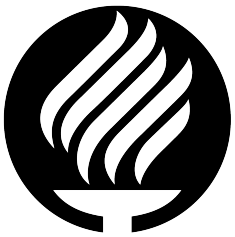
Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Querétaro
Departamento de Computación y Mecatrónica



Tecnológico de Monterrey

Prediciendo los Movimientos del Mercado de Criptomonedas Utilizando Sentiment Analysis y Machine Learning

por

Daniel Jesús Amezcua Sánchez

Proyecto Integrador para el Desarrollo de Soluciones Empresariales

Ingeniería
en
Sistemas Computacionales

Dr. Alfonso Gómez Espinosa Dr. Benjamín Valdés

Santiago de Querétaro, Querétaro, México
03/06/2020

.All models are wrong, but some are useful."

George Box

Abstract

El mercado de criptomonedas ha sido sujeto a diversos estudios en los últimos años por la creciente popularidad que diferentes criptomonedas han adquirido. Estos estudios buscan comprender el fenómeno de las criptomonedas y obtener una visión más clara de cuál es el futuro de estas monedas. Por otro lado, el constante flujo de críticas y opiniones que es inyectado en redes sociales diaramente, ha provocado que el Análisis Sentimental se convierta en una de las herramientas más utilizadas en diversos ámbitos de dominio social y económico. Los comportamientos y percepciones de la realidad de las personas, incluidos los inversionistas de criptomonedas, están influenciados en gran medida por las opiniones de otros. Esta combinación de circunstancias ha sido aprovechada previamente en trabajos que intentan predecir el movimiento y el valor de las criptomonedas dentro del mercado con base en el sentimiento manifestado por las personas en redes sociales. Mientras que la mayoría de estos estudios se centran en la red social Twitter, este trabajo explora la alternativa de la red social Reddit por las diferencias fundamentales que presenta ante otras redes sociales. Junto al sentimiento manifestado en noticias relacionadas a criptomonedas, se utilizaron datos sentimentales de la red social Reddit para constuir modelos clasificadores utilizando diferentes algoritmos de Machine Learning para evaluar si Reddit podía ser una fuente de información para construir modelos con un mejor desempeño que los logrados con la red social Twitter. Los resultados que se obtuvieron sugieren que este no es el caso, sin embargo, se lograron modelos que vencen el modelo aleatorio con certeza para las criptomonedas Bitcoin, Litecoin, Ripple y Ethereum.

Índice general

Capítulo 1

Introducción

A lo largo de el presente documento se describe el trabajo realizado en torno a la investigación de la posibilidad de utilizar información sentimental proveniente de noticias y de comentarios y publicaciones de la red social Reddit para entrenar modelos de Machine Learning y predecir los movimientos del mercado de las criptomonedas.

Esta es una contribución al trabajo de Valencia, Gómez y Valdés [1], quienes demostraron en su trabajo que un buen precursor a los movimientos dentro del mercado de las criptomonedas son los datos sentimentales obtenidos de la red social Twitter.

1.1. Contexto del problema

Las criptomonedas se han hecho un lugar importante en el mercado internacional desde su aparición en 2009 con Bitcoin [2]. Una criptomoneda es una moneda digital basada en criptografía. Como cualquier otra moneda, las criptomonedas son usadas para adquirir diferentes bienes o servicios. La característica más innovadora de las criptomonedas es que la confiabilidad, integridad y validez de las transacciones que las involucran se logran utilizando tecnología *blockchain*, en la cual una transacción es verificada por diferentes nodos de una red de computadoras. Convencionalmente, esta tarea es delegada a instituciones u organismos como bancos y gobiernos que se encargan además de emitir y distribuir las monedas. Con las criptomonedas, no existe la necesidad de una institución intermediaria como estas que se encargue de regular su uso.

