# Proyección del precio de criptomonedas basado en *Tweets* empleando LSTM

*Cryptocurrency price projection based on Tweets using LSTM*

Andrés Regal[1]* Juandiego Morzán[1] Carlos Fabbri[1] Gonzalo Herrera[1]
Gabriela Yaulli[1] Andrea Palomino[1] Claudia Gil[1]

Recibido 08 de mayo de 2019, Received: May 08, 2019
*Aceptado 15 de junio de 2019 Acepted: June 15, 2019*

**RESUMEN**

El modelamiento y predicción de series temporales constituye una tarea ardua y esencial para los procedimientos de optimización financiera. Numerosos estudios han sido elaborados con la finalidad de reducir la incertidumbre del inversor, mediante el pronóstico de precio de monedas y acciones. Sin embargo, el surgimiento de un nuevo tipo de monedas con características propias, conocidas como cryptocurrencies, plantea retos adicionales. En este sentido, el paper plantea analizar en qué medida las publicaciones en las redes sociales pueden capturar las expectativas colectivas de los inversores, y afectar el valor futuro de la moneda. Nuestro 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, proponemos el uso de un tipo de red neuronal recurrente, conocida como "*Long Short Term Memory*" (LSTM). La metodología empleada para el preprocesamiento de los datos y la aplicación de esta técnica de pronóstico de series temporales nos permite obtener una predicción con un Error Porcentual Absoluto Medio de 34.92%; lo que indica que la representación de la variable de percepción en redes social no ha sido la pertinente y, por lo tanto, motiva nuevos trabajos con la finalidad de modelar esta variable mediante el uso de otras técnicas de NLP.

Palabras clave: Cryptocurrencies, Twitter, LSTM.

***ABSTRACT***

*The modeling and prediction of time series is an arduous and essential task for financial optimization procedures. Numerous studies have been carried out to reduce investor uncertainty, by forecasting the price of currencies and shares. However, the emergence of a new type of coins with their own characteristics, known as cryptocurrencies, present additional challenges. In this sense, the paper seeks to understand up to what extent comments in social networks can capture the collective expectations of investors, and affect the future value of the currency. The objective is to predict the daily performance of a market based on two components: those that define the behavior of the cryptocurrency itself (volume, opening value, closing value, maximum value and minimum value) and the expectations and interactions of the environment, through the collected tweets. For this, the use of a type of recurrent neural network known as "Long Short Term Memory" (LSTM) is proposed. The methodology used for the preprocessing of the data and the application of this time series forecasting technique allows obtaining a prediction with a Mean*

[1] Universidad del Pacífico. Facultad de Ingeniería. Lima, Perú.
E-mail: a.regalludowieg@alum.up.edu.pe; j.morzansamame@alum.up.edu.pe; c.fabbrigarcia@alum.up.edu.pe; g.herreramedina@alum.up.edu.pe; cg.yaullih@alum.up.edu.pe; a.palominovargas@alum.up.edu.pe; c.giljuscamaita@alum.up.edu.pe
* Autor de correspondencia: a.regalludowieg@alum.up.edu.pe

*Absolute Percent of 34.92%; This indicates that the representation of the perception variable in social networks has not been that relevant and, therefore, motivates new works for better representation using other NLP techniques.*

*Keywords: Cryptocurrencies, Twitter, LSTM.*

## INTRODUCCIÓN

En el mundo económico y financiero, se pueden identificar diferentes tipos de mercados. Dentro de ellos, se encuentran dos mercados tradicionales particularmente interesantes para el presente análisis: el mercado de divisas y el mercado de valores. Por un lado, el mercado internacional de divisas es aquel en el que participantes –bancos, corporaciones, corredores e inversores minoristas de divisas– alrededor del mundo compran y venden diferentes monedas con la finalidad de facilitar el comercio internacional o reducir el riesgo asociado a las fluctuaciones de precios.

Por otro lado, el mercado de valores es un conjunto de mercados e intercambios donde empresas e inversores emiten y negocian acciones, bonos y otras clases de valores. En ese sentido, permite a las empresas recaudar dinero ofreciendo acciones y bonos corporativos; y a los inversores, participar en los logros financieros de las empresas, generando dinero a través de dividendos.

Ambos mercados –divisas y valores– presentan una serie de características que las hacen parecidas. Sin embargo, es su finalidad lo que las hace similares. En ambos casos, el inversionista busca maximizar su utilidad enfrentándose a diferentes combinaciones y variaciones de retornos esperados y riesgo. Entonces, el objetivo final es conseguir el mejor valor del *trade-off* entre retorno y riesgo.

No obstante, debido al comportamiento volátil e incierto de las divisas y valores, así como otros factores como la asimetría de información, los inversionistas se encuentran ante un conjunto de disyuntivas: cuáles divisas o valores comprar o vender, cuánto comprar o vender de cada uno y cuándo hacerlo. Además, surge la cuestión de qué información utilizar para apoyar su proceso de toma de decisiones.

Por ello, una serie de técnicas se han desarrollado para reducir la incertidumbre y para brindar soporte al proceso de decisión. Una de las más importantes es la técnica de pronósticos. Esta técnica busca predecir el valor de una variable en el tiempo t usando como datos de entrada los valores de la variable en una ventana de tiempos anteriores. Específicamente, en este tipo de mercados se busca predecir el valor de cierre de una divisa o acción, a partir de un conjunto de variables (e.g. valor de cierre, valor de apertura, valor máximo, valor mínimo, volumen) evaluados en una ventana de tiempo.

