dependen del tiempo para tener en cuenta los valores pasados y presentes antes de predecir un valor futuro. El modelo que se implementó en este proyecto fue realizado con LSTM (Long Short Term Memory), que es uno de los algoritmos más comunes para crear este tipo de redes neuronales recurrentes.

Es importante mencionar que las redes neuronales LSTM nos permiten definir cuánto tiempo en el pasado tendrá en cuenta el modelo para predecir un valor a futuro y después de haber experimentado con diferentes periodos de tiempo (horas), se decidió utilizar 3 hrs. para predecir un valor a futuro. Por otro lado, otra de las configuraciones de las redes neuronales LSTM es cuánto tiempo a futuro se quiere predecir y para este caso se ha experimentado con 3 configuraciones diferentes que han sido entrenadas para predecir 1 hr., 6 hrs. y 12 hrs. a futuro, dichas configuraciones fueron comparadas para su análisis posterior.

Otro aspecto importante a mencionar en esta etapa de modelado, es la división de la data en entrenamiento y prueba, para la fase de entrenamiento se utilizaron 27313 registros que corresponden al 80% del dataset y para la fase de prueba se utilizaron los 6829 registros restantes que corresponden al 20% del dataset, esto con el fin de probar la eficacia del modelo en sus 3 configuraciones.

Figura 36: Split del dataset (80 - 20).



Elaboración propia

Antes de seguir con la implementación de los modelos propiamente, es importante convertir la data para aplicar los modelos LSTM, esto implica convertirlos de una serie de datos de tiempo en una secuencia supervisada, una matriz 3d como tal con variables normalizadas, siendo la variable “Close” la variable objetivo. Para ello se codificó una función que recibe como primer parámetro toda la data, el segundo parámetro hace referencia al lookback que es la cantidad de horas en el pasado que el modelo tomará en

cuenta para predecir un valor futuro y el tercer parámetro es la cantidad futura de horas a predecir.

Figura 37: Código para convertir la data de una serie de datos de tiempo a una secuencia supervisada.

```
def series_to_supervised(data, n_in=1, n_out=1, dropnan=True):
  n_vars = 1 if type(data) is list else data.shape[1]
  df = DataFrame(data)
  cols, names = list(), list()
  # Secuencia de entrada (t-n, ... t-1)
  for i in range(n_in, 0, -1):
    cols.append(df.shift(i))
    names += [('var%d(t-%d)' % (j+1, i)) for j in range(n_vars)]

  # Secuencia de pronóstico (t, t+1, ... t+n)
  for i in range(0, n_out):
    cols.append(df.shift(-i))
    if i == 0:
      names += [('var%d(t)' % (j+1)) for j in range(n_vars)]
    else:
      names += [('var%d(t+%d)' % (j+1, i)) for j in range(n_vars)]

  # Juntamos todo
  agg = concat(cols, axis=1)
  agg.columns = names

  # Descartar filas con valores NaN
  if dropnan:
    agg.dropna(inplace=True)
  return agg
```

Elaboración propia

Debemos considerar también que el resultado de esta función será un dataframe que no muestra el nombre de las variables como tal, sino un nombre equivalente por motivos de orden y mejor visualización. A continuación, se muestra una tabla de equivalencia de nombre de las variables.

Tabla 14: Equivalencia de nombre de variables.

Variable	Equivalente
Close	var1

(continuación…)

Replies	var2
Likes	var3
Retweets	var4
Sentiment	var5
Tweet_Volume	var6
Volume	var7

Elaboración propia

También es necesario reordenar las columnas del dataframe y posicionar como primero columna a la variable “Close” ya que esta variable será la variable objetivo como se mencionó anteriormente.

Figura 38: Extracto del dataframe con las columnas reordenadas.

Date_merge	Close	Replies	Likes	Retweets	Sentiment	Tweet_Volume	Volume
2016-01-01 00:00:00	429.95	0	0	5	0.241254	13	0.000000
2016-01-01 01:00:00	432.68	0	5	14	0.188743	21	0.229500
2016-01-01 02:00:00	432.68	0	0	7	0.139074	19	0.000000
2016-01-01 03:00:00	432.68	0	3	8	0.131791	23	0.000000
2016-01-01 04:00:00	436.53	0	0	10	0.153862	21	63.503759

Elaboración propia

Se realizó la normalización de las variables de entrada utilizando la función “MinMaxScaler(feature_range=(0, 1))” de la librería sklearn que transformaron todas las variables en un rango de [0,1], es decir, el valor mínimo y máximo de una variable de entrada será 0 y 1 respectivamente. Este proceso se realiza ya que aquellas variables que se miden en diferentes escalas no contribuyen por igual a la función de ajuste y aprendizaje del modelo, y pueden terminar creando un sesgo. Adicionalmente, se utilizó la función “fit_transform()” para ajustar y transformar los datos de entrada al mismo

tiempo y convertir los puntos de datos para que la eficiencia del modelo no se vea afectada al realizar el ajuste y transformación de forma separada.

A continuación, se mostrará una tabla con las características del modelo predictivo.

Tabla 15: Características del modelo predictivo.

Característica	Valor
Épocas	250
Loss	mae
Optimizer	nadam
Batch_size	50
Verbose	2
Shuffle	False
Validation_split	0.2

Elaboración propia

El resumen de la estructura de la red LSTM utilizada en este proyecto podemos obtenerlo utilizando la función “summary()” que nos proporciona la librería Keras.

