MASTER UNIVERSITARIO EN CIENCIA DE DATOS

Predicción de Bitcoin con NLP y Twitter

CUNEF 2021-2022

Realizado por: Hugo Pasqual del Pobil

Dirigido por: Alejandro Baldominos y Leonardo Hansa

Madrid, 7 de junio de 2022

Resumen

El proyecto que nos ocupa consiste en predecir e intentar explicar los retornos del Bitcoin utilizando Procesamiento de Lenguaje Natural aplicado a *tweets* de Twitter. La motivación nace del auge de inversión por los criptoactivos y la influencia de las redes sociales sobre los mismos. La hipótesis planteada cosiste en demostrar y cuantificar la relación entre el activo y la red social. Mediante la limpieza y muestreo de los datos se selecciona un conjunto reducido de *tweets* de los cuales se calcula el sentimiento y se pondera la misma métrica. Seguidamente se procesan los datos para ser introducidos en dos modelos de red neuronal con distintas características. Cada uno de los modelos y su predicción son comparados y analizados para determinar la mejor manera de solucionar un problema de clasificación. Además de la predicción de los retornos se hace un estudio de los tópicos incluidos en los *tweets* y un análisis de la serie temporal del bitcoin.

Abstract

The project at hand consists of predicting and trying to explain Bitcoin's returns using Natural Language Processing applied to tweets. The motivation stems from the investment boom in crypto assets and the influence of social networks on them. The proposed hypothesis consists in demonstrating and quantifying the relationship between the asset and the social network. By cleaning and sampling the data, a reduced set of tweets is selected from which the sentiment is calculated, and the same metric is weighted. The data is then processed to be introduced into two neural network models with different characteristics. Each of the models and their prediction are compared and analysed to determine the best way to solve a classification problem. In addition to the prediction of returns, a study of the topics included in the tweets and an analysis of the time series of bitcoin is made.

Acrónimos

BTC – USD = Precio del bitcoin en dólares americanos (USD)

BTM = Modelo de Bi-Término para análisis de tópicos

ML = Machine Learning

NLP = Natural Language Processing

NN = Neural Network (Red Neuronal)

RNN = Recurrent Neural Network (Red Neuronal Recurrente)

EE. UU. = Estados Unidos de América

LSTM = *Long-Short Term Memory* (Red Neuronal)

CNN = Red Neuronal Convolucional

ARIMA = *Auto-Regressive Integrated Moving Average* (Modelo)

Repositorio GIT

- Adjunto al informe y resumen ejecutivo se incluye una carpeta comprimida con los cuadernos en Python utilizados en el proyecto. También se puede encontrar el repositorio en: https://github.com/hugopobil/tfm_hugopobil
- La carpeta de *code* incluye una subcarpeta de bitcoin donde se incluyen todo el código de referencia a este proyecto. Se incluye un *script* de soporte para el código.
- La carpeta de *data* con los datos utilizados, siguiendo una metodología de guardado en local y lectura de los datos al inicio de cada cuaderno.
- La carpeta de *models* contiene los cuadernos con los modelos de redes utilizados para la predicción del bitcoin y los modelos Bi-Término utilizados para el análisis de tópicos.
- Las carpetas de *functions, scr y plots_report* contienen las funciones definidas para la automatización de los procesos de análisis y graficación.

Tabla de contenido

1. Introducción

La idea general de este proyecto es aplicar un modelo de red neuronal para predecir el retorno del bitcoin utilizando datos de Twitter, para poder entender los movimientos y qué influencia puede tener la red social en la criptomoneda. *"Bitcoin es la moneda mas tuiteada por influencias, acumulando el 48% de las menciones en 2018 y el 50% en 2020."* (Ales Kovalevish, 2020). La hipótesis del proyecto se basa en la idea de que el movimiento del precio del bitcoin es influenciado por comentarios de usuarios en la red social de Twitter. Aplicando un modelo de red neuronal entrenamos un algoritmo que aprende la relación entre los dos datos y es capaz de hacer una aproximación de predicción introduciendo nuevos datos al modelo. El modelo se carga con datos de texto de Twitter y devuelve una clasificación positiva o negativa sobre el movimiento del precio del bitcoin.

La motivación de este proyecto se basa en los cambios que la industria financiera ha experimentado durante la última década con la introducción de nuevas monedas digitales que influencian los movimientos de la economía y los perfiles de los inversores. Existe en los mercados un creciente interés por las divisas digitales como el bitcoin, ethereum o ripple, entre todas las existentes. Este auge por invertir en este tipo de activos se ha traducido en una volatilidad anormal y por consecuente retornos anormales. Se han dado casos de retornos multimillonarios gracias a las rápidas crecidas del valor de estas monedas. Mediante el entrenamiento de un modelo de red, se aprovecha este rápido crecimiento y volatilidad para predecir y entender los retornos de las criptomonedas.

La capacidad de predecir el retorno de un activo usando las redes sociales se basa en la tendencia que tienen los inversores de criptomonedas de publicar su opinión sobre lo que ocurre en el mercado. *"La cuestión ya no se basa en si el sentimiento de los inversores influencia a los valores, sino como medir el sentimiento y cuantificar sus efectos."* (Baker and Wurgler, 2018).

Existen cuentas con millones de seguidores que se dejan aconsejar por los 'gurús' de las criptomonedas. La opinión de estas cuentas y las publicaciones de estas tienen influencia en los mercados, pues la demanda se asume que es impulsiva y emocional.

El precio de las monedas se determina por la oferta y la demanda de estas, en otras palabras, cuanto más positivo sea el sentimiento sobre una moneda en concreto, la teoría se traduce en que mayor será la crecida de su valor.

Este proyecto se centra únicamente en el bitcoin y la red social de Twitter. Siendo la moneda más popular y Twitter la red social por excelencia para publicar opiniones respecto a un tema en tendencia. Utilizando ambas fuentes de información permite generalizar los resultados.

Desde el inicio de bitcoin en 2009, en su undécimo año de existencia, el dinero digital o virtual que toma la forma de tokens se ha establecido como una moneda viable y una forma de inversión, el impacto económico de las criptomonedas es evidente en varias áreas de comunidades nacionales y globales. Actualmente, se estima que un 15% de la población mundial tiene activos digitales en sus carteras personales, elevando el valor total de estos activos a $2.2 trillones en 2022. (Reuters et al. 2022).

Desde enero de 2020, existen más de dos mil criptomonedas y casi 36,5 millones de personas que viven en los EE. UU. poseen algún tipo de criptomoneda. Aunque la criptomoneda en su conjunto no ha impactado a secciones más grandes de la economía como el mercado de valores, en 2017 se invirtieron cientos de miles de millones de dólares en criptomonedas, lo que la estableció aún más como una acción viable para invertir. De hecho, los expertos consideran que la criptomoneda como "oro digital" porque al igual que los metales preciosos, conserva su valor sin riesgo de depreciación.

Desde 2014 su valor ha aumentado de $400 a $40.000 con un máximo de $60.000 durante 2021, donde el precio muestra su mayor volatilidad. Como nuevo activo digital, bitcoin atrae a los jóvenes más que cualquier otro activo. Esta 'nueva' generación de inversores tiende a ser más activa en las redes sociales, transmitiendo sus pensamientos, emociones y opiniones a las plataformas de redes sociales. Este proyecto busca probar si se puede predecir bitcoin utilizando únicamente los comentarios y opiniones de las redes sociales.

En el pasado han existido varios casos de activos y su influencia en el precio por plataformas de redes sociales como Reddit, este es el caso de *GameStop*. *GameStop* es un minorista de videojuegos. Durante 2021, uno de los mayores *headge funds* apostó a la baja con las acciones cotizadas en bolsa debido al descenso de ventas físicas. Los pronósticos a la baja del precio cotizado se vieron afectados cuando un foro de Reddit llamado *WallStreetBets*

consiguió subir el precio del activo un 8.000% a través de miles de pequeños inversores. El incremento de volatilidad, traducido en un incremento del precio resultó en perdidas millonarias en los fondos de capital riesgo que apostaban contra la compañía. Esta oportunidad de inversión se podría haber identificado mediante el uso de técnicas de NLP y análisis de tópicos.

Sin embargo, Reddit es considerada como una plataforma de nicho, mientras que el contenido de Twitter está más diversificado por edad y antecedentes de usuario. La popularidad de bitcoin y el alcance y la utilidad de Twitter son una buena combinación para probar la hipótesis y entrenar un modelo para clasificar los retornos de la moneda.

El proyecto se estructura comenzando por la limpieza y muestreo de los datos. En la primera parte se detallan los pasos que se siguen para transformar los datos de texto y seleccionar una muestra estratificada de los datos que se van a usar.

La segunda parte consiste en analizar la serie temporal del bitcoin y computar el sentimiento de los *tweets* para buscar relaciones entre las variables. Además, se incluye un análisis de tópicos para identificar los temas de los *tweets* y ayudarnos a entender la relación entre Twitter y bitcoin.

Se definen dos modelos de red neuronal para el entrenamiento y predicción del bitcoin usando los *tweets* como única fuente de información. Se analizan los resultados para poder concluir sobre las predicciones y mejoras en los procesos descritos anteriormente.