Recientemente, ha aparecido un nuevo tipo de mercado que, aunque difiere de los mencionados anteriormente, guarda cierta similaridad: el mercado de criptomonedas. Las criptomonedas son monedas digitales seguras y descentralizadas, cuya creación está controlada por la criptografía. Específicamente, es un activo que sirve como medio de intercambio para asegurar transacciones financieras, controlar la creación de unidades adicionales y verificar la transferencia de activos.

El mercado de criptomonedas es similar al de divisas y acciones en el sentido que se transan en el mercado en donde interactúan las fuerzas de la oferta y demanda antes mencionadas. Además, se enfrentan a las mismas disyuntivas y se requiere de técnicas que permitan reducir la incertidumbre y brindar soporte al proceso de decisión antes descrito (cuáles, cuánto y cuándo). Nuevamente, el pronóstico funge un rol importante, dado que las variables utilizadas son virtualmente las mismas (valor máximo, valor mínimo, volumen).

No obstante, se pueden identificar diferencias entre ellos. Por un lado, la disponibilidad de información para el mercado de criptomonedas debe ser 24/365. Por otro lado, la naturaleza digital y global de las criptomonedas las vuelven aún más volátiles y susceptibles a grandes oscilaciones de valor. En ese sentido, se requiere y emite más información. Sin embargo, ante una mayor asimetría de información en este tipo de mercado, las expectativas de los inversores se ven moldeadas en mayor medida

por las expectativas de otros inversores o ciertas comunidades [1].

Entonces, surge una interrogante: ¿qué fuente de información existe que cumpla con ser 24/365 y plasme las interacciones y expectativas de los inversores? Con el desarrollo de la Web 2.0, se generaron espacios en la web que cumplen con estos requisitos: las redes sociales. Las redes sociales también operan 24 horas al día, 7 días a la semana, 365 días al año, por lo que los inversores de criptomonedas naturalmente dirigen su atención a fuentes como Reddit y Twitter cuando debaten sobre nuevos desarrollos y posibles valoraciones [2]. Al confiar en estas redes para obtener noticias de última hora, los operadores de criptomonedas pueden administrar mejor sus riesgos y tomar decisiones más inteligentes.

En el presente trabajo, se propone una metodología que busca combinar información del comportamiento de las criptomonedas en sí, así como de las expectativas e interacciones de inversores en las redes sociales. Además, dado el mayor grado de volatilidad y oscilaciones de las criptomonedas, se deben considerar tanto ventanas cortas (memorias a corto plazo) como ventanas largas de tiempo (memoria de largo plazo). Para ello, se plantea elaborar una técnica de pronósticos basada en *Long Short Term Memory* (LSTM) y *Recurrent Neural Networks* (RNN) que utilice como predictores los valores previos asociados a las criptomonedas y a las expectativas construidas entorno a las mismas (expresadas a través de *Sentiment Analysis* de *tweets*). El resto del artículo está organizado de la siguiente manera: (ii) Estado del Arte, (iii) Metodología, (iv) Resultados, (v) Conclusiones y recomendaciones.

## CRIPTOMONEDAS Y BLOCKCHAIN

En [3] se define las criptomonedas como activos digitales que funcionan como un medio de intercambio basado en sistemas de criptografía que aseguran las transacciones, controlan la creación de monedas y verifican la transferencia segura de activos. En [4] se propone que la diferencia esencial que existe entre las criptomonedas del resto de monedas es que las primeras están basadas en el principio de control descentralizado, es decir, no es necesaria la existencia de un organismo central que controle la moneda y legitimice su valor. El análogo a este organismo dentro del contexto de criptomonedas, es el sistema mismo de encriptación, comúnmente conocido como Blockchain.

Las criptomonedas datan del 2008, año en que Satoshi Nakamoto publica su paper titulado "Bitcoin: A Peer-to-Peer Electronic Cash System" [5], sentando las bases necesarias para el funcionamiento de un sistema de monedas electrónicas sin necesidad de un aparato central.

Como ya ha sido mencionado, la tecnología que hace posible la presencia de Bitcoin y otras criptomonedas en la economía actual es conocida como Blockchain. El trabajo de [6] modela sencillamente el sistema. Todas las transacciones forman parte de una red y son escuchadas por todos los participantes. Todo el historial de intercambios se registra dentro del Blockchain y puede ser verificado por todos los participantes. La unidad llamada "block" o bloque incluye una cantidad fija de las transacciones más recientes y el valor "hash" del bloque previo. Este bloque crea una representación irreversible de estas transacciones mediante una función "hash" y se convierte en el último bloque de la cadena momentáneamente hasta que aparezca un bloque más reciente que incluya las transacciones nuevas. Generar el bloque "ganador" que encripta todas las transacciones toma una cierta cantidad de tiempo, normalmente fijada en 10 minutos. Es por esto que los bloques "ganadores" rara vez provienen del mismo creador y elimina la posibilidad de que una persona pueda inventar el bloque ganador ilimitadas veces aprovechándose de este poder. La dinámica de Blockchain se ilustra en la Figura 1.