Figura 39: Estructura de la red LSTM.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm (LSTM)                 (None, 10)                720

 dense (Dense)               (None, 1)                 11

=================================================================
Total params: 731
Trainable params: 731
Non-trainable params: 0
_________________________________________________________________
```

Elaboración propia

4.1.4.1. Modelo Predictivo Configurado para Predecir 1 Hora a Futuro

Esta configuración considera un lookback de 3 horas y las 7 variables del dataframe por lo que la cantidad de observaciones (n_obs) fue de 21. Aplicando la función de conversión de series temporales a una secuencia supervisada para predecir 1 hora futuro, tenemos el siguiente dataframe.

Figura 40: Data convertida a una secuencia supervisada (Configuración a 1hr.)

	var1(t-3)	var2(t-3)	var3(t-3)	var4(t-3)	var5(t-3)	var6(t-3)	\
34136	0.350536	0.003110	0.004134	0.002536	0.586586	0.080048	
34137	0.348152	0.003365	0.004673	0.003084	0.576988	0.087869	
34138	0.349397	0.003849	0.003738	0.002948	0.588959	0.089969	
34139	0.349521	0.004346	0.005219	0.004333	0.602097	0.090441	
34140	0.350052	0.000956	0.002570	0.001710	0.583753	0.087187	

	var7(t-3)	var1(t-2)	var2(t-2)	var3(t-2)	...	var5(t)	var6(t)	\
34136	0.002636	0.348152	0.003365	0.004673	...	0.602097	0.090441	
34137	0.005077	0.349397	0.003849	0.003738	...	0.583753	0.087187	
34138	0.001695	0.349521	0.004346	0.005219	...	0.591668	0.096110	
34139	0.000142	0.350052	0.000956	0.002570	...	0.595685	0.104404	
34140	0.002559	0.348698	0.002957	0.003915	...	0.608722	0.104246	

	var7(t)	var1(t+1)	var2(t+1)	var3(t+1)	var4(t+1)	var5(t+1)	\
34136	0.000142	0.350052	0.000956	0.002570	0.001710	0.583753	
34137	0.002559	0.348698	0.002957	0.003915	0.002916	0.591668	
34138	0.000515	0.350159	0.002294	0.004274	0.002650	0.595685	
34139	0.000761	0.353997	0.002345	0.003717	0.002514	0.608722	
34140	0.004528	0.356101	0.001020	0.001966	0.001510	0.596222	

	var6(t+1)	var7(t+1)
34136	0.087187	0.002559
34137	0.096110	0.000515
34138	0.104404	0.000761
34139	0.104246	0.004528
34140	0.088027	0.005705

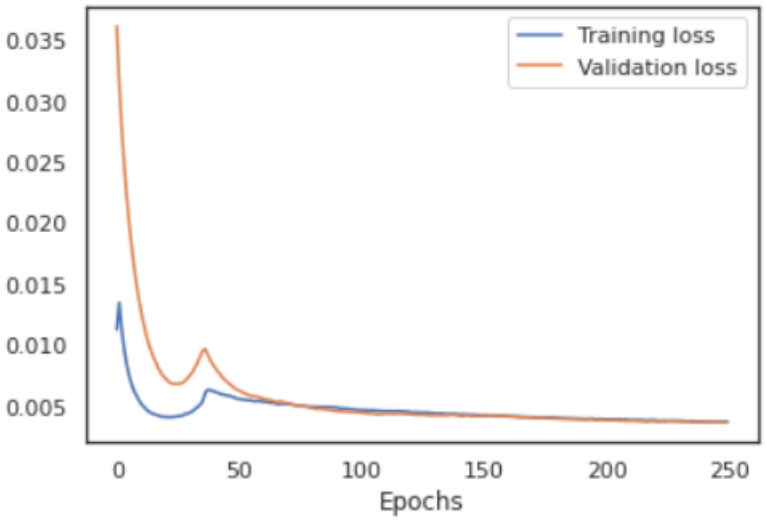
Elaboración propia

Como se mencionó anteriormente, el modelo está configurado para considerar 3 horas en el pasado (t-3, t-2 y t-1) de cada variable de estudio y así predecir un valor futuro (t+1) ya que para esta configuración deseamos predecir 1 hora a futuro.

Al finalizar el entrenamiento del modelo, tenemos 2 métricas importantes, Training Loss y Validation Loss. El Training Loss se utiliza para evaluar cómo un modelo

de aprendizaje profundo se ajusta a los datos de entrenamiento, es decir, evalúa el error del modelo sobre el conjunto de entrenamiento. Por el contrario, el Validation Loss se utiliza para evaluar el rendimiento de un modelo de aprendizaje profundo en el conjunto de validación. Podemos graficar ambas métricas utilizando la librería Matplotlib.

Figura 41: Training loss y Validation loss del modelo predictivo a 1 hr.



Elaboración propia

4.1.4.2. Modelo Predictivo Configurado para Predecir 6 Horas a Futuro

Esta configuración también considera un lag de 3 horas y las 7 características del dataframe por lo que la cantidad de observaciones (n_obs) fue de 21. Aplicando la función de conversión de series temporales a una secuencia supervisada para predecir 6 horas a futuro, tenemos el siguiente dataframe.

Figura 42: Data convertida a una secuencia supervisada (Configuración a 6hr.)

	var1(t-3)	var2(t-3)	var3(t-3)	var4(t-3)	var5(t-3)	var6(t-3)	\
34131	0.354071	0.001848	0.004318	0.002118	0.596085	0.082463	
34132	0.355485	0.005086	0.009729	0.003730	0.578918	0.080101	
34133	0.351861	0.005646	0.010227	0.003774	0.582317	0.075954	
34134	0.351721	0.001644	0.002813	0.001515	0.589231	0.073067	
34135	0.352327	0.001147	0.002028	0.001727	0.581930	0.075219	

	var7(t-3)	var1(t-2)	var2(t-2)	var3(t-2)	...	var5(t+5)	var6(t+5)	\
34131	0.006085	0.355485	0.005086	0.009729	...	0.602097	0.090441	
34132	0.000976	0.351861	0.005646	0.010227	...	0.583753	0.087187	
34133	0.005506	0.351721	0.001644	0.002813	...	0.591668	0.096110	
34134	0.002198	0.352327	0.001147	0.002028	...	0.595685	0.104404	
34135	0.001232	0.350536	0.003110	0.004134	...	0.608722	0.104246	

	var7(t+5)	var1(t+6)	var2(t+6)	var3(t+6)	var4(t+6)	var5(t+6)	\
34131	0.000142	0.350052	0.000956	0.002570	0.001710	0.583753	
34132	0.002559	0.348698	0.002957	0.003915	0.002916	0.591668	
34133	0.000515	0.350159	0.002294	0.004274	0.002650	0.595685	
34134	0.000761	0.353997	0.002345	0.003717	0.002514	0.608722	
34135	0.004528	0.356101	0.001020	0.001966	0.001510	0.596222	

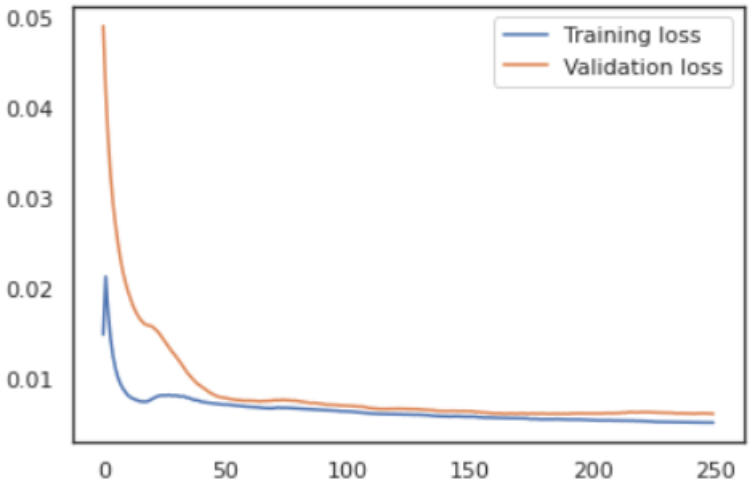
	var6(t+6)	var7(t+6)
34131	0.087187	0.002559
34132	0.096110	0.000515