2. Estado del arte

Originalmente se propusieron intentos de predecir el precio de bitcoin utilizando celdas de memoria a largo plazo (LSTM), red neuronal convolucional (CNN) y modelos híbridos CNN-LSTM sin realizar un análisis de sentimiento (Li y Dai 2020; Livieris et al. 2021; Kwon et al. 2019).

Pant (2018) propuso un trabajo inicial para investigar el uso del análisis de sentimientos de los datos de Twitter para la predicción de precios; Galeshchuk et al. (2018) y Serafini et al. (2020). Para eliminar el problema del gradiente de fuga que se observa en las redes neuronales recurrentes (RNN), Pant (2018) propuso utilizar un predictor de RNN con variaciones de LSTM y *Gated Recurrent Unit* (GRU). Este trabajo dio como resultado una correlación moderada entre el aumento del sentimiento negativo y la consiguiente caída de los precios. Se realizó una comparación de los enfoques basados en modelos LSTM y ARIMA en Serafini et al. (2020), en el que se afirma que el modelo ARIMA se comporta mejor.

Más comparaciones con otros modelos y en diferentes criptomonedas se presentaron en Valencia et al. (2019) y Wołk (2019). Según Valencia et al. (2019), los datos de Twitter no son suficientes para predecir el precio de bitcoin por sí solos, pero pueden ayudar cuando se combinan con otros datos de mercado.

Los estudios realizados anteriormente intentan resolver un problema de regresión mediante la predicción del retorno con dos decimales de ajuste. El problema de intentar buscar una correlación precisa entre los datos de Twitter y el retorno son las métricas de resultados, las cuales devuelven *scores* muy bajos debido a la poca correlación. Intentar resolver un problema de clasificación para predecir cuando el bitcoin sube o baja debería obtener mejores métricas en los resultados.

3. Muestreo de los datos y limpieza

Este capítulo detalla los datos que se utilizan en el proyecto y el proceso de limpieza que se aplica a los datos de tipo texto. El proceso es muy importante, pues la calidad y resultado final son dependientes del correcto procesamiento de los datos.

3.1 Obtención de los datos

Tanto los datos financieros históricos de bitcoin como los *tweets* provienen de Kaggle, un conocido proveedor de datos para usos prácticos. La descarga de *tweets* utilizando la API de Twitter (tweepy) tiene un uso limitado de descargas. Teniendo en cuenta la cantidad de datos necesaria para este proyecto y el fin educativo, se utiliza la base de datos de Kaggle. Los datos se descargan y almacenan en la carpeta de datos del proyecto (data) para ser leídos por el cuaderno y proceder a limpiar y transformar el texto.

Tweets

El conjunto de datos de *tweets* incluye trece variables y 2.347.470 *tweets* únicos, desde el 10 de febrero de 2021 hasta el 2 de febrero de 2022. Los *tweets* se filtran y solo se incluyen aquellos con los hashtags “BTC”, “Bitcoin”, “btc”, “bitcoin” y “BITCOIN” en el cuerpo del texto del *tweet*. De esta manera se utilizan únicamente los *tweets* relevantes para la predicción.

Bitcoin

El *dataset* de bitcoin incluye el precio de cierre ajustado y la fecha, con inicio de los datos el 17 de septiembre de 2014 y finalización el 28 de enero de 2022. El modelo y el proyecto se centran únicamente en el año 2021, los datos previos al margen de estudio se utilizan para el análisis de la evolución desde la creación de la criptomoneda.

3.2 Muestreo Estratificado

El *dataset* de *tweets* contiene aproximadamente dos millones de valores únicos. La máquina de entrenamiento y procesado de los modelos tiene una capacidad limitada. Esta restricción nos obliga a reducir el tamaño de los datos utilizando un muestreo que se define a continuación.

El muestreo estratificado consiste en reducir los datos iniciales a una muestra reducida siguiendo un patrón para elegir unos datos que comparten unas mismas características. En este caso queremos incluir un mismo número de *tweets* por día, ordenados por fecha mediante la aplicación de una función de grupo y muestra de doscientos *tweets* individuales. Los datos se reducen al uno por ciento de su tamaño original (dos millones de *tweets*), con el tamaño del conjunto de datos reducido a veinte-tres mil *tweets*.

La muestra estratificada surge de la necesidad de incluir en la muestra datos para todos los días del año. El problema que nos ocupa de predicción consiste en clasificar el retorno del bitcoin para cada uno de los días que se consideran (2021). El método de muestra estratificada nos permite reducir la dimensión original de los datos para incluir datos para todos los días de la serie del bitcoin.

Si los *tweets* no están ordenados y existe una cantidad diferente de *tweets* para cada día, los datos son inconsistentes, pues el modelo recibe distinta cantidad de información para los distintos retornos y no es capaz de identificar una relación consistente. Para intentar una predicción del rendimiento de bitcoin, se agregan datos por día y se procesan en el modelo. La reducción del conjunto de datos consecuentemente mejora los tiempos de computación.

3.3 Limpieza de los datos

Los *tweets* pueden contener texto sin formato, menciones, *hashtags*, enlaces y signos de puntuación. Antes de continuar procesando los *tweets* para introducirlos en los modelos y calcular el sentimiento de estos es necesario limpiarlos y transformarlos en una versión más simple y fáciles de procesar. Si el texto de los *tweets* incluyera simbología, números y otro carácter que no forma parte de las palabras, podría conllevar a errores y por lo tanto una predicción errónea. A continuación, se describe el proceso seguido en este proyecto para la limpieza de los *tweets*.

El primer paso es crear una nueva variable donde se incluyan todos los *tweets* del muestreo, creando una copia del dataset y procediendo a la limpieza individual de cada *tweet*, mediante un proceso iterativo que elimina lo siguiente:

- *Hashtags*, menciones, *links y* puntuación
- Filtrado de caracteres no alpha-numéricos
- Números y emojis

El objetivo es terminar con un texto que únicamente incluya las palabras escritas por el autor haciendo referencia al bitcoin y eliminar lo que no es relevante. Esto se debe a que los *tweets* continúan un procesado de tokenización en el cual se distinguen las palabras y se elimina las preposiciones que no aportan información relevante. La tokenización no sabe distinguir entre los distintos tipos de palabras, esto se debe a que este proceso de tokenizar un texto consiste en la división o separación en fragmentos una cadena de texto, esta división se basa en los espacios entre las distintas palabras.

Algunos modelos de procesamiento de lenguaje natural permiten identificar los emojis y los signos de puntuación para identificar el sentimiento. Al fin y al cabo, los emojis tienen una finalidad de representar emociones a través de iconos. Este proyecto se centra en procesar los *tweets* únicamente con modelos que admiten texto. Más adelante se detallan los modelos utilizados y el tipo de dato que admiten.

	tweets_clean	tweets
0	amazing monopoly crypto cryptocurrency cryptoc...	Blue Ridge Bank shares halted by NYSE after #b...
1	blockchain by cryptocurrency bitcoin crypto bl...	😎 Today, that's this #Thursday, we will do a "...
2	analyst says bitcoin likely reach in q bitcoin...	Guys evening, I have read this article about B...
3	penn asia is developing a digital asset called...	$BTC A big chance in a billion! Price: \487264...
4	dwtsiahtjdlpdapbnakyipfmwwmjf australia s cent...	This network is secured by 9 508 nodes as of t...
5	flm flm bullish allready profit hold on only m...	💹 Trade #Crypto on #Binance \n\n📌 Enjoy #Cashb...
6	stan drunkenmiller shared his views on bitcoin...	<'fire' & 'man'>\n#Bitcoin #Crypto #...
7	btc buying pressure alert price trading around...	🔄 Prices update in $EUR(1hour)$: \n\nBTC - ...
8	watch csgo live streaming rerun team vitality ...	#BTC #Bitcoin #Ethereum #ETH #Crypto #cryptotr...
9	keep knocking btc bitcoin	.@Tesla's #bitcoin investment is revolutionary...

Figura 1.1 Comparación limpieza de *tweets*

La Figura 1.1 muestra la diferencia entre los *tweets* antes de ser procesados (columna derecha) y el resultando del proceso de limpieza (columna izquierda) definido anteriormente. Se observa la desaparición de los emoticonos, signos de puntuación y las mayúsculas, además de la simbología de menciones típica de Twitter.

4. Procesamiento de Lenguaje Natural

4.1 Análisis de Sentimiento

Una vez que los datos se han limpiado y están listos para ser procesados, el siguiente paso es crear nuevas variables que nos ayuden a comprender la ingesta de los datos a los modelos. El primer paso es procesar *tweets* e identificar su sentimiento mediante la métrica de polaridad. Los *tweets* pueden ser tener sentimiento positivo, neutro o negativo, según las palabras clave y el contexto de la expresión utilizada en el texto.

Después del preprocesamiento, se utiliza VADER (Hutto y Gilbert 2015) para asignar puntajes de sentimiento a los *tweets*. Utilizan enfoques similares Abrahán et al. (2018) y Kraaijeveld y De Smedt (2020). VADER califica cada *tweet* con una puntuación de polaridad negativa, positiva, neutral y compuesta. El puntaje compuesto es una suma de los puntajes de sentimiento individuales, ajustado de acuerdo con un conjunto de reglas y normalizado para caer dentro del rango [−1,+1]. Sin embargo, para los fines de este estudio, solo se incluyen puntajes de polaridad positiva, neutral y negativa (se elimina la compuesta) en los conjuntos de datos de capacitación y evaluación.