La ola de criptomonedas desencadenada por el éxito de Bitcoin desde el 2009 ha recibido mucha atención de los medios y los inversionistas debido a las características de estos activos, su potencial como herramientas transaccionales y sus impresionantes fluctuaciones de precios [6]. En los últimos dos años, la capitalización total de mercado del mercado de criptomonedas ha crecido en 11,600% [8]. Este crecimiento exponencial es el resultado de la creciente especulación de los inversores por un lado y de la introducción de nuevas criptomonedas por otro. En el presente estudio trabajaremos con la criptomoneda líder del mercado: Bitcoin.

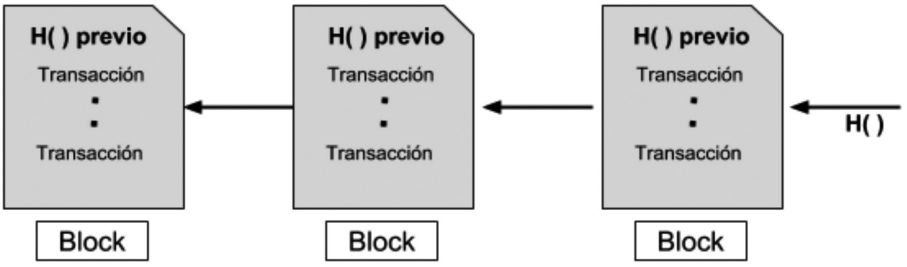


Figura 1. Esquema de la dinámica de Blockchain.

## ESTADO DEL ARTE

En esta sección, se señalarán y discutirán distintos estudios relacionados a los temas centrales de esta investigación, con el objetivo de diseñar correctamente la metodología del estudio y comprobar la originalidad del mismo. Se plantea revisar el Estado del Arte en un orden lógico específico. Primero se expondrá un primer grupo de investigaciones pasadas en las que se haya propuesto la predicción de precios en un mercado financiero; sean estos precios de monedas, criptomonedas, acciones u otros. El enfoque aquí es reconocer y registrar las principales técnicas y marcos aplicados en la predicción de valores que dependen de leyes de oferta y demanda. Paralelamente, establecer ciertos umbrales de comparación o *benchmarks* para tener disponibles al concluir sobre nuestros resultados.

Una vez presentadas las publicaciones del grupo anterior, se pasará a examinar un grupo más selecto de estudios que presenten el detalle de incluir variables provenientes de redes sociales, o que, de alguna otra manera, consiguen modelar el efecto de interacciones sociales dentro de una predicción de valores en mercados financieros. Esto permitirá dirigirse mejor al objetivo del presente estudio que busca, justamente, apalancarse de información social para ofrecer una predicción certera sobre los precios futuros de las criptomonedas.

El trabajo realizado por [6] ofrece un estudio muy didáctico con respecto a las bases teóricas que hacen posible las criptomonedas para justificar la inclusión de variables de *blockchain* en el análisis. En su caso buscan comparar las redes neuronales bayesianas con otros modelos predictivos en su capacidad para predecir el precio de Bitcoin. Lo novedoso de su estudio es que una cantidad importante de variables trata netamente de la tecnología que soporta a la criptomoneda. Por eso encontramos atributos como: tamaño promedio de bloque, transacciones por bloque, tiempo de confirmación medio, ratio de hash, entre otras.

Además, se introduce la estrategia de Rollover para las redes bayesianas sobre la cual queda oportunidad de indagar más pues no es el foco de esa investigación, pero muestra resultados interesantes. Esta técnica consiste en escoger una "ventana" de los datos ordenados temporalmente y entrenar el algoritmo siempre con una cantidad fija de datos en vez de con todo el *dataset*. La idea de esto es que la información más relevante con respecto al comportamiento de la moneda se encuentra en los últimos registros disponibles y no es necesario analizar todo el histórico disponible.

La conclusión del trabajo es que se encontró evidencia empírica que indica que la volatilidad en los precios del Bitcoin se debe, en una gran parte, a la información del blockchain directamente involucrada en la oferta y demanda del Bitcoin y no a otras variables macro-financieras como se pensaba.

En el trabajo de [9] se realiza un estudio centrado en predecir el precio de las acciones en operaciones intradía (inversiones en acciones que se cierran antes que acabe la jornada). Para esto propone un sistema de predicción que utiliza *Singular Spectrum Analysis* (SSA) para descomponer las series temporales, *Support Vector Regression* (SVR) en el proceso de aprendizaje y aproximación, y *Particle Swarm Optimization* (PSO) para optimizar los parámetros iniciales del SVR.

Compara los resultados de este enfoque híbrido con los de modelos ampliamente utilizados para

pronósticos financieros como WT-FFNN, ARMA, PolyReg y Naive. El estudio resulta interesante, pues se centra en la utilización de métodos no paramétricos, como el SSA, para el análisis y predicción de series de tiempo financieras ruidosas. Los resultados de la aplicación de este modelo sobre seis series temporales de precios de acciones demuestran su superioridad frente a los demás, alcanzando el menor *Mean Average Error* (MAE), *Mean Absolute Percentage Error* (MAPE) y el *Root Mean Squared Error* (RMSE). Esto demuestra las ventajas que se obtienen de fusionar distintas técnicas en el desarrollo del modelo de pronóstico de la variable en estudio.