La posibilidad latente de una descentralización de la moneda es un fenómeno sin precedentes y de naturaleza disruptiva, por lo que la postura que los gobiernos de diferentes países en todo el

mundo han tomado ante las criptomonedas y las medidas que han adoptado ante estas varían de país en país. Mientras que en países como Bolivia y Vietnam el uso de criptomonedas es ilegal, otros países como España y Bielorrusia permiten la libre inversión sobre estas bajo la premisa de que es un incentivo para fomentar la inversión en compañías de tecnología. [3] [4]

A pesar de las restricciones que se han impuesto al uso de criptomonedas en diferentes países, estas han logrado adquirir presencia en el mercado internacional y han sido gradualmente adoptadas como forma de pago en una amplia variedad de servicios y plataformas [4] [5].

La capitalización de mercado de las criptomonedas entró en crecimiento exponencial en el año 2016 [6], y, de acuerdo al sitio CoinMarketCap, en Mayo de 2020 [7], esta alcanzó los 248 mil millones de dólares. Al compararlo con el mercado de acciones y divisas, el mercado de las criptomonedas tiene un comportamiento parecido al mercado de acciones, lo cual lo describe como un mercado de alto riesgo; también, demuestra ser un mercado frágil al no exhibir ningún tipo de estructura como sí lo hacen sus similares, quienes muestran agrupaciones de acuerdo a la región geográfica donde se analicen. [8] El comportamiento caótico e inestable que han tenido los valores de las criptomonedas durante su corto tiempo de vida hacen de las estrategias de compraventa alrededor de las criptomonedas difíciles de desarrollar y riesgosas de llevar a cabo. Por lo mismo, este mercado está constantemente sujeto a estudios que buscan esclarecer su complejidad y tener un mejor entendimiento de cómo es y hacia dónde va.

Una forma interesante de observar este mercado es a través de la conectividad entre las criptomonedas que lo conforman. Esta conectividad se ha mantenido fuerte desde el surgimiento del mercado [9], y es más pronunciada en los tiempos de alta volatilidad de los precios, lo cual es una propiedad que no se presenta en otros mercados [10]. Sin embargo, en términos generales, esta conectividad se ha vuelto más compleja en función del tiempo [11] , lo que sugiere una independización gradual de las criptomonedas a sus homólogas más influyentes, e ahí el considerar la importancia de considerarlas a cada una por separado. Al igual que en el mercado de valores, el continuo flujo de información que está disponible al público en redes puede tener una influencia considerable en la opinión de los inversionistas del mercado de criptomonedas.

La mera naturaleza innovativa de las criptomonedas y la incertidumbre de la seriedad de sus usos en el futuro, sumado a su carencia de valor intrínseco y otros factores sociales [12], hacen que la compraventa de las criptomonedas se base meramente en la especulación, lo cual hace que los agentes inversores sean fácilmente influenciados a vender o comprar criptomonedas [13]. Dicho esto, podemos entender que el que existan grandes cantidades de información acerca del

sentimiento de las personas con respecto a diferentes criptomonedas -especialmente en redes sociales y noticias disponibles en la red, nos brinda una oportunidad de disminuir la incertidumbre respecto al próximo paso que tomará un inversionista de criptomonedas.

.

1.2. Objetivo del trabajo

Este trabajo pretende evaluar la red social Reddit y noticias como fuentes de información para predecir el movimiento de precio para diferentes criptomonedas dentro del mercado de criptomonedas.

1.3. Estructura del trabajo

Este trabajo está compuesto por 7 Capítulos que presentan el trabajo sobre el que se realizó esta investigación, así como los detalles experimentales y procedurales que involucraron la consecución del objetivo.

El Capítulo 2 se enfoca en describir el trabajo realizado previamente en el área y otros trabajos de interés que aportaron a la investigación llevada a cabo.

El Capítulo 3 presenta las limitantes observadas en los trabajos descritos en el Capítulo 2 que sirvieron como motivación y justificación de este trabajo.

El Capítulo 4 presenta el diseño detallado de la solución que permitió obtener los modelos requeridos para obtener resultados.