VADER ofrece ventajas que incluyen las siguientes: es de código abierto y libre; está validado por humanos y ajustado para el contenido de Twitter y también se ha demostrado que tiene un desempeño competitivo con anotadores humanos y ha superado varios puntos de referencia, especialmente en contenido de redes sociales (Hutto y Gilbert 2015).

El texto y la expresión se pueden asociar con emociones. VADER intenta con el uso de algoritmos computacionales procesar texto para decodificar y cuantificar la emoción contenida en los datos. El análisis de sentimiento tiene dos enfoques principales: léxico y aprendizaje automático. El objetivo léxico consiste en mapear palabras a sentimientos y construir un grupo de palabras con su respectivo sentimiento.

El sentimiento toma una forma categórica (negativo, neutro y positivo), pero también puede ser numérico y expresarse como un rango porcentual entre cero y uno, para el puntuaje de intensidad, que se puede traducir a la forma categórica para una mejor comprensión.

Métrica de polaridad

Mediante el uso del analizador de sentimientos (*polarity_scores*) aplicado a todos los *tweets* previamente limpiados y procesados, obtenemos la puntuación de polaridad. Este puntaje, como se mencionó anteriormente, es una métrica numérica expresada como un porcentaje de la intensidad del sentimiento.

Métrica ponderada

Al obtener la métrica de sentimiento, el analizador no puede distinguir que *tweets* son más relevantes. En de las redes sociales los usuarios interactúan unos con otros, lo cual conlleva a que algunos usuarios tengan más relevancia y supuesta influencia sobre el bitcoin con sus publicaciones que otros usuarios. Para poder ponderar el sentimiento de los *tweets*, utilizamos una fórmula simple que pondera las métricas utilizando el número de seguidores.

$$Métrica\ ponderada = intensidad\ de\ sentimiento * \frac{(número\ de\ seguidores + 1)}{1000}$$

Aplicando la anterior fórmula cada una de las métricas se multiplica por el número de seguidores y se divide entre mil. Hemos considerado la división por mil como base de relevancia para el número de seguidores, en la descarga de datos se incluye para cada *tweet* una variable con la información. Cualquier *tweet* publicado por un usuario con menos de mil seguidores, verá su métrica de sentimiento reducida porcentualmente. De esta manera, los usuarios con poca relevancia reciben menos puntuaje. Un usuario con veinte mil seguidores multiplica por veinte la métrica de sentimiento.

Es importante tener en cuenta la relevancia de los *tweets* para poder entender mejor su influencia en la criptomoneda. Si se consideran todos los *tweets* por igual, no se estaría representando una predicción de la realidad. Como se ha mencionado anteriormente, aquellos usuarios con más seguidores tienen una capacidad de influenciar más los cambios de precio de las monedas a través de sus opiniones, un claro ejemplo es Elon Musk, el 14 de junio de 2021 se publica en El Mundo la noticia de que un *tweet* dispara el bitcoin hasta los $40.000. Consecuentemente, se aplica la formula anterior para influenciar y ponderar las métricas de sentimiento en los usuarios con mayor número de seguidores.

Mediante la variación de la variable de base de referencia (1000 seguidores) se puede influenciar la métrica se sentimiento para probar distintos escenarios. Aumentando la base se influencia en ambos grupos, reduciendo la ponderación en los extremos de las publicaciones.

	text	sample_date	user_followers	compound	score
0	debunking bitcoin myths by crypto...	2021-02-05	301.0	0.0000	0.000000
1	blockchain by cryptocurrency ...	2021-02-05	301.0	0.0000	0.000000
2	bitcoin braces for as inverse head and shou...	2021-02-05	16546.0	0.4019	6.650239
3	bitcoin bitcoin btc btcusd	2021-02-05	4154.0	0.0000	0.000000
4	weekend read keen to learn about crypto ...	2021-02-05	16813.0	0.4939	8.304435

Figura 1.2 Extracto de dataset de *tweets* con métricas de sentimiento

Se observa en la Figura 1.2 que el *score* que mayor puntuaje recibe es aquel que mayor número de seguidores tiene, gracias a la fórmula aplicada anteriormente. Se destacan algunos *tweets* que no reciben *score*, esto se debe al algoritmo de análisis de sentimiento, que no detecta intensidad en las palabras. Comprobando el número de tweets que tienen score cero nos damos cuenta de que ocho mil *tweets* aproximadamente no han recibido ningún score. Esta cifra representa un 35% de los datos que no aportan información para el testeo de la hipótesis. La manera de solventar esta pérdida de datos es mediante el aumento de la muestra inicial. En la sección de muestreo y limpieza se define un muestreo estratificado de dos cientos *tweets* por día. Aumentando esta cifra, el porcentaje de pérdida de información se vería reducido indirectamente con un mayor número de *tweets* para procesar.

5. Análisis del Bitcoin

5.1 Bitcoin

Bitcoin, un sistema de moneda electrónica descentralizado, representa un cambio radical en los sistemas financieros después de su creación en 2008 por Satoshi Nakamoto. bitcoin representa una innovación de IT basada en el avance de las redes *peer-to-peer* y los protocolos criptográficos. Debido a sus propiedades, bitcoin no es administrado por ningún gobierno o banco. Como cualquier otra moneda, una peculiaridad de bitcoin es facilitar las transacciones de servicios y bienes, atrayendo a un gran número de usuarios y mucha atención de los medios.

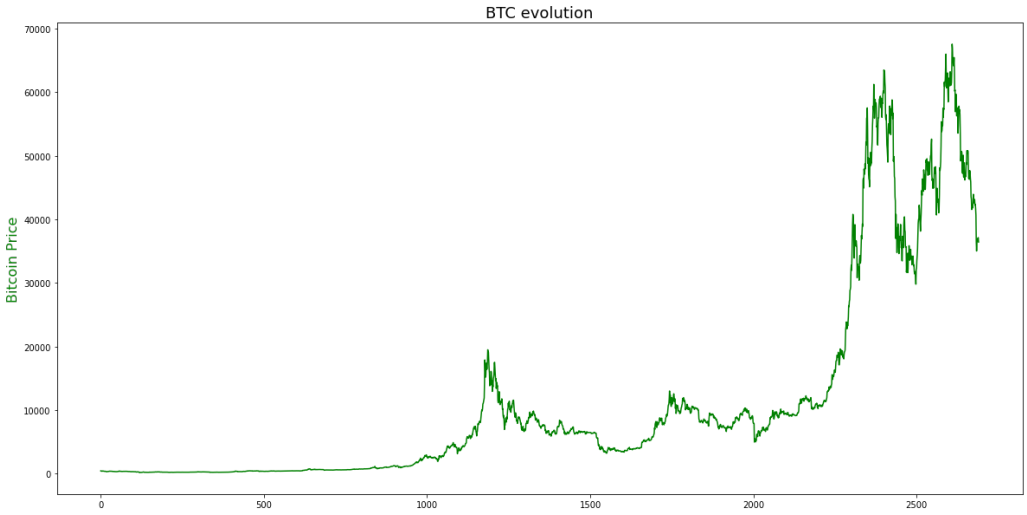


Figura 2.1 Evolución Precio del Bitcoin (2014 a 2022)

5.2 Agrupación de los datos

En esta etapa del proceso, nos encontramos con dos bloques principales de datos, bitcoin y *tweets*. Estos conjuntos de datos tienen una frecuencia de tiempo distinta. Esto significa que los *tweets* se pueden escribir en cualquier momento del día y la frecuencia del precio de cierre ajustado de bitcoin es por días. Para poder trabajar con ambos conjuntos de datos y extraer información de valor, convertimos el conjunto de datos de *tweets* en datos que se puedan contrastar con la frecuencia diaria del bitcoin.

La agrupación de los *tweets* utiliza la fecha de publicación aplicando la suma de los *scores*. Al tener *scores* negativas y positivas, la suma devolverá el sentimiento acumulado por hora. Debido a la volatilidad de la criptomoneda, dejamos en frecuencia horaria la agregación del sentimiento.

5.3 Correlación de los datos

Con ambos conjuntos de datos alineados para graficar, realizamos correlaciones estadísticas básicas para ver si existe una relación entre los movimientos en el precio de bitcoin y las puntuaciones de sentimiento de los *tweets* agrupados por hora para distintos momentos en el tiempo.