En la investigación de [10] se utilizan siete reglas financieras que apoyan en la toma de decisiones respecto a la venta y compra de acciones: *Simple Moving Average, Exponential Moving Average, Moving Average Convergence/Divergence, Relative Strength Index, Stochastic Oscillator, Bollinger Bands* y *Accumulation/Distribution Line.*

Con la información de precios diarios de una acción, cada regla especifica si vender, comprar o no hacer nada respecto a la misma. Con resultados anteriores de un stock de acciones se entrenó un algoritmo de *Random Forest* y un modelo híbrido entre *Gradient Boosting* y *Random Forest: Gradient Boosted Random Forest,* de manera que, en base a las reglas mencionadas, se dé un "veredicto final".

En los resultados, destacó el desempeño del modelo de *Gradient Boosted Random Forest* en rentabilidad, superando al clásico *buy-and-hold* y a los otros modelos propuestos asociados con cada regla financiera.

En [11] se busca pronosticar el precio de la electricidad en Colombia, la cual tiene un precio altamente volátil. Para ello, se proponen dos modelos: una red neuronal artificial (ANN), cuya variable de entrada será únicamente el precio diario de la electricidad, y una ANN, cuyas variables de entradas serán el precio diario de la electricidad y el nivel medio de embalses. Ambas redes se comparan con un modelo Autorregresivo Condicional Heterocedástico Generalizado (GARCH), utilizando los datos de 150 días anteriores.

Se da como resultado que la red neuronal con dos variables de entradas tenía un mejor desempeño en el pronóstico. Además, se encontró que el modelo GARCH requería muchos reajustes, por lo que el uso de redes neuronales era más sencillo, e incluso daba mejores resultados fuera de la muestra. Para el lector que desee revisar más estudios de esta índole, se le invita a consultar [12, 13].

Ahora se pasará a discutir el segundo grupo de estudios, que afinan aún más el ámbito de la investigación y lo llevan a un terreno más social apoyándose, por lo general, en técnicas de análisis de sentimientos.

Para el lector que se encuentre escéptico ante el uso de variables provenientes de redes sociales para predecir precios de criptomonedas, lo invitamos a referirse a [14, 15, 16, 17] que son solo algunas de las investigaciones en donde quedó comprobado la propiedad de plataformas como Twitter para monitorear el sentimiento de los inversores y cambios enlos precios del mercado de Bitcoin y otras criptomonedas.

El estudio de [7] es la primera prueba que se ofrece sobre la efectividad de datos de redes sociales para predecir precios de una criptomoneda alternativa *(alt-coin)* como ZClassic. En la investigación, logran reunir *tweets* sobre la criptomoneda bajo análisis y, con técnicas de *sentiment-analysis*, identifican si el *tweet* indicaba algo positivo, negativo o neutro sobre la moneda. Tomando estos *tweets* (agrupados por hora), como dato, luego se procede a aplicar un modelo, *Extreme Gradient Boosting* o XGBoost, para predecir el precio horario de la criptomoneda. Los resultados del estudio son alentadores, pues consiguen demostrar que tomando únicamente datos de Twitter es posible predecir con una alta confianza las fluctuaciones de la criptomoneda.

Esto permite concluir a los investigadores que el mercado de *alt-coins* tiene una fuerte dependencia de las expectativas de los inversionistas de criptomonedas. También se desprende que Twitter es una plataforma en donde se puede capturar exitosamente el sentimiento de los inversionistas y convertir estos en señales tempranas de las fluctuaciones en *alt-coins*.

En [18] se lleva a cabo un estudio que tiene como principal objetivo proponer una estrategia para determinar si el precio de una criptomoneda se

incrementará o disminuirá, correlacionando esta variable con datos obtenidos de plataformas de redes sociales. En este caso particular, se utilizan *tweets* extraídos durante un periodo de tiempo determinado. Para los procesos de clasificación de textos y análisis de sentimientos se comparan los resultados obtenidos de la aplicación de SVM, Regresiones logísticas y Naive Bayes.

Cabe resaltar que en el estudio se proponen dos enfoques para el entrenamiento de los clasificadores: formar los vectores característicos utilizando directamente los textos de cada *tweet* o, generar los mismos con las puntuaciones positivas y negativas de las palabras de cada post, obtenidas de APIs de terceros. Los resultados muestran que, con el primer enfoque, Naive Bayes obtuvo el mejor resultado entre los algoritmos de clasificación, alcanzando un *accuracy* por día de 95% y por hora de 76,23%. Mientras que, siguiendo el segundo enfoque, con regresión logística se obtuvo el mayor *accuracy* por día (86%) y por hora (98,58%).

[19] ofrece un estudio bastante similar, aplicando únicamente métodos naive. Su aporte radica en los distintos rangos de tiempo que consideraban para agrupar los *tweets* (desde 5 minutos a 4 horas). Su modelo demostró que agrupar los *tweets* en periodos de 30 minutos, ofrecía el mejor rendimiento con 79% de *accuracy*.

## METODOLOGÍA

En esta sección se detallará el preprocesamiento de datos, modelos a aplicar y los métodos de validación a utilizar en este trabajo. Dentro del preprocesamiento se describe el tratamiento de variables continuas y textuales a utilizar. El modelo principal será una red neuronal recurrente con arquitectura LSTM. Como métricas de validación se utilizará el RMSE y el MAPE. La estructura completa de la metodología se detalla en la Figura 2.