34133	0.104404	0.000761
34134	0.104246	0.004528
34135	0.088027	0.005705

Elaboración propia

Como se mencionó anteriormente, el modelo está configurado para considerar 3 horas en el pasado (t-3, t-2 y t-1) de cada variable de estudio y así predecir un valor futuro que en este caso es 6 horas a futuro (t+1, t+2, t+3, t+4, t+5 y t+6).

Las métricas (Training loss y Validation loss) obtenidas en este modelo podemos graficarlas utilizando la librería Matplotlib.

Figura 43: Training loss y Validation loss del modelo predictivo a 6 hr.



Elaboración propia

4.1.4.3. Modelo Predictivo Configurado para Predecir 12 Horas a Futuro

Esta configuración también considera un lag de 3 horas y las 7 características del dataframe por lo que la cantidad de observaciones (n_obs) fue de 21. Aplicando la función de conversión de series temporales a una secuencia supervisada para predecir 12 horas a futuro, tenemos el siguiente dataframe.

Figura 44: Data convertida a una secuencia supervisada (Configuración a 12hr.)

	var1(t-3)	var2(t-3)	var3(t-3)	var4(t-3)	var5(t-3)	var6(t-3)	\
3	0.003812	0.0	0.000000	0.000027	0.667265	0.000682	
4	0.003952	0.0	0.000017	0.000076	0.626718	0.001102	
5	0.003952	0.0	0.000000	0.000038	0.588366	0.000997	
6	0.003952	0.0	0.000010	0.000043	0.582743	0.001207	
7	0.004149	0.0	0.000000	0.000054	0.599785	0.001102	

	var7(t-3)	var1(t-2)	var2(t-2)	var3(t-2)	...	var5(t+11)	var6(t+11)	\
3	0.000000	0.003952	0.0	0.000017	...	0.610163	0.001627	
4	0.000027	0.003952	0.0	0.000000	...	0.586247	0.001785	
5	0.000000	0.003952	0.0	0.000010	...	0.608804	0.001470	
6	0.000000	0.004149	0.0	0.000000	...	0.597719	0.001627	
7	0.007448	0.004149	0.0	0.000020	...	0.604396	0.001522	

	var7(t+11)	var1(t+12)	var2(t+12)	var3(t+12)	var4(t+12)	var5(t+12)	\
3	1.172780e-07	0.004034	0.0	0.000000	0.000081	0.586247	
4	0.000000e+00	0.004002	0.0	0.000000	0.000076	0.608804	
5	5.746620e-06	0.003981	0.0	0.000003	0.000054	0.597719	
6	1.144265e-03	0.004108	0.0	0.000014	0.000060	0.604396	
7	2.147853e-05	0.004099	0.0	0.000003	0.000065	0.592826	

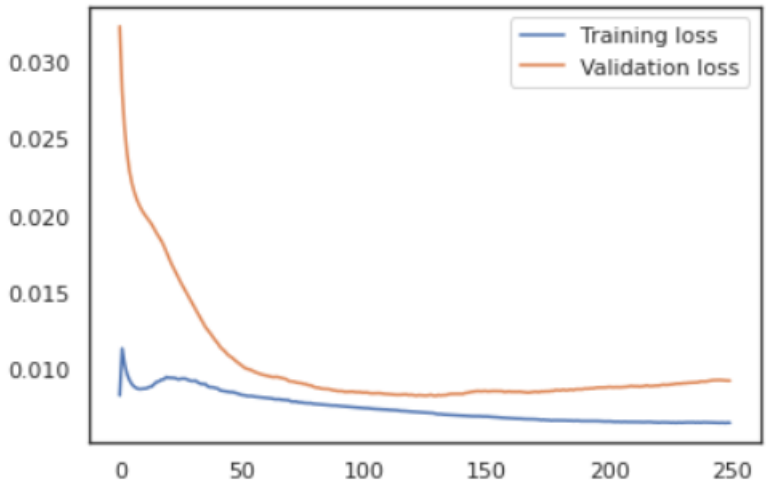
	var6(t+12)	var7(t+12)
3	0.001785	0.000000
4	0.001470	0.000006
5	0.001627	0.001144
6	0.001522	0.000021
7	0.001680	0.000054

Elaboración propia

Como se mencionó anteriormente, el modelo está configurado para considerar 3 horas en el pasado (t-3, t-2 y t-1) de cada variable de estudio y así predecir un valor futuro que en este caso es 12 horas a futuro (t+1, t+2, t+3, t+4, t+5, t+6, t+7, t+8, t+9, t+10, t+11 y t+12).

Las métricas (Training loss y Validation loss) obtenidas en este modelo podemos graficarlas utilizando la librería Matplotlib.

Figura 45: Training loss y Validation loss del modelo predictivo a 12 hr.



Elaboración propia

4.1.5. Evaluación

Para evaluar las 3 configuraciones del modelo predictivo implementado se utilizaron 2 métricas de error, RMSE y MAPE. Con estas métricas se pudieron obtener valores comparables entre las configuraciones en términos porcentuales (MAPE) y validación interna del error del modelo (RMSE). En la Tabla 16, se presentan los valores resultantes de cada configuración del modelo predictivo.

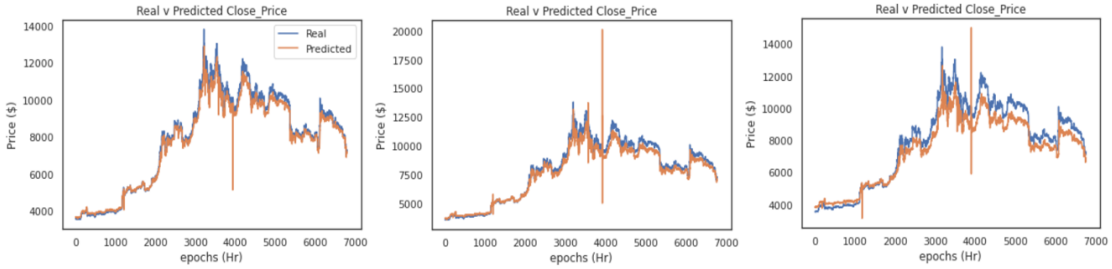
Tabla 16: Comparación de resultados de las 3 configuraciones del modelo predictivo implementado (RMSE y MAPE).

Configuración	RMSE	MAPE
1 hora a futuro	227.413	0.022
6 horas a futuro	372.079	0.035
12 horas a futuro	617.936	0.057

Elaboración propia

A continuación, se mostrarán gráficamente los resultados obtenidos en las diferentes configuraciones considerando el precio real del Bitcoin y la variable "Close_Price" de toda la data de prueba.

Figura 46: Predicciones obtenidas con las 3 configuraciones de la red LSTM.



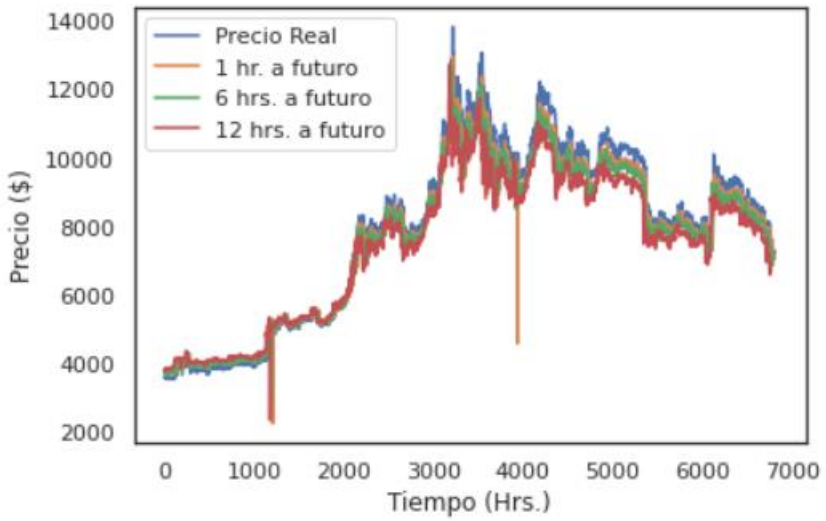
a) b) c)

Elaboración propia

El primer gráfico de la Figura 46 (a), muestra la predicción obtenida por el modelo configurado 1 hr. a futuro que presenta los mejores resultados, coincidiendo con una mayor precisión la línea de color azul que corresponde a la variación del precio real del Bitcoin con la línea de color anaranjado que corresponde a la predicción realizada del precio del Bitcoin. El segundo y tercer gráfico (b) y (c) corresponden a las predicciones obtenidas por los modelos configurados a 6 hrs y 12 hrs. respectivamente, se puede observar que la precisión de las líneas de color anaranjado es inferior a la precisión lograda con la primera configuración lo que confirma los valores obtenidos con la métrica RMSE.

Es posible comparar visualmente el comportamiento real del Bitcoin con las predicciones realizadas por las 3 configuraciones del modelo predictivo teniendo en cuenta la data de testing como se muestra a continuación en la Figura 47.

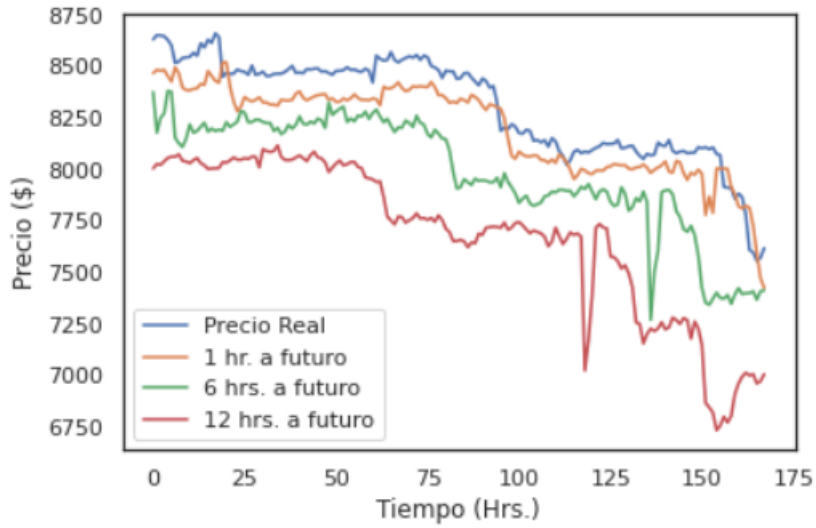