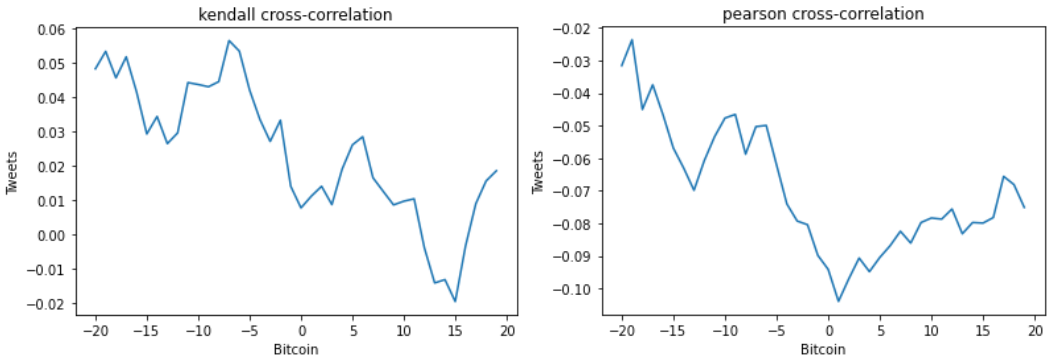


Figura 2.2 y 2.3 Gráficos de correlación

Las correlaciones de Kendall y Pearson entre bitcoin y el sentimiento de *tweet* muestran cierta correlación, aunque muy pequeña, entre el-10% y 6%. La correlación de rango de Kendall se usa para probar las similitudes en el orden de los datos cuando se clasifican por cantidades. Otros tipos de coeficientes de correlación usan las observaciones como base de la correlación, el coeficiente de correlación de Kendall usa pares de observaciones y determina la fuerza de la asociación según el patrón de concordancia y discordancia entre los pares. Por otro lado, la correlación de Pearson es una medida de correlación lineal entre dos conjuntos de datos. Es el cociente entre la covarianza de dos variables y el producto de sus desviaciones estándar; por lo tanto, es esencialmente una medida normalizada de la covarianza, de modo que el resultado siempre tiene un valor entre −1 y 1. Al igual que con la propia covarianza, la medida

solo puede reflejar una correlación lineal de variables e ignora muchos otros tipos de relaciones o correlaciones.

La correlación de rangos de Kendall (no paramétrica) es una alternativa a la correlación de Pearson (paramétrica). Estos resultados nos adelantan que intentar modelar una regresión tendría un resultado pobre y de baja significancia. Los modelos de clasificación son más simples, pero primero se deberá definir una variable categórica para poder introducirla al modelo como objetivo de predicción que se detalla la sección de Metodología.

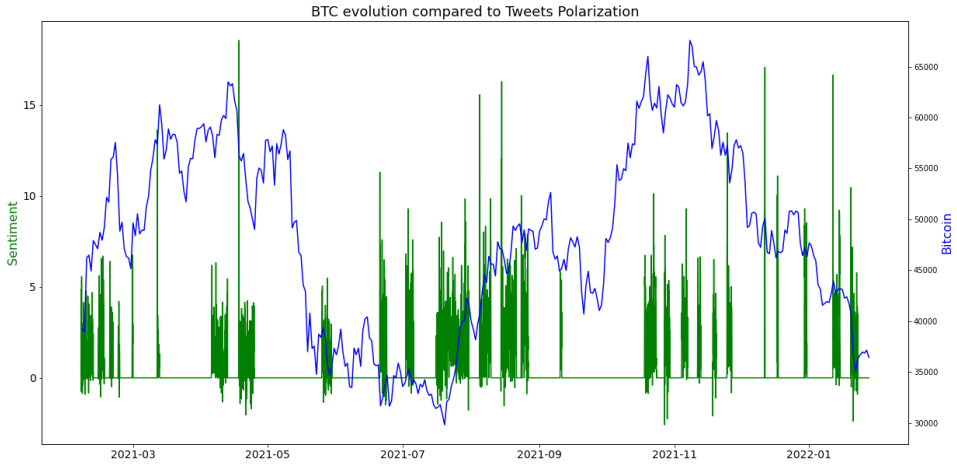


Figura 2.4 Bitcoin y Sentimiento de *Tweets*

La figura 2.4 muestra la comparación entre la evolución del precio de bitcoin y las puntuaciones compuestas antes de transformarse utilizando la relevancia del usuario. A primera vista, la intensidad del sentimiento tiene mayor frecuencia cuando el bitcoin experimenta su período alcista más largo y estable.

Después de dar a cada compuesto el score de su relevancia ponderada, podemos observar en la Figura 2.5 que durante el período de junio a noviembre de 2021, el precio de bitcoin experimenta un aumento de 3.000 USD a 60.000 USD. Un período en el que las puntuaciones de sentimiento tienen su pico más alto por ponderación. Nuestro modelo intentará encontrar la relación entre el sentimiento del *tweet* y el rendimiento de la criptomoneda. De un análisis inicial podemos ver una relación, el sentimiento recibe un puntuaje mayor en julio de 2021, momento en el cual el bitcoin inicia una crecida hasta noviembre de 2021.

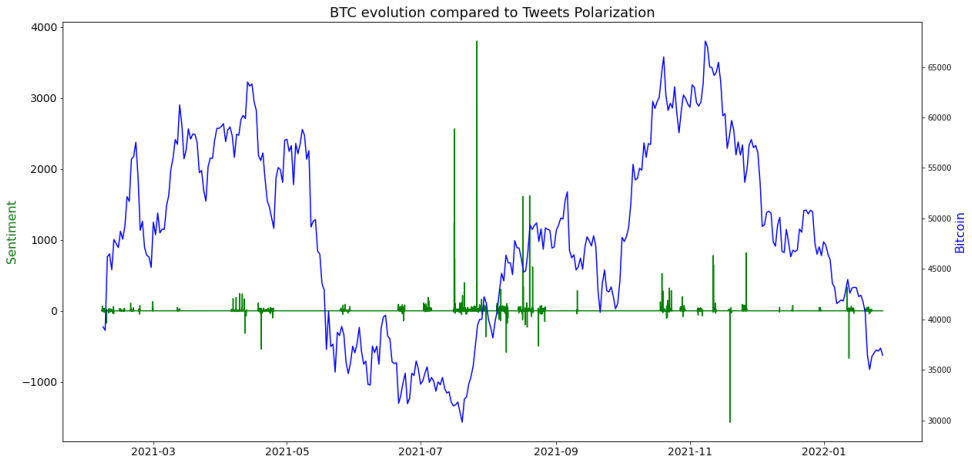


Figura 2.5 Bitcoin y Sentimiento Ponderado por Seguidores

La conclusión de esta primera fase de visualización de relación entre los conjuntos de datos es que existe poca relación. Predecir el retorno del bitcoin con un problema de regresión utilizando el score de análisis de sentimiento plantea un problema de predicción mucho más complejo. El siguiente paso será procesar los datos de texto con técnicas de NLP para introducirlos en un modelo de clasificación.

5.4 Correlación entre variables

Para poder calcular la correlación exacta entre las variables de bitcoin y la polaridad primero es necesario escalar los datos. Una vez procesados se visualiza la siguiente relación y se computa la correlación y se obtiene un resultado del -16.01%, utilizando la correlación de Pearson.

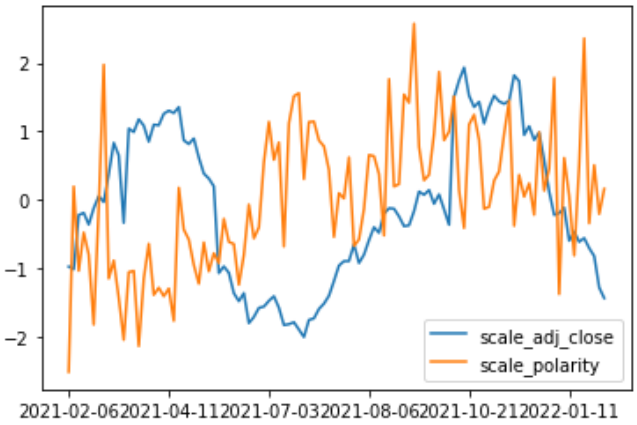


Figura 2.6 Gráfico representado ambas variables escaladas

5.5 Análisis de Estacionalidad

Para aplicar el análisis de estacionalidad del bitcoin convertimos los datos a frecuencia mensual. De esta manera podemos analizar los datos de tendencia y sus componentes estacionales para todo el conjunto de datos desde 2014 hasta 2022. Aunque estamos utilizando únicamente los datos de 2021 a 2022, es importante tener una idea de la tendencia y la estacionalidad de los datos durante todos los años. La descomposición necesita datos de distintos años para poder identificar las características.