### A. Dataset

La primera fase de esta metodología se enfoca en la recolección de datos. Para los datos cuantitativos de precios y volumen transado en un periodo se utilizará la base de datos de Kaggle "Bitcoin Historical Data", la cual cuenta con información histórica minuto a minuto del precio de apertura, el precio máximo y mínimo de Bitcoin desde 2012 a marzo de 2018. Si bien es posible extraer toda la base de datos, para este estudio se enfocará en los datos históricos para 2018. Esta base de datos consiste de cinco campos principales: el precio de apertura del periodo, el precio máximo, el precio mínimo, el precio de cierre del periodo y el volumen transado. Se consiguió un total de 512 000 observaciones. En la Figura 3 se muestra el precio de cierre del BitCoin como serie temporal, de esta gráfica destaca la falta de estacionalidad en media y varianza, lo que incentiva el uso de redes neuronales para capturar de manera no lineal el comportamiento de los datos. Asimismo, en la Figura 4, se muestran los histogramas para los predictores (los precios de apertura, máximo y mínimo) en los que se aprecia una distribución asimétrica a la derecha. Esta misma asimetría se explica al tener precios más altos alrededor de 2012 y una tendencia decreciente Esta información se encuentra agregada en intervalos de un minuto, por lo que es particularmente atractivo utilizar modelos orientados a series de tiempo.

Una vez almacenada esta información se recolectará datos de redes sociales. Esta información en formato de comentarios o *tweets* será especialmente útil para poder cuantificar el efecto de la especulación y la influencia de las redes sociales en la demanda de

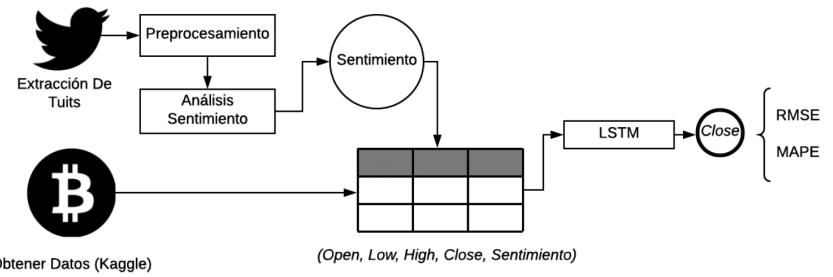


Figura 2. Esquema de la metodología propuesta.


Figura 3. Serie de precios de cierre de BitCoin.

criptomonedas y, por lo tanto, su precio. La cantidad de *tweets* extraídos correspondientes al periodo de la variación del precio de las acciones de las criptomonedas fue de 41 002. Estos datos textuales tienen que ser transformados a una representación cuantitativa. Para llegar a esta representación, previamente se realiza un preprocesamiento para reducir el ruido en el set de *tweets*. Este es detallado a continuación.

**B. Pre-procesamiento**

En primer lugar, se removieron todos los espacios en blanco excedentes y se transformó todo el texto a letras minúsculas. El segundo paso consistió en remover todos los caracteres no alfabéticos; signos de puntuación y los símbolos "#" y "@", correspondientes a *hashtags* y menciones a usuarios usados en esta plataforma social. Asimismo, con ayuda de expresiones regulares (RegEx) se eliminaron los *links* de páginas web incluidos en los *tweets*.

El tercer paso consideró un caso común en las publicaciones en redes sociales: la utilización de una o más letras repetidas en la redacción. Se pasó a reducir a un máximo de dos letras repetidas por cada palabra (i.e "*hellooooo*" → "*helloo*"). Cabe resaltar que, si bien en este ejemplo no se lleva la palabra "*hellooo*" a su forma correcta "*hello*", se consigue llegar a una estandarización para todas las ocurrencias con más de dos letras repetidas.

En el cuarto paso, las palabras extremadamente comunes (artículos, preposiciones, conjunciones, entre otros) fueron eliminadas de los *tweets* debido a que por sí solas no generan un valor, basándose en la pertenencia a un corpus en inglés de *stopwords*, definido por *Natural Language Toolkit*. Este fue complementado con un diccionario de elaboración propia para mejorar la tarea de filtrado ya que se consideró relevante para este estudio los nombres de las criptomonedas y los adjetivos que las acompañaban.

Posteriormente, se llevó a cabo el proceso de lematización de palabras. Con la finalidad de que esa sea más exacta, se comenzó por un etiquetado *gramatical POS (part of speech)* de cada palabra dentro del corpus, basado tanto en su definición como en su contexto.

Una vez que su función gramatical (verbo, sustantivo, adjetivo, entre otros) sido determinada, se obtiene el lema de cada palabra, o forma de diccionario. Este procedimiento permite agrupar las inflexiones y variaciones de una palabra para que pueda analizarse como un solo elemento.

Por último, se pasó a la tokenización, donde se separó el documento en unidades menores de análisis, llamadas *tokens*. De esta forma cada *tweet* queda representado por una estructura similar a un vector de características, conformado por sus palabras o *tokens*. Finalmente, esta estructura es utilizada para vectorizar numéricamente el documento, con la herramienta TextBlob.