El Capítulo 5 presenta los métodos para comparar los modelos obtenidos a partir de la solución con el objetivo deseado y los trabajos realizados previamente en el área.

El Capitulo 6 presenta los modelos resultantes más relevantes para la investigación.

En el Capítulo 7 se discuten los resultados obtenidos, se presentan las conclusiones a partir de estos resultados a la vez que se define la dirección que deben tomar los trabajos futuros que tengan base en este trabajo.

Capítulo 2

Estado del Arte

2.1. Contexto del área

2.1.1. Análisis Sentimental

El Análisis Sentimental es una rama de el estudio del Procesamiento del Lenguaje Natural, el cual a su vez es un campo de las ciencias computacionales. Se encarga de analizar y extraer información subjetiva como opiniones, sentimientos, evaluaciones, actitudes y emociones a partir de el lenguaje escrito. Debido a la creciente disponibilidad de datos como críticas y opiniones que manifiestan las personas en diferentes redes sociales, el Análisis Sentimental ha ganado popularidad y su uso se ha extendido a las ciencias de la administración y las ciencias sociales [14].

Esta técnica ha sido aplicada en diversos estudios que incluyen desde obtener y clasificar críticas de aplicaciones móviles [15], hasta el desarrollo de estrategias de marketing y la predicción del mercado de valores utilizando opiniones manifestadas en redes sociales [16].

Existen diferentes formas para realizar Análisis Sentimental:

- Análisis Sentimental a Nivel de Documento
- Análisis Sentimental a nivel de oraciones
- Análisis Sentimental basado en aspectos
- Análisis Sentimental comparativo

La elección de una técnica de Análisis Sentimental depende de la naturaleza de los textos a examinar. La mayoría de estas técnicas incorporan algoritmos de aprendizaje supervisado y no supervisado para catalogar un extracto de texto como positivo o negativo. Por otro lado, estas técnicas requieren de un léxico de sentimientos, el cual se compone de un conjunto de palabras que tienen asignado un sentimiento positivo o negativo. Este léxico de sentimientos varía según el campo de estudio y es uno de los recursos más importantes para el análisis sentimental [17].

2.2. Trabajos Relacionados

La predicción del mercado de valores ha sido un área de interés que ha atraído a investigadores de múltiples disciplinas por la importancia que este tiene dentro del sistema económico que rige al mundo. Esta tarea resulta increíblemente compleja debido a la naturaleza caótica del mercado. El estado del arte del arte para el mercado de valores fue descrito por Shah, Isah, y Zulkernine [18], quienes sitúan diferentes técnicas desarrolladas dentro de cinco diferentes enfoques: estadístico, reconocimiento de patrones, aprendizaje automático, análisis sentimental, o un híbrido conformado por las anteriores. Dentro de la categoría de Aprendizaje Automático, Random Forests y Support Machine Learning destacan en la tarea de predecir los movimientos del mercado de valores, como lo muestra también el trabajo de Ballings, Van den Poel, Hespeels y Gryp [19], quienes comparan el desempeño de múltiples clasificadores simples y ensamblados en la misma tarea.

Para el alcance de este trabajo, nos enfocaremos en las técnicas de aprendizaje automático y de análisis sentimental aplicados al mercado de criptomonedas. Debido a la corta edad que posee este mercado, para nuestro conocimiento existen un número limitado de trabajos que aportan a la literatura. Sumado a esto, la fragilidad y volatilidad que caracteriza a este mercado hace que la mayor parte de la literatura ofrezca trabajos orientados a predecir los movimientos del mercado y no los valores exactos de las criptomonedas; es decir, tratan la tarea como un problema de clasificación y no como un problema de regresión. Dicho esto, los trabajos más relevantes al respecto, orientados tanto a predecir el valor como el movimiento del precio, se describen a lo largo de las siguientes líneas.