Figura 47: Comparación de las predicciones considerando toda la data de prueba.



Elaboración propia

Además, para analizar los resultados gráficos con más detalle es necesario realizar un zoom en la data de tal forma que se puedan visualizar las diferentes precisiones obtenidas del estudio (RMSE). La Figura 48 muestra lo mencionado anteriormente considerando 168 hrs., equivalentes a los últimos 7 días de la data de prueba (test).

Figura 48: Comparación de las predicciones de las 3 configuraciones del modelo con el precio real de los últimos 7 días.



Elaboración propia

En la Figura 48, podemos observar claramente que el modelo predictivo configurado para predecir 1 hr. a futuro (color amarillo) es el más cercano al

comportamiento (fluctuación) del precio real del Bitcoin (color azul) con un RMSE de 227.413, es decir, existe una diferencia de $ 227.413 entre el valor predicho por el modelo y el valor real del Bitcoin.

4.2. DISCUSIÓN

A partir de los resultados obtenidos se acepta la hipótesis que señala que el modelo predictivo predice adecuadamente el comportamiento de la criptodivisa Bitcoin aplicando análisis de sentimientos en Twitter y variables de alcance para medir la influencia de los tweets en el criptomercado.

Badiola Ramos (2019) en sus resultados señala que el sentimiento promedio de los tweets no es una buena variable para predecir el valor del Bitcoin lo que concuerda con los resultados encontrados en esta investigación, ya que se pudo observar que el coeficiente de correlación Spearman entre la variable que mide el sentimiento promedio de los tweets (Sentiment) y el precio de cierre del Bitcoin (Close) es bajo con un valor de 0.215071.

En cuanto a la correlación entre la fluctuación del precio del Bitcoin y los sentimientos del público en los tweets, Pagolu et al. (2017) sostienen que la correlación mencionada si existe y es fuerte, sin embargo, en esta investigación se encontró que, si bien es cierto existe dicha correlación, no es lo suficientemente fuerte como para considerar al sentimiento promedio de los tweets como una variable principal en la predicción del comportamiento del Bitcoin, esto puede deberse al objeto de estudio y su propia naturaleza ya que, por un lado, tenemos como objeto de estudio a la fluctuación del precio del Bitcoin (criptodivisa) con una naturaleza muy volátil y, por otro lado, a la fluctuación del precio de acciones empresariales que es mucho más estable. En adición, Wołk (2020) concluye que las fluctuaciones del precio de una criptomoneda dependen en

gran medida del sentimiento en las redes sociales (Twitter) lo cual difiere con el presente estudio y esto puede explicarse con el enfoque de estudio ya que para ese estudio se consideraron sentimientos negativos generales (incluidos los sentimientos ponderados) indicando que las noticias negativas tienen un mayor peso e importancia para la predicción del precio del Bitcoin.

En lo que respecta a la selección de variables para el modelo predictivo. Abraham et al. (2018) menciona que el volumen de tweets es mejor indicador predictivo del precio del Bitcoin que el sentimiento de los tweets, lo que concuerda con los resultados obtenidos en el presente estudio ya que se demostró que el volumen de tweets es una variable altamente relacionada con el precio de cierre del Bitcoin (Close) con un coeficiente de correlación Spearman de 0.651952.

Por otro lado, en cuanto a los resultados obtenidos del modelo predictivo en sus diferentes configuraciones, se observó que el mejor resultado lo obtuvo el modelo predictivo configurado para predecir 1 hora a futuro con un RMSE de 227.413, este resultado es relativamente mejor que los resultados obtenidos por Badiola Ramos (2019) y Ferdiansyah et al. (2019). En este punto es necesario entender algunos de los posibles motivos por los cuales existe una diferencia de resultados. En primer lugar, el periodo de estudio considerado en la primera investigación mencionada es diferente ya que solo se considera un intervalo de tiempo de aproximadamente 1 año y medio (agosto del 2017 - enero de 2019) mientras que para este trabajo de investigación se consideró un periodo de aproximadamente 4 años. Segundo, el lookback (tiempo en el pasado que considera la red para predecir un valor a futuro) utilizado por el primer trabajo mencionado es de 1 día (24 horas) lo que difiere en gran medida del lookback utilizado en este trabajo que es de 3 horas por tener un enfoque de scalping (inversión a corto plazo). Tercero, las variables de entrada consideradas para el modelo predictivo son distintas, si bien es cierto que