**C. Análisis de Sentimiento**

Una vez terminada esta limpieza, la tarea siguiente será realizar el Análisis de Sentimientos para cada

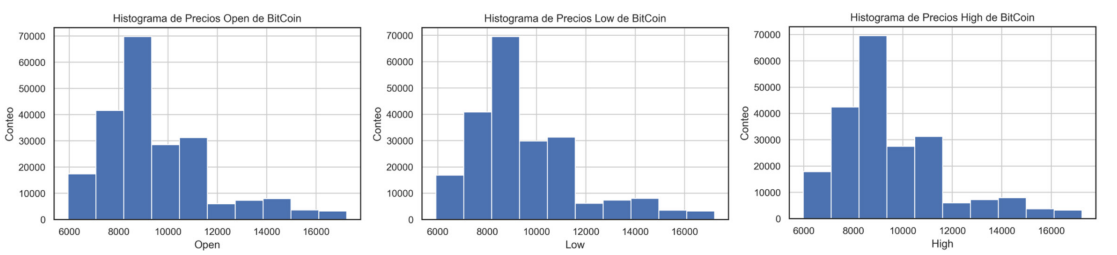


Figura 4. Histogramas de los precios de apertura, máximos y mínimos.

representación vectorial de los *tweets* de la base de datos. Para esto, se utilizará el paquete *TextBlob Sentiment*. Este genera un puntaje de la polaridad de cada texto comprendido en un rango de –1,0 a 1,0, donde 0 indica neutralidad, +1 una actitud muy positiva y –1 una actitud muy negativa. Con estos valores, se calculará un promedio de la polaridad de los *tweets* por cada día. Este es el dato que corresponde al *input* que se introducirá en la red neuronal.

**D. Aplicación de ML para predicción**

Una vez terminada la fase de recolección de datos, se aplicará el modelo para predecir el precio de cierre de cada criptomoneda. Una red neuronal recurrente LSTM tiene como objetivo aprender dependencias a largo plazo; es decir, aprender las dependencias de valores futuros de una secuencia en función a los valores anteriores. En [20] introducen el concepto de una red LSTM, la cual busca eliminar el problema que nace al tratar de integrar t-n valores para predecir el valor $t + 1$ de una secuencia. Como toda red neuronal recurrente, una LSTM tiene la forma de una cadena con unidades (o celdas) repetidas (Figura 5) pero en vez de tener una estructura simple dentro de cada unidad, en una LSTM se tiene 4 componentes que interactúan entre sí (Figura 6).

Fuente: https://colah.github.io/

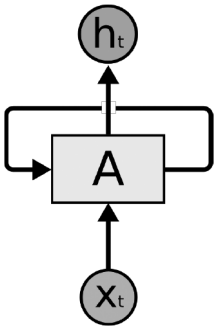


Figura 5. Estructura de una red neuronal recurrente.

Fuente: https://colah.github.io/

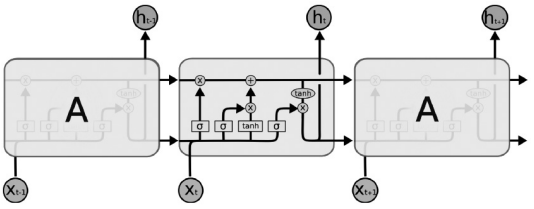


Figura 6. Estructura de una red LSTM.

El *core* de una red LSTM se basa en el estado de una celda. La red LSTM modifica el estado de la celda al añadir o quitar información del estado de la celda anterior $C_{t-1}$ para producir un nuevo estado $C_{t-1}$. El primer paso en una LSTM consiste en decidir qué información va a permanecer en el estado de la celda. Este procedimiento se realiza con una capa de activación sigmoidal, la cual toma la salida $h_{t-1}$ y la entrada $x_t$, y representa con valores en el intervalo [0,1] la necesidad de información (un valor más cercano a uno es más necesario que uno cercano a cero) para cada valor del estado $C_{t-1}$. Matemáticamente, este procedimiento se detalla en la ecuación (1), donde $W_f$ representa la matriz de pesos y $b_f$ el *bias* para el "*forget gate*".

$$f_t = \sigma\left(W_f \cdot [x_t, h_t - 1] + b_f\right) \tag{1}$$

El siguiente paso es decidir qué información retener en el estado de la celda. Para esto se utilizan dos capas, una sigmoide con la que se decide qué valores se actualizarán y una tangente hiperbólica que creará un nuevo vector de valores candidatos a ser añadidos al estado de la celda (ecuaciones (2) y (3)). Estos dos valores serán combinados más adelante para la actualización del estado de la celda.

$$i_t = \sigma\left(W_i \cdot [x_t, h_t - 1] + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot [x_t, h_t - 1] + b_C\right) \tag{3}$$

Al terminar estos cálculos, el nuevo estado $C_t$ se obtiene al multiplicar $C_{t-1}$ por $f_t$, olvidando la información innecesaria y añadiendo la nueva información al multiplicar $i_t$ por la matriz de candidatos $\alpha$ (ecuación (4)). Estos candidatos son escalados por $i_t$, ya que este representa cuánto se decidió actualizar cada valor del estado anterior.

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4}$$

Finalmente, se necesita calcular la salida. Esta se debe basar en el estado actual, pero no toda la información es parte de la salida. Para realizar este filtrado, se aplica una capa sigmoidal para decidir qué valores del estado actual formarán parte de la salida. Luego, se aplica una tangente hiperbólica (para tener valores entre –1 y 1) y se multiplica estos valores por la activación de la capa sigmoidal para solo tener como output los valores seleccionados (ecuaciones (5) y (6)).

$$o_t = \tanh\left(W_0 \cdot [x_t, h_t - 1] + b_0\right]) \quad (5)$$

$$h_t = o_t * \tanh\left(h_t\right) \quad (6)$$

Esta arquitectura es particularmente interesante, ya que ataca el problema del "*vanishing gradient*" [21, 22]; donde al aplicar *back propagation* en redes recurrentes el vector de gradiente se vuelve tan pequeño que la actualización de pesos toma demasiado tiempo o se excede la cantidad de memoria disponible. Esta arquitectura permite entender patrones interesantes en las secuencias de datos como las que se presentan con series temporales de precios de criptomonedas. A continuación, se muestra en la Figura 7 la arquitectura de red aplicada en el presente trabajo (que se resume en 32 procesadores por capa).