Chowdhury, Rahman, y Rahman [20] proponen utilizar los modelos creados por: Ensemble Learning Method, Gradient Boosted Tree Model, Neural Net Model y K-Nearest Neighbor (K-NN) para predecir el valor de cierre de nueve diferentes criptomonedas tomando como datos

sus precios de cierre diarios a lo largo de dos años. Ensemble Learning Method resalta entre los otros modelos al realizar predicciones con hasta el 0.924 de exactitud. Regal, Morzán y Fabri [21] agregan publicaciones de la red social Twitter y utilizan redes neuronales de arquitectura Long Short Term Memory en su trabajo para predecir los precios que tuvo Bitcoin durante 2018, encontrando que no es una aproximación adecuada para atacar la tarea al tener un error estándar del 34.28 %. Por otro lado, Vo, Nguyen y Ock [22] realizaron Análisis Sentimental a noticias relacionadas a criptomonedas y, junto datos históricos del mercado predijeron los valores de las criptomonedas utilizando series temporales con redes neuronales LSTM y Support Vector Regression. Trabajaron con información de un período de tiempo de poco más de un año encontraron que, cuando las redes LSTM son entrenadas con datos sentimentales y del mercado, predicen con un error mANE del 1.36 % los valores de las criptomonedas.

Lamon, Nielsen y Redondo [23] utilizan en un enfoque ligeramente diferente, utilizando publicaciones de Twitter y encabezados de noticias relacionados a las criptomonedas de un período de 67 días para predecir las fluctuaciones de Bitcoin, Ethereum y Litecoin. Etiquetan las publicaciones y encabezados de noticias no como positivos o negativos, sino con los cambios de precio que sufrieron las criptomonedas en los días siguientes a la fecha de publicación. De esta forma obtienen además una predicción del valor de las criptomonedas. Utilizan las dos fuentes de datos por separado para entrenar modelos de Regresión Logística, Naive Bayes y Support Vector Machine, encontrando que Naive Bayes y Regresión Logística ofrecen una mayor precisión a la hora de predecir las fluctuaciones con hasta un 75.8 % de precisión en Ethereum. Sin embargo, ninguno de los modelos resultó ser preciso para predecir los valores de las criptomonedas.

Wolk [24] analizó en su trabajo la relación que hay entre la presencia de una criptomoneda en una red social con su valor en el mercado. Utilizó el número de publicaciones realizadas en la red social Twitter con respecto a diferentes criptomonedas y, junto a las tendencias de búsqueda de las mismas obtenidas de Google Trends, encontró una relación evidente entre el número de publicaciones y búsquedas relacionadas a una criptomoneda con su valor en el mercado al probar con diferentes períodos de tiempo y diez modelos de aprendizaje automático, siendo Support Vector Regression y Multilayer Perceptron Neural Networks los que mejor desempeño tuvieron. Similarmente, Abraham, Higdon, y Nelson [25] utilizaron publicaciones de Twitter y datos de Google Trends de un período de 30 días para incorporar datos sentimentales a un modelo de regresión múltiple para proyecciones de los valores de Bitcoin y Ethereum, encontrando como resultado que utilizar el volumen de búsquedas y publicaciones es un mejor predictor que el sentimiento contenido en estos. Así mismo, señalan que sus resultados pueden verse

predispuestos por la fuerte tendencia a aumentar de precio que tuvieron estas monedas durante el período investigado.

Valencia, Gómez y Valdés [1] utilizaron publicaciones en un periódo de tiempo de 80 días de la red social Twitter para medir el sentimiento manifestado con respecto a diferentes criptomonedas y, junto a datos históricos del mercado, compararon las predicciones de modelos entrenados con Support Vector Machine, Random Forest y Neural Networks, de los cuales Support Vector Machine y Random Forest resultan ser los mejores, logrando una exactitud de hasta el 65 % cuando los modelos son entrenados incluyendo datos del análisis sentimental. De forma similar, Hitam, Ismail y Saeed [27] utilizaron una versión optimizada de Support Vector Machine para predecir los movimientos de diferentes criptomonedas utilizando solamente datos históricos del mercado en un período de tiempo de 5 años, logrando una exactitud de hasta el 97 % en la criptomoneda Ethereum.