algunas variables son parecidas, en ninguna de las otras investigaciones se consideraron variables de alcance para medir la influencia de los tweets en el criptomercado, en este sentido, el presente trabajo propone 3 nuevas variables de entrada o indicadores altamente relacionados con la fluctuación del precio del Bitcoin (Likes [ρ = 0.716701], Replies [ρ = 0.686800] y Retweets [ρ = 0.554935]), que son incluso mejores variables de entrada, a excepción de la variable “Retweets”, que el volumen de tweets considerada como una de las variables más significativas en la predicción del precio de Bitcoin.

V. CONCLUSIONES

PRIMERA

Se logró predecir adecuadamente el comportamiento de la criptodivisa Bitcoin aplicando análisis de sentimientos en Twitter considerando 3 configuraciones distintas del modelo para predecir 1hr., 6 hrs. y 12 hrs. a futuro, concluyendo que la configuración de predicción 1 hora a futuro es la que mejores resultados obtuvo comparada con las otras 2 configuraciones, con un RMSE de 227.413 y un MAPE de 0.022. Sin embargo, la desventaja es que el resultado no es lo suficientemente bueno con respecto al RMSE por lo que no se recomienda basar decisiones de inversión en el criptomercado de Bitcoin utilizando solamente los resultados del modelo propuesto en este trabajo.

SEGUNDA

Se recopiló y preprocesó la data histórica de Bitcoin y los tweets referidos a Bitcoin. En el caso de la data histórica del Bitcoin se utilizó la plataforma Gemini, antes ya mencionada, de la cual se extrajo un archivo ".csv" que contenía 8 columnas ("Unix Timestamp", "Date", "Symbol", "Open", "High", "Low", "Close" y "Volume") de las cuales solo se utilizaron 2 ("Close" y "Volume") como variables de entrada para el modelo. Para la data sobre Bitcoin de Twitter se utilizó un dataset ya existente y disponible en la plataforma Kaggle, este dataset contenía 9 columnas ("Id", "User", "Fullname", "Url", "Timestamp", "Replies", "Likes", "Retweets", "Text") de las cuales, para el modelo final solo se utilizaron 3 ("Replies", "Likes" y "Retweets") como variables de alcance; sin embargo, utilizando este mismo dataset se crearon 2 columnas adicionales ("Sentiment" y "Tweet_Volume").

En un primer análisis se revisaron las columnas de cada dataset para entender su fin, conocer los tipos de datos de cada una y la cantidad de registros presentes. Se

utilizaron gráficos para cada variable inicial con el objetivo de ver su comportamiento y comprenderlo de una forma más gráfica. El segundo análisis fue más detallado ya que primero se seleccionó solo la data comprendida entre 2016 y 2019 en ambos datasets, en el caso de la data extraída de Twitter solo se trabajó con tweets en inglés para lo cual se utilizó la librería "Whatthelang" y se aplicó un criterio de limpieza basado en la identificación de "promote" y "no promote" tweets.

TERCERA

Se aplicó el análisis de sentimientos y la selección de características. Para el análisis de sentimientos se modificó el clasificador VADER, al que se le añadieron 20 términos o expresiones comúnmente usadas en la comunidad Bitcoin siguiendo el mismo método que los autores originales utilizaron, para representar de mejor manera el sentimiento del criptomercado respecto al Bitcoin. Con el clasificador modificado, se procedió a realizar el análisis de sentimientos agregando una nueva columna al dataset ("sentiment") que sirvió para implementar el modelo predictivo en sus diferentes configuraciones. Adicionalmente, se realizó un análisis correlativo utilizando el índice de correlación de Spearman para medir la relación estadística entre las variables respecto a la variable objetivo "Close" (precio de cierre del Bitcoin), de un total de 10 variables consideradas inicialmente para el análisis, solo 7 variables fueron las más significativas ("Replies", "Likes", "Retweets", "Sentiment", "Tweet_Volume", "Close" y "Volume"). Además, se explicó el motivo de la consideración de variables de alcance ("Replies", "Likes" y "Retweets") y su relevancia en este trabajo a modo de propuesta para medir la influencia de cada tweet y el efecto del mismo en el sentimiento del criptomercado.

CUARTA

Se implementó el modelo de predicción con sus distintas configuraciones utilizando redes neuronales recurrentes de tipo LSTM (Long Short Term Memory) ya que los datos obtenidos dependen del tiempo para tener en cuenta los valores pasados y presentes antes de predecir un valor futuro. Se trabajó con un lag temporal o lookback de 3 hrs. para predecir un valor a futuro, y se configuraron 3 modelos predictivos para predecir 1 hr., 6 hrs. y 12 hrs. a futuro. También, se dividió la data en entrenamiento y prueba considerando una proporción de 80 – 20, por lo que 27313 registros se utilizaron para la fase de entrenamiento y 6829 para la fase de prueba. Se utilizaron 250 épocas para el entrenamiento, para la función de pérdida (loss) se utilizó el mean absolute error (mae) y como optimizador se utilizó nadam ya que obtuvo los mejores resultados al compararlo con otros optimizadoras.