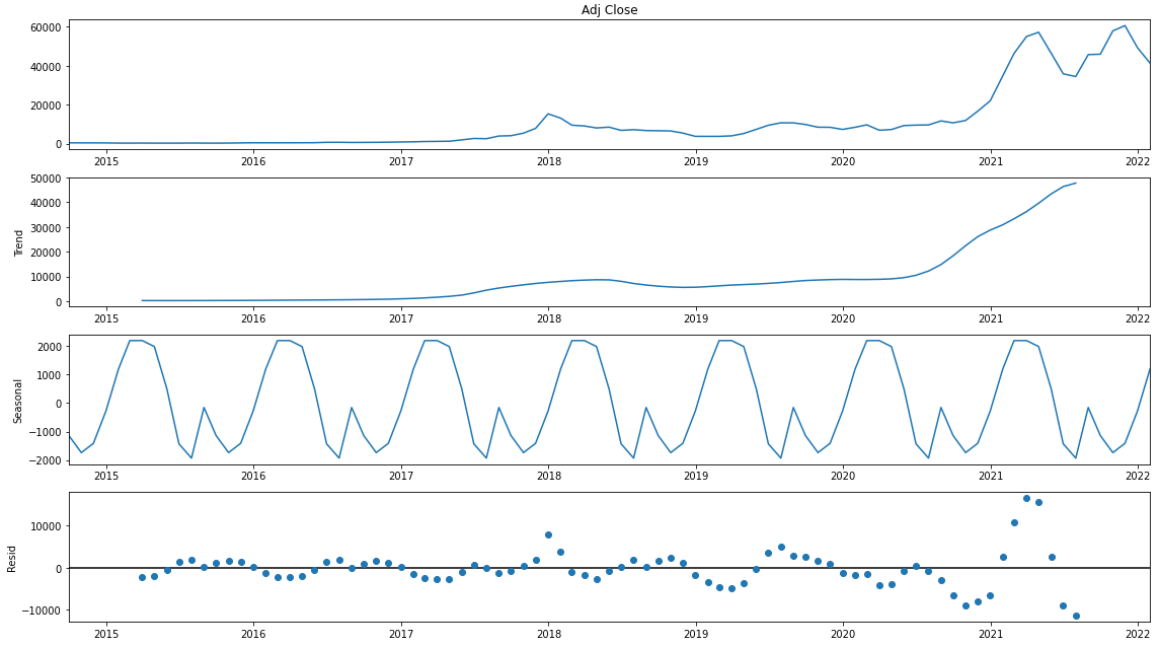


Figura 2.7 Análisis de Estacionalidad de Bitcoin (2014 a 2022)

La Figura 2.7 muestra una tendencia alcista desde sus inicios al precio actual. Esto se debe a la popularidad que ha ganado y el alto nivel de transacciones que ha experimentado que ha llevado a la subida del nivel de precio y por lo tanto la comparación de 2014 a 2022 muestra esta tendencia que continúa creciendo.

El componente estacional anual nos indica que el precio del bitcoin tiende a crecer durante los primeros meses del año con un repunte en julio y agosto para los últimos meses del año volver a subir. Los residuos muestran mayor volatilidad a partir de 2021, coincidiendo con un mayor cambio en el precio del bitcoin.

6. Análisis de Tópicos

El análisis de tópicos consiste en el procesado de texto mediante técnicas de NLP para agrupar los distintos *tweets* en tópicos que contienen palabras similares y poder visualizar en una línea temporal cuando el tópico gana relevancia. Los tópicos nos indican los distintos temas que se incluyen en los datos de Twitter. Gracias a la visualización en línea temporal, los tópicos nos ayudan a entender cuando el tema al que hacen referencia gana relevancia en un punto en el tiempo. Cada uno de los tópicos modelados pueden tener más o menos relación con la variable del bitcoin. Los tópicos tienen la finalidad de identificar temas, no explicar el movimiento de una variable independiente.

Para poder obtener los tópicos de nuestros datos partimos de los datos de texto previamente limpiados y procesados en los primeros pasos del proyecto. Para analizar los tópicos únicamente necesitamos el cuerpo del texto. Las métricas de sentimiento no son relevantes. Comenzamos aplicando la tokenización de los *tweets* y creamos una nueva variable llamada *tokens* que contiene todas las palabras relevantes del texto. Durante la tokenización se eliminan las preposiciones y otras palabras contenidas en una lista en inglés descargada de la librería de NLP par Python NLTK. Seguidamente se eliminan los tokens "bitcoin" en diversas formas ("BTC", "btc", etc.) dado que no aportan información nueva sobre las tendencias en la criptomoneda. Esto se debe a que el modelo de tópicos utiliza el contaje de las palabras más frecuentes en los documentos (*tweets)*.

El siguiente paso es crear un diccionario de tokens utilizando una de las funciones definidas en el script de apoyo y crear una lista de tuplas con la palabra y la frecuencia de las 10 palabras más comunes. Las palabras más frecuentes en los *tweets* tienen mucha relación con el bitcoin, lo cual es lo esperado. Se genera un *bag of words*, utilizando el diccionario de palabras. El *bag of words* consiste en una matriz que contiene todas las palabras únicas y la frecuencia de aparición en cada uno de los *tweets*. De esta manera se crea una matriz numérica que puede ser introducida en el modelo para obtener los tópicos.

Modelo Bi-Término

En la mayoría de los modelos de tópicos, los tópicos se representan como grupos de palabras correlacionadas con la correlación básicamente revelada por patrones de co-ocurrencia de palabras en los documentos. Por ejemplo, una vez que se observan las palabras "ipad" y "iphone" con frecuencia, uno puede decir que tienen sentidos cercanos y posiblemente pertenecen a un mismo tema, aunque no sepa el significado exacto de estas. Los modelos de tópicos convencionales explotan los patrones de co-ocurrencia de palabras para revelar la estructura semántica latente de un corpus de forma implícita al modelar la generación de palabras en cada documento.

Estos enfoques son sensibles a la brevedad de los documentos, ya que los patrones de coincidencia de palabras en un solo documento breve son escasos y no confiables. En cambio, si agregamos todos los patrones de co-ocurrencia de palabras en el corpus, sus frecuencias son más estables y revelan más claramente la correlación entre las palabras. Con esta idea, desarrollamos el modelo de tema bitérmino, que adopta una forma novedosa de revelar los componentes del tema latente en un corpus mediante el modelado directo de la generación de patrones de co-ocurrencia de palabras.

El modelo Bi-Término (BTM) se aplica para distintos números de tópicos. Variando este número el modelo agrega los datos y obtiene agrupaciones más específicas para mayor número de tópicos o más generales para el contrario. El título del tópico no es representativo, se utiliza como visualización de los distintos tópicos.

topic_id	title	tweets_count	tweets_theta_count	topic_no
0	stake,shock,cap,wmz,ranoq,mechanism,scammer,un...	1730	3094.059988	3
1	shock,unbelievable,highs,jque,wmz,mark,xr,stak...	8063	6190.948887	2
2	gi,shock,heard,stake,unbelievable,ren,hex,libr...	7820	6189.453191	1
3	gfnrbaz,bayc,unbelievable,pullback,qcpgvl,dxy,...	1088	2329.208876	4
4	shock,bitcoinprice,proof,mechanism,ueamxtibgk,...	214	1111.329057	5

Figura 3.1 Tabla de tópicos

Definición de cada uno de los tópicos:

- **Tópico 1**: Tópico relacionado con cuentas anónimas de criptomonedas
- **Tópico 2**: Tópico general sobre el bitcoin
- **Tópico 3**: Tópico con menciones a política y filtraciones
- **Tópico 4**: Tópico relacionado con cuentas de famosos
- **Tópico 5**: Tópico general sobre criptomonedas

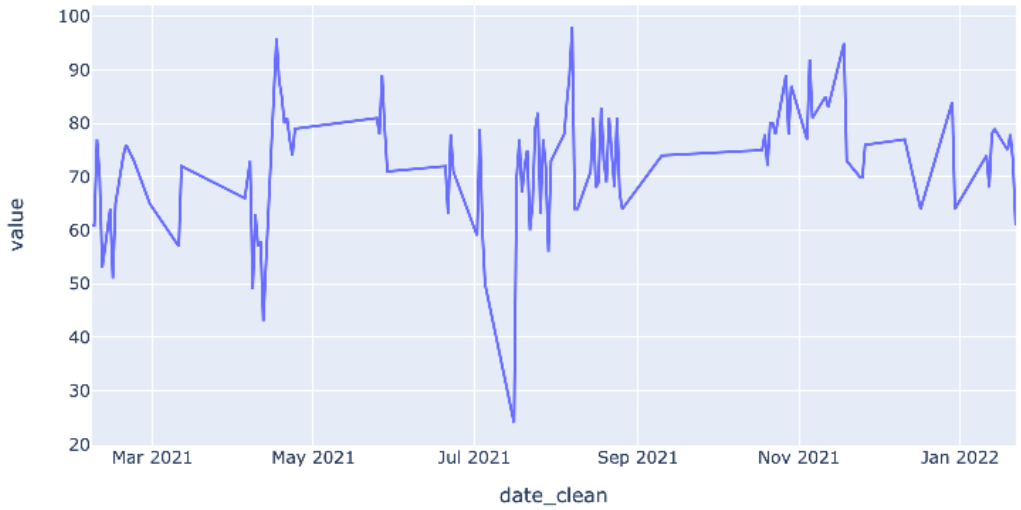


Figura 3.2 Gráfico temporal de tópico uno

Podemos observar en la Figura 8 como este tópico pierde relevancia durante el mes de julio de 2021. Esta bajada está relacionada con la estacionalidad analizada anteriormente, durante los meses de julio y agosto el bitcoin parece mostrar menos volatilidad, lo cual se puede traducir en un menor número de transacciones y movimientos. A partir de la información que aportan los tópicos se podría modelar un problema de regresión con el bitcoin.

Este proyecto intenta predecir un problema de clasificación; el uso de los tópicos únicamente nos sirve para ayudarnos a entender los movimientos de la criptomoneda y la influencia que Twitter puede tener sobre el precio. Al analizar los distintos tópicos se observa que en el tópico 4 existe un periodo de tiempo anormal, cuando el tópico tiene un valor de 120. Esto representa el número de *tweets* correspondiente a ese tópico. El tópico cuatro, el cual está relacionado con cuentas de más influencia genera ese periodo anormal por la publicación de estas respecto al bitcoin.