La tercera y última fase de esta metodología se centra en evaluar el performance del modelo. Para esto se utilizarán dos medidas de error: RMSE y MAPE (ecuaciones (7) y (8)). Con estos se podría obtener valores comparables con otros modelos en términos porcentuales (MAPE) y validación interna del error del modelo (RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad (7)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \hat{y}_i\right|}{y_i} \quad (8)$$

## RESULTADOS

En esta sección se detallarán los resultados obtenidos al utilizar una red neuronal recurrente LSTM para la predicción de precios de BitCoin. Para este trabajo se utilizó una arquitectura recurrente con una capa de *dropout* y una capa densa para realizar la regresión. Para el *dropout* se utilizó una probabilidad de "cierre" para una neurona de 25%. Se utilizó una capa densa (o *fully connected*) sin función de activación para obtener el valor final de la regresión.

En la Figura 8, se presenta el ajuste de las predicciones de la red al valor real. Al tener miles de registros, detectar las diferencias visualmente es complicado. La gráfica de ambas series presenta buen ajuste excepto en picos en subidas y bajadas repentinas (característicos de una criptomoneda).

Para analizar de manera más cuantitativa estos resultados, es necesario calcular el RMSE y MAPE de este modelo. En este caso se utilizó 50 épocas (cantidad de veces que se completa una iteración), 32 neuronas (o *hidden units*). Se obtuvo un RMSE de 0.0704 y un MAPE de 34.92%. Este último es bastante elevado, orientando el análisis a la perspectiva de un inversionista, un error de +–34% es inaceptable, representa un riesgo demasiado alto para la compensación que puede traer en la siguiente hora. Asimismo, al analizar la relación entre las curvas de loss y validation loss, estas no convergen. Esta falta de convergencia nos indica que el problema que se está tratando de entrenar es demasiado complejo para la arquitectura que se utiliza actualmente. Un incremento en el número de procesadores o capas podría traer mejor performance.

Finalmente, los resultados no han sido favorables. Se obtuvo un MAPE y RMSE superior a los benchmarks establecidos en la revisión de literatura, pero todavía hay espacios de mejora como se mencionó anteriormente mediante el testeo de distintas arquitecturas de red.

| Layer (type) | Output Shape | Param # |
|---|---|---|
| lstm_1 (LSTM) | (None, 32) | 4736 |
| dropout_1 (Dropout) | (None, 32) | 0 |
| dense_1 (Dense) | (None, 1) | 33 |

Total params: 4,769
Trainable params: 4,769
Non-trainable params: 0

Figura 7. Arquitectura de la red.

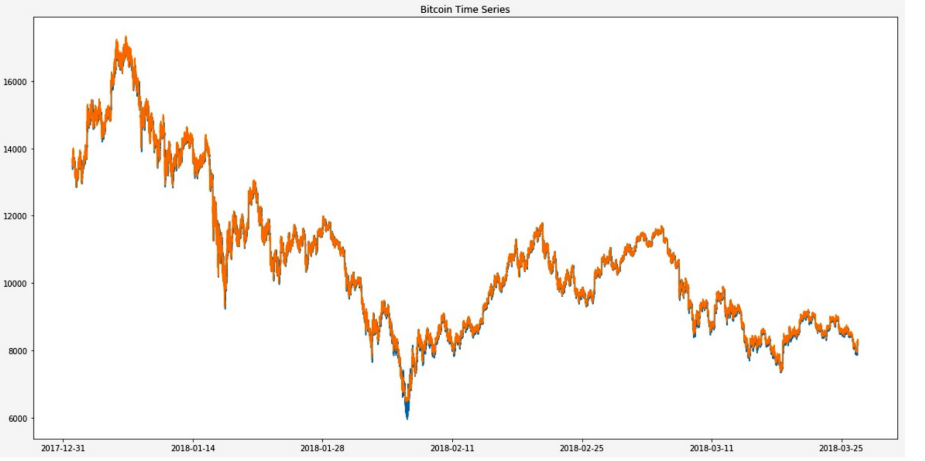


Figura 8. Serie temporal de precios de Bitcoin reales (azul) y precios predichos (naranja) por la red.

## CONCLUSIONES Y RECOMENDACIONES

Las conclusiones para este trabajo se pueden segmentar en dos vertientes, las relacionadas al análisis de sentimiento y las relacionadas a la red neuronal en sí. En esta sección se discutirá ambos conjuntos de conclusiones, detallando puntos fuertes y débiles de la metodología y oportunidades de mejora.