QUINTA

Se evaluó y comparó el performance de las diferentes configuraciones del modelo predictivo utilizando el RMSE y MAPE, y se determinó que la configuración del modelo más óptima es la configuración que predice 1 hora a futuro con un RMSE de 227.413, es decir, la diferencia entre el valor real y el predicho por el modelo es de $227.413 aproximadamente y un MAPE de 0.022 comparándolo con las otras 2 configuraciones. mientras que la configuración para predecir 6 hrs. a futuro ocupó el segundo lugar con un RMSE 372.079 y un MAPE de 0.035.

VI. RECOMENDACIONES

Aplicar otra librería o método para el reconocimiento del idioma de los tweets ya que la librería WhatTheLang no siempre reconoce el idioma de un tweet correctamente, se registraron casos en los que el reconocimiento del idioma era completamente erróneo.

Construir un diccionario de léxicos específicos del dominio de Bitcoin mucho más grande para VADER ya que 20 expresiones no son suficientes para representar de manera óptima el sentimiento del criptomercado.

Considerar para el análisis de sentimientos no solo a la población de habla inglesa sino también a la población de habla hispana u otras lenguas ya que el Bitcoin es un activo comerciado en casi todo el mundo y gran parte de los actores interactúan en otros idiomas.

Recopilar información no solo de Twitter sino también de otras plataformas como Reddit que es la segunda plataforma con más presencia de la comunidad de criptomonedas en el mundo.

VII. REFERENCIAS BIBLIOGRÁFICAS

Abraham, J., Higdon, D., Nelson, J., Ibarra, J., Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). *Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis Volumes and Sentiment Analysis*. *1*(3).

Arévalo Ascanio, J., & Estrada López, H. (2017). La toma de decisiones. Una revisión del tema. *Gerencia De Las Organizaciones. Un Enfoque Empresarial*. https://doi.org/10.17081/bonga/2824.c8

Awoke, T., Rout, M., Mohanty, L., & Satapathy, S. C. (2021). Bitcoin Price Prediction and Analysis Using Deep Learning Models. In *Lecture Notes in Networks and Systems* (Vol. 134, Issue February 2021). Springer Singapore. https://doi.org/10.1007/978-981-15-5397-4_63

BAdiola Ramos, J. (2019). *¿Podemos comerciar Bitcoin usando análisis de sentimiento sobre Twitter?* *1*(Junio), 56. https://repositorio.comillas.edu/xmlui/handle/11531/27168

BBVA. (2022). Criptomonedas: Ranking de usuarios y cajeros de criptomonedas en el mundo. *BLOCKCHAIN, CRIPTOMONEDAS*. https://www.bbva.ch/noticia/criptomonedas-ranking-de-usuarios-y-cajeros-de-criptomonedas-en-el-mundo/

Berrón Ruiz, E., & Régil López, M. V. (2018). Twitter como instrumento para fomentar la participación del profesorado en los cursos formativos. *@Tic. Revista D'Innovació Educativa*, *20*, 43. https://doi.org/10.7203/attic.20.10646

Bhagya Laxmi, K., Yamini, B., Rakshitha, C., & Keerthi, D. (2020). Twitter Sentiment Analysis Using VADER On Python. *International Journal for Research in Applied*

Science and Engineering Technology, *8*(9), 1023–1026. https://doi.org/10.22214/ijraset.2020.31647

bitcoin.org. (2022). *¿Cómo funciona Bitcoin?* https://bitcoin.org/es/como-funciona

Bucquet, P., Lermite, M., & Jo, A. (2019). *How many active crypto traders are there across the globe ?*

coindesk.com. (2022). *About Bitcoin*. https://www.coindesk.com/price/bitcoin/

CoinMarketCap. (2022). *Criptodivisas*. https://coinmarketcap.com/es/all/views/all/

CoinTracking. (2021). *The Best 85 Crypto Twitter Accounts to Follow*. CoinTracking. https://blog.cointracking.info/best-85-crypto-twitter-accounts-to-follow/

Delgado Mohatar, Ó., & Ortigosa Juárez, Á. (2019). Diseño e implementación de un monedero para criptomoneda Bitcoin [UNIVERSIDAD AUTONOMA DE MADRID]. In *UNIVERSIDAD AUTONOMA DE MADRID*. http://zaguan.unizar.es/TAZ/EUCS/2014/14180/TAZ-TFG-2014-408.pdf

Dixon, S. (2022). *Worldwide revenue of Twitter from 2010 to 2021*. https://www.statista.com/statistics/204211/worldwide-twitter-revenue/

Dolader, C., Bel, J., & Muñoz, J. (2017). La blockchain : fundamentos, aplicaciones y relación con otras tecnologías disruptivas. *Economía Industrial*, *405*, 33–40.

ESAN. (2017). *Bitcoin y las criptomonedas: Rentabilidad, riesgo y confianza*. https://www.esan.edu.pe/conexion-esan/bitcoin-y-las-criptomonedas-rentabilidad-riesgo-y-confianza

Espino Timón, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso.: Vol. I*

[Universitat Oberta de Catalunya]. http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117memòria.pdf