En el apéndice se encuentran las gráficas de los tópicos dos, tres y cinco. No se han incluido en el análisis principal debido a la poca capacidad explicativa hallada en los resultados.

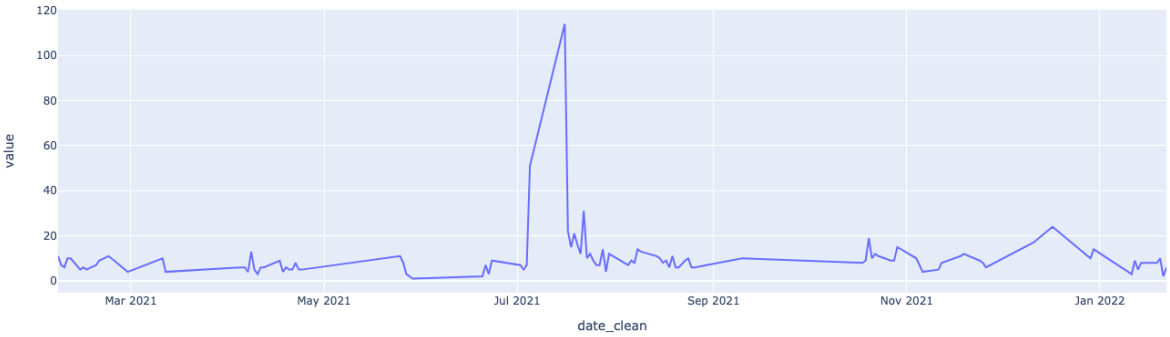


Figura 3.3 Gráfico temporal del tópico cuatro

	date_clean	cleaned_tweets	topic_prob
12679	2021-08-18	Supply On Exchange Wallets D Bitcoin EXCHANGES Coinbase Binance Bitfinex Huobi Kraken OKEX UIlUQzvqKj	0.359864
9131	2021-07-23	Footage drone performance taken rehearsal Tokyo Olympics Opening Ceremony This played actual ceremony Looks interesting Tokyo OlympicGames Japan Japanese w xJt JJPn	0.382020
15425	2021-10-23	GAS Trend reversal h timeframe Last Price Binance h Volume BTC BTC Binance GAS Nvcrq Gojj	0.413985
6354	2021-06-21	FIL BTC Volume increase detected Spike BTC h Vol BTC Get ALL alert NPcHHX r crypto bitcoin binance altcoins hodl defi trading freedom blockchain btc trig dnt tnb icn eng fuel storm lend bchsv xzc ethbull iwVyYODnWC	0.357386
18141	2021-11-26	You Cloud Mining using site IemY RVt w mining dogemining trxmining btcmining money earnmoney register bitcoin ethereum btc eth	0.463416

Figura 3.4 Tabla con muestra de *tweets* tópico cuatro

Se observa que la mayoría de los *tweets* tienen un contenido positivo respecto al bitcoin. Volviendo a la evolución del bitcoin durante el mismo periodo de tiempo se puede observar que el bitcoin crece durante meses hasta alcanzar sus máximos. De todos los tópicos el 4 es el que mejor representa el cambio de precio para 2021.

A la hora de utilizar esta herramienta, este tópico es interesante para mantener una monitorización y observar cuando aumente su número de *tweets* contabilizados para esperar una crecida del valor del bitcoin. Los tópicos no son una ciencia cierta, los datos del pasado pueden no aplicar al futuro. En concreto el tópico cuatro tiene muchas menciones a Hash, esto representa el número de transacciones capaces de ser encriptadas y aseguradas por Blockchain.

Anteriormente, se observan comentarios positivos y aumento de la mención sobre la encriptación y el elevado número de transacciones, pudiéndose así explicar el cambio de precio del bitcoin.

7. Metodología

Preparación de los datos

Una vez tenemos los *tweets* limpios y listos para ser introducidos en el modelo, definimos como variable objetivo el retorno del bitcoin, siendo esta variable una clasificación positiva o negativa (neutro en caso de mantenerse igual) según el comportamiento del precio en un periodo de 7 días.

Variando el periodo de 7 días de los retornos del bitcoin a un menor tiempo podemos influenciar el resultado de predicción. Se intuye que el *tweet* tiene más influencia cuanto menor sea el margen de tiempo del retorno calculado en días a partir de la fecha de publicación del *tweet*. Se computa semanalmente como punto de inicio de entrenamiento.

Definimos las variables del modelo:

- **X**: *tweets* limpios y tokenizados en secuencia
- **Y**: Clasificación del retorno del bitcoin a 7 días

El siguiente paso es distribuir las variables X e Y en sets de entrenamiento y validación. Lo hacemos mediante la función *train_test_split*, que redimensiona a un 20% los datos de validación, los cuales se usan para comprobar las predicciones y calidad del modelo. Las dimensiones son las siguientes:

- **Set de entrenamiento** = 16.452 *tweets*
- **Set de validación** = 4.113 *tweets*

Procedemos a tokenizar los *tweets* de entrenamiento y validación, este proceso separa por tokens la composición del *tweet*. Para ello, ajusta un máximo de 20.000 *'features'*. Esto se realiza por motivos de computación, y se asume que 20.000 es un número balanceado para no influir en los resultados de predicción. El siguiente paso convierte la tokenización en secuencia y reduce a cincuenta el número máximo de palabras por *tweet*. En el análisis de los *tweets* se pudo observar que la mayoría de los *tweets* están compuestos por menos de 50 palabras.

Modelo de red neuronal 1

El modelo de red 1 consiste en una red de siete capas convolucionales de una dimensión, estos modelos son muy usados para el reconocimiento de la actividad humana en los textos. Se aplica *MaxPooling* para la reducción de los datos y LSTM que ayuda a la red neuronal al aprendizaje de las secuencias en orden. En nuestro caso asumimos que el orden de las palabras en cada uno de los *tweets* tiene relevancia. Las personas tienden a expresar sus emociones e ideas con impulsividad, lo cual se representa al comienzo de cada uno de los *tweets*.

Entrenamos el modelo mediante la compilación usando *categorical_crossentropy* que y un optimizador 'Adam'. El algoritmo Adam combina las bondades de Adagrad y RMSprop. Se mantiene un factor de entrenamiento por parámetro y además de calcular RMSprop, cada factor de entrenamiento también se ve afectado por la media del momentum del gradiente. La evolución de la precisión y perdida durante las épocas es la siguiente:

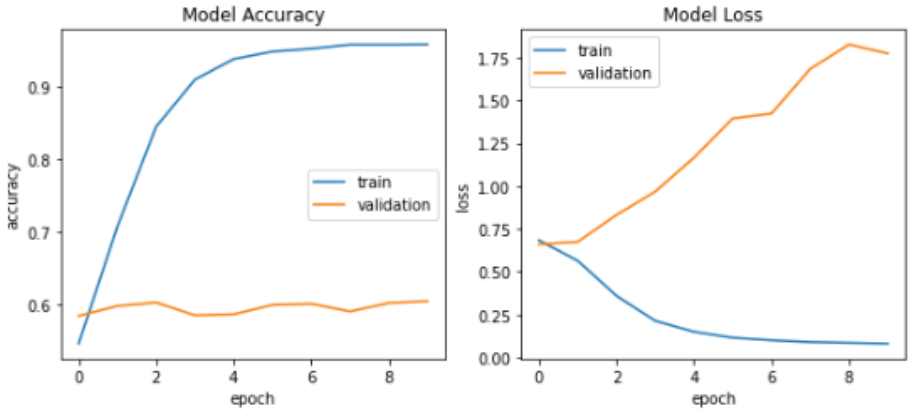


Figura 4.1 Historial de entrenamiento modelo de red 1

Se observa un problema de sobre ajuste y una **precisión del 60%** estable. La métrica es la esperada debido al problema que se intenta solucionar. Una predicción del bitcoin no debería devolver resultados mayores de los encontrados.

La matriz de confusión representa la predicción del modelo con la clasificación real de cada uno de los *tweets*. Esta matriz nos ayuda a entender cuantos días de retorno positivo o negativo de bitcoin el modelo ha sido capaz de predecir usando los *tweets* como única fuente de información.

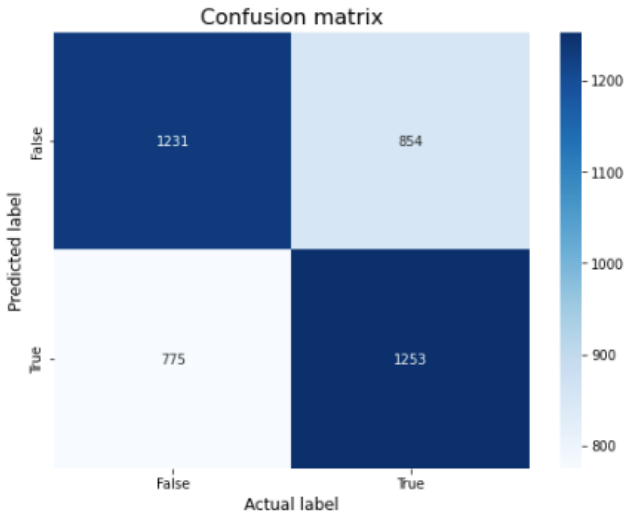


Figura 4.2 Matriz de confusión modelo de red 1

Una vez analizados los resultados del primer modelo, procedemos a entrenar un segundo modelo para intentar reducir el problema de sobre ajuste e intentar mejorar los resultados de predicción. Con un nuevo modelo de red con mayor número de épocas y pocas capas se espera poder mejorarlos.