En cuanto a las expectativas y percepción de las personas, modelada por el análisis de sentimientos de *tweets,* se puede apreciar que su inclusión en el modelo LSTM para la predicción del valor del *Bitcoin* no fue muy influyente. Esto se ve explicado en tanto que el modelo no logra generalizar bien el comportamiento del sentimiento como serie temporal. En ese sentido, se reconoce la importancia de explorar otras formas de modelar y caracterizar la percepción de las personas. En primer lugar, indagar sobre otras formas de analizar el sentimiento de *tweets* no etiquetados, diferentes de *TextBlob*. En segundo lugar, verificar el cambio en el funcionamiento del LSTM al considerar el análisis de sentimiento como una variable entera. En tercer lugar, incluir otras caracterizaciones encontradas en el tratamiento de lenguaje natural como puede ser *topic modeling.*

Relacionado a la red neuronal, la aplicación de una LSTM tiene buen sustento teórico. El *gap* que existe en la literatura relacionado a su aplicación para predecir precios de *Bitcoin* motivó el estudio de esta técnica. Sin embargo, como se vio en la sección anterior, los resultados no fueron favorables. Así como se comentó anteriormente, la arquitectura actual no se ajusta bien a la complejidad del problema. Utilizar una capa recurrente en conjunto con una de *dropout* y una *fully connected* era una estructura muy sencilla para el problema que se estaba trabajando. En ese sentido, utilizar más capas intermedias y refinar la probabilidad de *dropout* (así como la cantidad de neuronas por capa) podría llevar a una mejora significativa en los resultados. El MAPE de 34.92% no es aceptable por dos razones: es un porcentaje de error respecto a la data muy alto y al analizar los volúmenes transados en *Bitcoin*, este 34% representa una cantidad de dinero significativa.

## REFERENCIAS

[1] J.M. Macedo. “Are We in a cryptocurrency bubble? A comparison with the 2000 dotcom bubble”. 2017.

[2] K. Leinz. ”Who Is Buying Bitcoin? This Charts Reveals the Answer-Money”. Time, 24 Ene. 2018.

[3] Greenberg. “A CRYPTO CURRENCY-Money you can’t trace”. Forbes, 40. 2011.

[4] I. Allison. “If Banks Want Benefits of Blockchains. They Must Go Permissionless”. 2015.

[5] S. Nakamoto. “Bitcoin: A Peer-to-Peer Electronic Cash System”. Cryptography Mailing list at https://metzdowd.com. 2009

[6] H. Jang and J. Lee. "An Empirical Study on Modeling and Prediction of Bitcoin Prices With Bayesian Neural Networks Based on Blockchain Information". In IEEE Access. Vol. 6, pp. 5427-5437. 2018.

[7] T. Ray Li, A.S. Chamrajnagar, X.R. Fong, N.R. Rizik and F. Fu. "Sentiment-Based Prediction of Alternative Cryptocurrency Price Fluctuations Using Gradient Boosting Tree Model". 2018.

[8] Cuadro de CoinMarketCap: https://coinmarketcap.com/currencies/zclassic/, (consultado el 26 de Junio, 2018).

[9] S. Lahmiri. "Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression". Applied Mathematics and Computation. Vol. 320, pp. 444-451. 2018.

[10] Q. Qin, Q.G. Wang, J. Li and S.S. Ge. "Linear and nonlinear trading models with gradient boosted random forests and application to Singapore stock market". Journal of Intelligent Learning Systems and Applications. Vol. 5 Nº 1, pp. 1-10. 2013.

[11] F. Villada, D.R. Cadavid, J.D. Molina. "Pronóstico del precio de la energía eléctrica usando redes neuronales artificiales". Revista facultad de ingeniería. Vol. 44, pp. 111-118. 2014.

[12] C.Y. Yeh, C.W. Huang and S.J. Lee. "A multiple-kernel support vector regression approach for stock market price forecasting". Expert Systems with Applications. Vol. 38 Nº 3, pp. 2177-2186. 2011.

[13] B.M. Henrique, V.A. Sobreiro and H. Kimura. "Stock Price Prediction Using Support Vector Regression on Daily and Up to the Minute Prices". The Journal of Finance and Data Science. 2018.

[14] A. Meucci. "'P' Versus 'Q': Differences and Commonalities between the Two Areas of Quantitative Finance". 2011.

[15] D. Garcia and F. Schweitzer. "Social signals and algorithmic trading of Bitcoin". Royal Society open science. 2015.

[16] Y.B. Kim, J.G. Kim, W. Kim, J.H. Im, T.H. Kim, S.J. Kang and C.H. Kim. "Predicting fluctuations in cryptocurrency transactions based on user comments and replies". 2016.

[17] R.C. Phillips and D. Gorse. "Predicting cryptocurrency price bubbles using social media data and epidemic modelling". In Computational Intelligence (SSCI), 2017 IEEE Symposium Series on 2017 Nov 27 (pp. 1-7). IEEE. 2017.

[18] S. Colianni, S. Rosales and M. Signorotti. "Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis". CS229 Project. 2015.

[19] E. Stenqvist and J. Lönnö. "Predicting Bitcoin price fluctuation with Twitter sentiment analysis". Dissertation. 2017.

[20] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". Neural Computation. Vol. 9 Nº 8, pp. 1735-178. 1997

[21] S. Hochreiter. "Untersuchungen zu dynamischen neuronalen Netzen". Master's thesis. Institut fur Informatik, Technische Universitat. Munchen. 1991.

[22] Y. Bengio, P. Simard and P. Frasconi. "Learning long-term dependencies with gradient descent is difficult". IEEE Transactions on Neural Networks. Vol. 5 Issue 2, pp. 157-166. 1994