European Central Bank. (2022). Virtual currency schemes – a further analysis. In *European Central Bank* (Issue February). https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemesen.pdf

Ferdiansyah, Othman, S. H., Zahilah Raja Md Radzi, R., Stiawan, D., Sazaki, Y., & Ependi, U. (2019). A LSTM-Method for Bitcoin Price Prediction: A Case Study Yahoo Finance Stock Market. *ICECOS 2019 - 3rd International Conference on Electrical Engineering and Computer Science, Proceeding*, *June*, 206–210. https://doi.org/10.1109/ICECOS47637.2019.8984499

Fintech-Observatorio. (2017). Monederos Bitcoin: cómo almacenar bitcoins de forma segura. *Seminario Fintech*.

García López, M. (2018). *Aplicación de modelos de redes neuronales al modelado y predicción del precio de la electricidad en españa*. 24.

Gartner. (2022). *Predictive Modeling*. https://www.gartner.com/en/information-technology/glossary?glossarykeyword=predictive modeling&glossarycontext=ac

Gemini. (2022). *Bitcoin Historical Data*. https://www.cryptodatadownload.com/data/gemini/

Gomez Rodriguez, J. L. (2020). *Bitcoin , un activo de inversión alternativo | Bitcoin , an alternative investment asset*. *November*. https://doi.org/10.13140/RG.2.2.28801.68968

Hernández Sampieri, R., & Mendoza Torres, C. P. (2018). Metodología de la

investigación. In *Metodología de la investigación : las rutas cuantitativa, cualitativa y mixta*.

Hutto, C. J., & Gilbert, E. (2014). VADER : A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International AAAI Conference*, 216–225.

IBM. (2021). *Conceptos básicos de ayuda de CRISP-DM*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Investing.com. (2021). *Bitcoin*. Fusion Media Ltd. https://es.investing.com/crypto/bitcoin/historical-data

Izquierdo Cervera, E. (2018). *Bitcoin*. 4.

Jiang, X. (2020). Bitcoin Price Prediction Based on Deep Learning Methods. *Journal of Mathematical Finance*, *10*(01), 132–139. https://doi.org/10.4236/jmf.2020.101009

Kaggle. (2022). *Bitcoin tweets - 16M tweets*. https://www.kaggle.com/datasets/alaix14/bitcoin-tweets-20160101-to-20190329

Khedr, A. M., Arif, I., Pravija Raj, P. V., El-Bannany, M., Alhashmi, S. M., & Sreedharan, M. (2021). Cryptocurrency price prediction using traditional statistical and machine-learning techniques: A survey. *Intelligent Systems in Accounting, Finance and Management*, *28*(1), 3–34. https://doi.org/10.1002/isaf.1488

La Ley. (2022). Criptomonedas: ¿Existe regulación legal en el Perú? *La Ley*. https://laley.pe/art/13493/criptomonedas-existe-regulacion-legal-en-el-peru#:~:text=La legislación peruana no prohíbe,normativo que reglamente estas operaciones.

Lee, D. K. C., Guo, L., & Wang, Y. (2018). Cryptocurrency: A new investment

opportunity? *Journal of Alternative Investments*, *20*(3), 16–40. https://doi.org/10.3905/jai.2018.20.3.016

Lee, D., Lim, M., Park, H., Kang, Y., Park, J. S., Jang, G. J., & Kim, J. H. (2017). Long short-term memory recurrent neural network-based acoustic model using connectionist temporal classification on a large-scale training corpus. *China Communications*, *14*(9), 23–31. https://doi.org/10.1109/CC.2017.8068761

Li, Y., & Dai, W. (2019). Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model. *The Journal of Engineering - The 3rd Asian Conference on Artifical Intelligence Technology (ACAIT, 2019)*, *148*, 148–162.

Marín Vilca, D. G., & Pineda Torres, I. A. (2019). *Modelo predictivo Machine Learning aplicado a análisis de datos Hidrometeorológicos para un SAT en Represas*. 42.

Mena Roa, M. (2021). *La adopción de las criptomonedas en el mundo*. Statista. https://es.statista.com/grafico/18425/adopcion-de-las-criptomonedas-en-el-mundo/

Microsoft. (2022). *Modelos de minería de datos (Analysis Services - Minería de datos)*. https://docs.microsoft.com/es-es/analysis-services/data-mining/mining-models-analysis-services-data-mining?view=asallproducts-allversions&viewFallbackFrom=sql-server-2017#bkmk_mdlDefine

Noboa, B., & Navas, G. (2022). *ESTADO DEL ARTE ACTUAL CON RESPECTO AL BITCOIN, ¿QUÉ ES? Y ¿CÓMO FUNCIONA?* [UNIVERSIDAD POLITÉCNICA SALESIANA SEDE QUITO]. https://dspace.ups.edu.ec/bitstream/123456789/21960/1/UPS - TTS633.pdf

Osman, M. (2021). *Estadísticas Impresionantes de Twitter y Datos Importantes Sobre Nuestra Red Favorita*. KINSTA BLOG. https://kinsta.com/es/blog/estadisticas-

twitter/

Paez Guarnizo, E. P., & Monroy, A. F. (2020). Implementación De Un Modelo De Análisis De Sentimientos Con Respecto a La Jep Basado En Minería De Datos En Twitter. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2017). Sentiment analysis of Twitter data for predicting stock market movements. *International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 - Proceedings*, 1345–1350. https://doi.org/10.1109/SCOPES.2016.7955659