Modelo de red neuronal 2

Nuestra segunda red se compone de seis capas utilizando las mismas características que la anterior. En este modelo cambian la capa bidireccional de LSTM, permitiendo al modelo aprenden en ambas direcciones el orden de las palabras, y un componente *Dropout* que permite terminar el entrenamiento en caso de no mejorar el aprendizaje. Se añaden las siguientes características como intención de mejorar las predicciones, al entrenar con cincuenta épocas, se incluye el *Dropout* para mejorar los tiempos de computación.

La segunda red procesa los *tweets* de la misma manera que la red neuronal uno, tokeniza y convierte en secuencia, con un máximo de 50 palabra por *tweet*. Las dimensiones de los sets de entrenamiento, testeo y validación cambian debido a que para esta red se usa el set de validación para mejorar el modelo entrenado.

- **Set de entrenamiento**: 12.339 *tweets*
- **Set de Validación**: 4.113 *tweets*
- **Set de Testeo**: 4.113 *tweets*

Entrenamos y evaluamos el modelo de red neuronal 2. Obtenemos una **precisión del 53%.** Los gráficos de precisión del modelo y perdida son los siguientes:

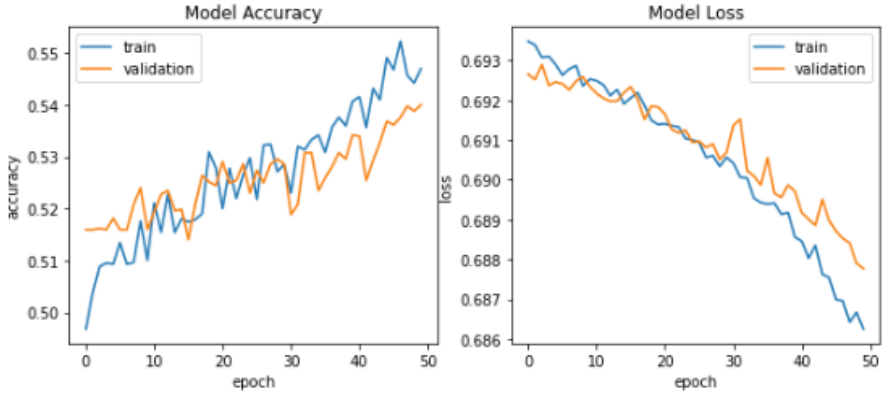


Figura 4.3 Historial de entrenamiento modelo de red 2

Conseguimos eliminar el problema de sobre ajuste, pero la precisión del modelo disminuye en comparación con el 60% del modelo previo. Podemos asumir que el aumento de complejidad y características de la red no ha conseguido mejorar los dos objetivos propuestos. Y que para el problema de clasificación el primer modelo devuelve mejores resultados.

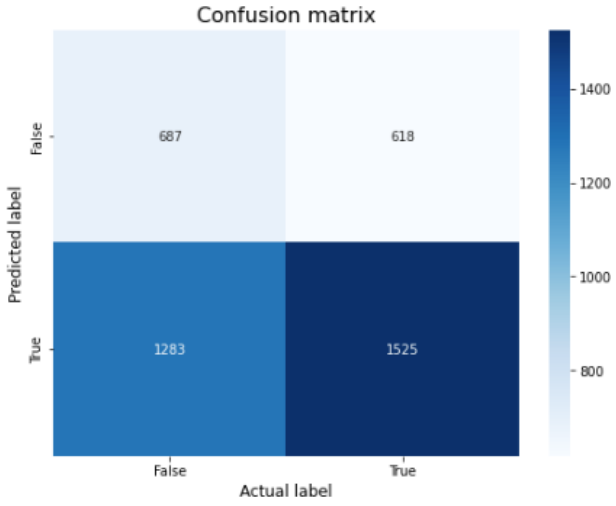


Figura 4.4 Matriz de confusión modelo de red 2

Como se puede observar, las predicciones de la categoría positiva son buenos, a diferencia de los negativos. Este resultado de predicción únicamente serviría para identificar los periodos de crecimiento del valor del bitcoin. La criptomoneda como cualquier otro activo tiene dos direcciones y es de interés contar con un modelo capaz de identificar ambas.

8. Conclusiones

Una vez analizado todo el proceso de modelado y estudio de las variables, teniendo en cuenta las predicciones y las métricas utilizada para determinar la validez de los experimentos, los resultados sugieren que existe una relación entre los *tweets* y los movimientos de precio del bitcoin. Se acepta la hipótesis como válida basándome en los resultados de predicción y la precisión obtenida, pero se recomienda una mayor búsqueda de los parámetros óptimos para mejorar la capacidad de predicción de los modelos.

El cambio de las variables en el proceso seguido de ingesta y muestreo de los datos permite aumentar el número de datos utilizado en los modelos y consecuentemente conocer e identificar nuevos o existentes patrones en la composición de los *tweets* que aportaría más información en las predicciones. La base de referencia utilizada para ponderar la métrica de sentimiento de los *tweets* permite explorar la variable para influir los resultados. El estudio de la influencia de la base en las correlaciones entre el bitcoin y los *tweets* ayudaría a encontrar la relevancia óptima.

En relación con el análisis de tópicos, el análisis de estos ha conllevado a un mejor entendimiento de los temas hallados en los tweets y la influencia que tienen en el bitcoin. Durante los análisis, el tópico cuatro mostró un periodo anormal de crecimiento de *tweets* coincidiendo con un aumento de volatilidad que se deriva en un aumento del valor del bitcoin. Aumentando el número de tópicos en los que se categorizan los *tweets* podría identificar nuevos temas que ayuden a explicar el movimiento del bitcoin. Durante la búsqueda de tópicos se halló que un menor número generalizaba los temas con precisión. Aumentar los tópicos sería eficiente con un aumento de los datos de muestra, de esta manera se incluiría más información en los modelos y se podría identificar temas nuevos.

Durante los entrenamientos y el análisis de resultados, con la matriz de confusión se encuentra una mejor capacidad de predicción de los retornos positivos y negativos en el modelo simple de red neuronal, aunque mostrando efectos de *overfitting*. Este problema se soluciona con la complejidad y características de la segunda red, aunque afectando negativamente a los resultados de predicción. El segundo modelo resulta de una mejor capacidad de predicción para los periodos de crecimiento, convirtiéndolo en un modelo poco eficiente para evaluar el futuro movimiento de un activo con dos direcciones, positiva y negativa.