Pauli, P. A., & Soliani, V. I. (2019). Análisis de sentimiento, comparación de algoritmos predictivos y métodos utilizando un lexicon español. *Instituto Tecnológico de Buenos Aires-ITBA*, 1–47.

PCMag. (2022). *PCMAG*. Bitcoiner. https://www.pcmag.com/encyclopedia/term/bitcoiner#:~:text=A proponent of the Bitcoin,%2C developer%2C investor or user.

Pérez Sanjuán, A. (2019). *Aplicación para el análisis de sentimientos y tendencias en redes sociales*. 45. http://hdl.handle.net/10251/126316

Puig, X. (2022). *Volatilidad*. https://www.eleconomista.es/diccionario-de-economia/volatilidad

Raamkumar, A. S., Erdt, M., Vijayakumar, H., Rasmussen, E., & Theng, Y. L. (2018). Understanding the Twitter usage of humanities and social sciences academic journals. *Proceedings of the Association for Information Science and Technology*, *55*(1), 430–439. https://doi.org/10.1002/pra2.2018.14505501047

Rahouti, M., Xiong, K., & Ghani, N. (2018). Bitcoin Concepts, Threats, and Machine-

Learning Security Solutions. *IEEE Access*, *6*(c), 67189–67205. https://doi.org/10.1109/ACCESS.2018.2874539

Regal, A., Morzán, J., Fabbri, C., Herrera, G., Yaulli, G., Palomino, A., & Gil, C. (2019). Proyección del precio de criptomonedas basado en Tweets empleando LSTM. *Ingeniare. Revista Chilena de Ingeniería*, *27*(4), 696–706. https://doi.org/10.4067/s0718-33052019000400696

Restrepo Betancur, L. F., Garcia Henao, G., & Arboleda Zapata, E. (2020). *EL PODER DE TWITTER EN LA COMUNICACIÓN INVESTIGATIVA EN LAS ÁREAS DE LA EDUCACIÓN, MARKETING Y POLÍTICA*. 1–14.

Rodriguez Garagorri, M. (2017). *Análisis de tecnologías Bitcoin y Blockchain*. 61. http://hdl.handle.net/10609/72606

Rosenbrock, G., Trossero, S., & Pascal, A. (2019). Técnicas de análisis de sentimientos aplicadas a la extracción de opiniones en el lenguaje español. *XXI Workshop de Investigadores En Ciencias de La Computación (WICC 2019, Universidad Nacional de San Juan).*, *November*, 291–300.

RPP. (2021). Las criptomonedas en el 2021: este es el balance en el año. *Tecnología*. https://rpp.pe/tecnologia/mas-tecnologia/las-criptomonedas-en-el-2021-este-es-el-balance-en-el-ano-noticia-1375266?ref=rpp

Sabalionis, A., Wang, W., & Park, H. (2021). What affects the price movements in Bitcoin and Ethereum? *Manchester School*, *89*(1), 102–127. https://doi.org/10.1111/manc.12352

Sangeeth, K., & Manoj. (2017). *Whatthelang library*. https://github.com/indix/whatthelang

Sarmiento, J., & Garcés, J. (2016). Criptodivisas en el entorno global y su incidencia en Colombia. *Revista Lebret*, *8*, 1–11.

Sattarov, O., Muminov, A., Lee, C. W., Kang, H. K., Oh, R., Ahn, J., Oh, H. J., & Jeon, H. S. (2020). Recommending cryptocurrency trading points with deep reinforcement learning approach. *Applied Sciences (Switzerland)*, *10*(4). https://doi.org/10.3390/app10041506

Sheikh, H., Azmathullah, R. M., & Rizwan, F. (2018). Proof-of-Work Vs Proof-of-Stake: A Comparative Analysis and an Approach to Blockchain Consensus Mechanism. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, *887*(Xii), 2321–9653. www.ijraset.com

SNGULAR. (2022). *CRISP-DM: La metodología para poner orden en los proyectos*. https://www.sngular.com/es/data-science-crisp-dm-metodologia/

Sobrino Sande, J. C. (2018). *Análisis de Sentimientos en Twitter*.

Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis. *Diva*, 37. http://www.diva-portal.org/smash/record.jsf?pid=diva2:1110776%0Ahttp://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-209191

Subramanyam, J. (2019). *Prediction Models: Traditional versus Machine Learning*. Gartner. https://blogs.gartner.com/jitendra-subramanyam/prediction-models-traditional-versus-machine-learning/

Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062.

https://doi.org/10.1002/asi.23833

Tovanich, N., Soulie, N., Heulot, N., & Isenberg, P. (2022). The Evolution of Mining Pools and Miners', Behaviors in the Bitcoin Blockchain. *IEEE Transactions on Network and Service Management*. https://doi.org/10.1109/TNSM.2022.3159004

Twitter. (2022). *Glosario Twitter*. https://help.twitter.com/es/resources/glossary

Valdez Alvarado, A. R. (2019). Machine Learning para Todos. *Reseaechgate*, *January 2019*, 61. https://doi.org/10.13140/RG.2.2.13786.70086

Wołk, K. (2020). Advanced social media sentiment analysis for short-term cryptocurrency price prediction. *Expert Systems*, *37*(2), 1–16. https://doi.org/10.1111/exsy.12493

Zarraluqui Matos, I. (2018). *Análisis De Las Criptomonedas En La Economía Actual*.

Zocaro, M. (2021). *LA MINERÍA DE CRIPTOMONEDAS Y SU TRIBUTACIÓN EN ARGENTINA*. *0*(0), 10. https://www.bitcoinmining.com/es/