Apéndice

Gráfico temporal de Tópico 2

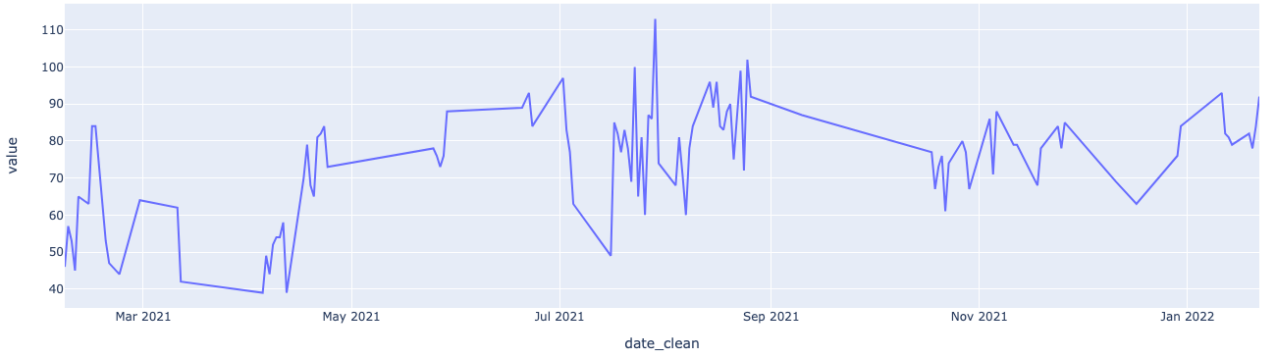


Gráfico temporal de Tópico 3

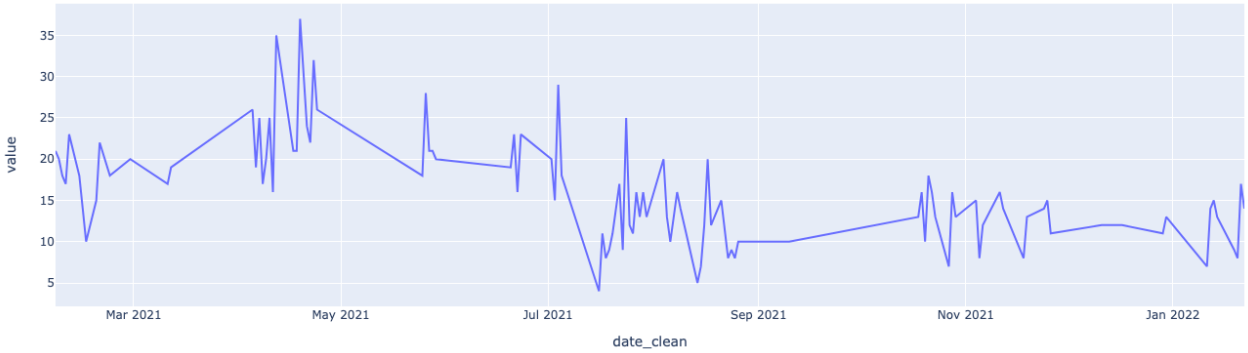
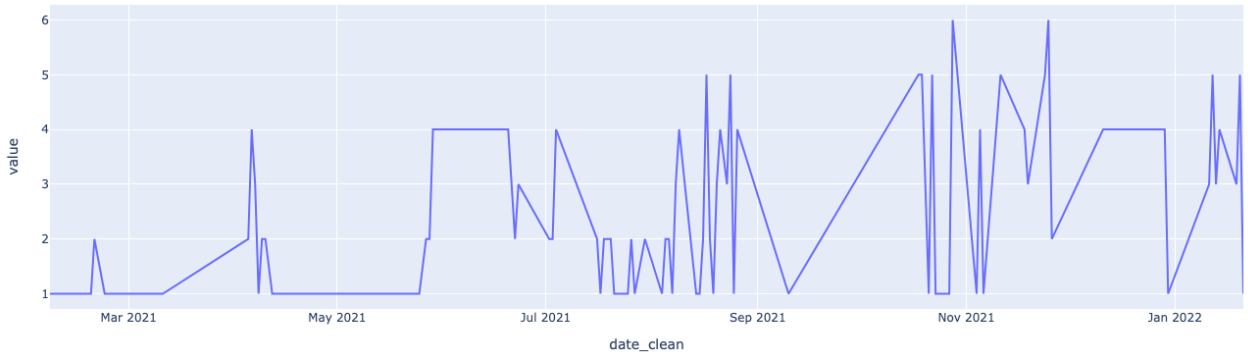
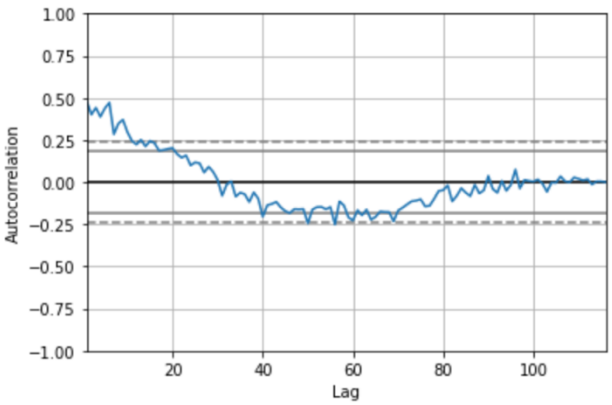


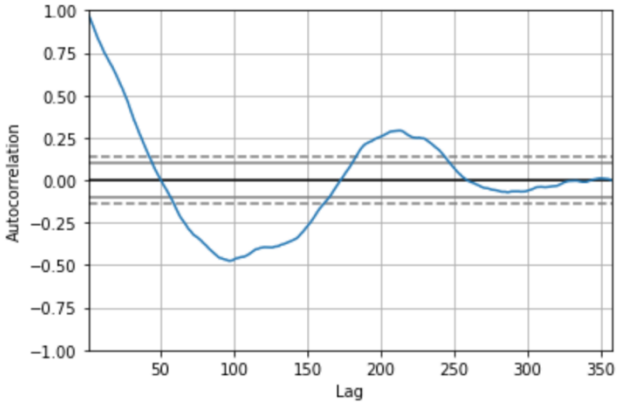
Gráfico temporal de Tópico 5



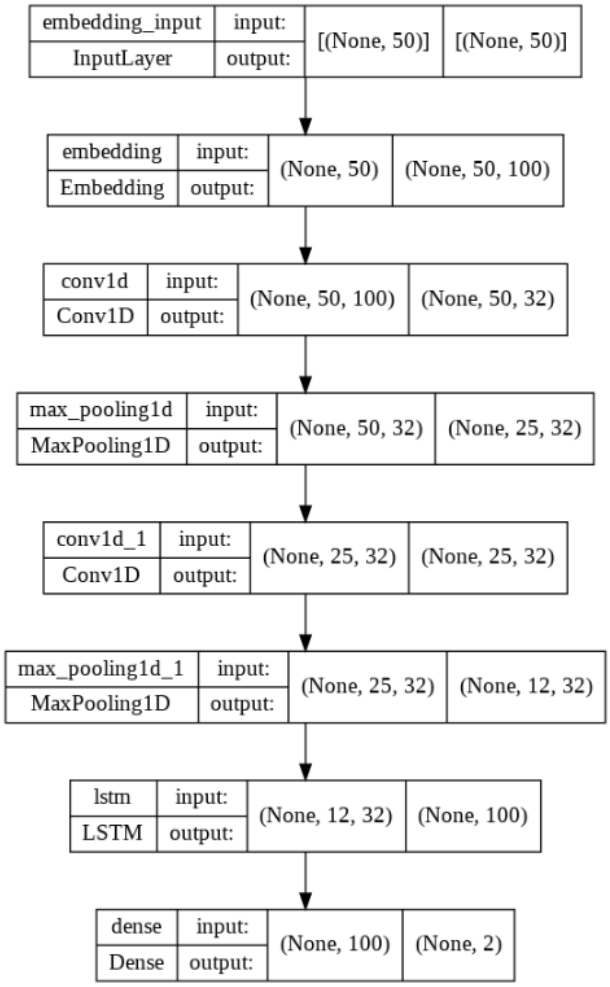
Tweets autocorrelación de sentimiento



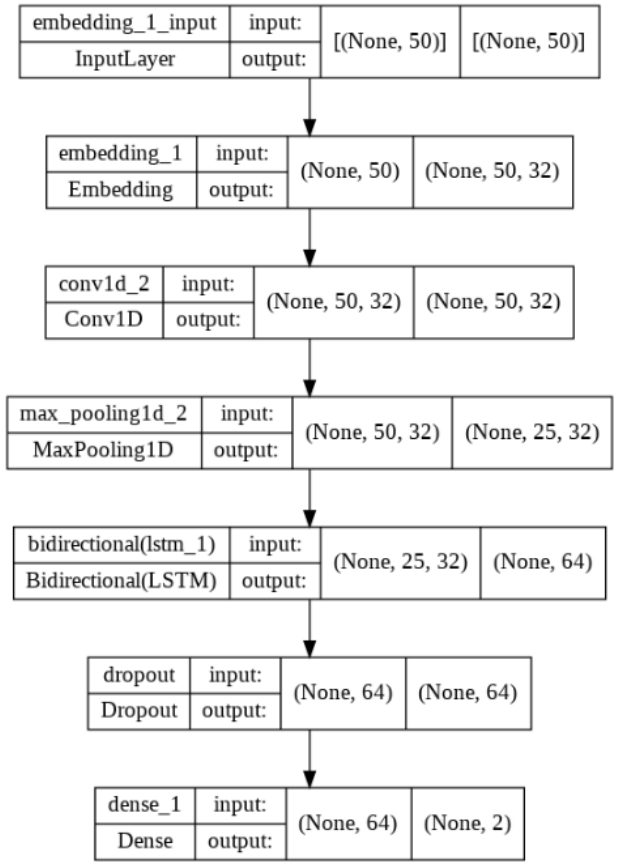
Bitcoin Autocorrelación



Estructura de Red Neuronal 1



Estructura de Red Neuronal 2



Referencias

Links:

https://www.pelicoin.com/blog/what-is-the-economic-impact-of-cryptocurrency

https://www.investopedia.com/ask/answers/032615/what-are-some-examples-stratified-random-sampling.asp

https://medium.com/@piocalderon/vader-sentiment-analysis-explained-f1c4f9101cd9

https://money.usnews.com/investing/articles/the-history-of-bitcoin

https://www.elmundo.es/economia/macroeconomia/2021/06/14/60c71107fc6c833a628b4583.html

Literatura:

KHLAED MOKNI, AHMED BOUTESKA & MOHAMMED SAHBI NAKHLI (2022), Investor sentiment and Bitcoin relationship: A quantile-based analysis, *The North American Journal of Economics and Finance*. 2022.

LI, YAN & DAI, WEI (2020). Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model. *The Journal of Engineering*. 2020.

HUTTO, C.J. & GILBERT, ERIC (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and social media*, ICWSM 2014.

CRITIEN, J.V., GATT, A. & ELLUL, J. Bitcoin price change and trend prediction through twitter sentiment and data volume. *Finance Innovation 8, 45*, 2022.

IBRAHIM, AHMED & KASHEF, RASHA & LI, MENGLU & VALENCIA, ESTEBAN & HUANG, ERIC. (2020). Bitcoin Network Mechanics: Forecasting the BTC Closing Price Using Vector Auto-Regression Models Based on Endogenous and Exogenous Feature Variables. *Journal of Risk and Financial Management*, 2020.