2.3. Conclusión

Un factor común que resalta entre los diferentes trabajos mencionados anteriormente que utilizan la red social Twitter, es que el período de tiempo que sirvió como entrenamiento para los modelos, es relativamente corto. Debido a la inestabilidad que presenta el mercado de criptomonedas, estos modelos pueden atrapar el comportamiento de estos períodos de tiempo, más no tener un buen desempeño en otros períodos de tiempo, como es mencionado en el trabajo [25]. Por lo que un análisis que involucre Análisis Sentimental de un período de tiempo de más de 3 meses aún es necesario.

Otro punto a resaltar es la efectividad la efectividad lograda con las redes neuronales Multilayer Perceptron y los algoritmos con base en vectores de soporte.

Capítulo 3

Planteamiento del Problema

Este capítulo discute los principales problemas y limitantes observados en los trabajos descritos en el Capítulo 2, los cuales justifican el objetivo de este trabajo y la forma en que se realizó.

3.1. El problema de las identidades automatizadas

Se han utilizado redes sociales como Twitter y Facebook, así como encabezados y contenidos de noticias en trabajos previos para intentar construir modelos que predigan los movimientos y los precios tanto en el mercado de criptomonedas como en el mercado de divisas y valores.

Un problema que dificulta el análisis sentimental dentro de estas redes sociales es que existe una gran cantidad de información proveniente de identidades automatizadas, o bots, que en lugar de expresar una opinión genuina respecto a un tema, buscan modelar la opinión general del público de acuerdo a los intereses de los grupos de personas que emiten a estas identidades[27]. Este volumen de publicaciones actúan como ruido a la información que se busca analizar, y esta puede ser una de las razones por las que en algunos de los estudios que utilizan la red social Twitter para predecir los precios y los movimientos del mercado de criptomonedas, el número de publicaciones es un mejor predictor que el sentimiento general proveniente de estas.

3.2. La recolección de datos con Twitter

Otra limitante que se encuentra en la red social Twitter, es que la disponibilidad de los datos está restringida para información proveniente de fechas previas a la fecha en la que se quiere

realizar la extracción [28]. Esto significa que el tiempo de ejecución de la recolección de datos es equivalente al período de tiempo del que se quieren obtener datos. Este obstáculo impide que se puedan entrenar modelos con información sentimental histórica de largos períodos de tiempo, como sí se puede hacer con información histórica del mercado, pues esta última está disponible en su totalidad en diferentes sitios web.

Esta limitante tiene un impacto directo en los conjuntos de datos que se han utilizado en trabajos previos relacionados a la predicción del mercado de criptomonedas utilizando información de esta red social, pues el volumen del conjunto de datos utilizado en cada investigación se acota a las fechas en las que la investigación relacionada fue llevada a cabo.

3.3. Experimentación utilizando otras redes sociales

Como se mencionó previamente, la mayoría de los trabajos que utilizan información proveniente de redes sociales para predecir los movimientos en el mercado de criptomonedas exploran principalmente las plataforma Twitter.

Reddit es una red social enfocada a foros de discusión. Los usuarios se "suscriben" a foros que debaten temas de su interés. A diferencia de la red social Twitter, Reddit es una red social más apropiada para encontrar opiniones críticas en diferentes materias [29]. La disponibilidad de los datos está disponible para cualquier fecha [30] por lo que se pueden entrenar modelos a partir de datos sentimentales de un intervalo largo de tiempo.

El volumen de publicaciones y el número de usuarios activos es similar al de la red social Twitter. La cuenta más relevante de Bitcoin en Twitter cuenta con más de 1 millón de seguidores [31], mientras que la comunidad más relevante de Bitcoin en Reddit cuenta con cerca de 1.5 millones de lectores [32]. Comparaciones similares se encuentra a lo largo de las cuentas y comunidades de diferentes criptomonedas.

Por las características que presenta, Reddit pareciera un buen candidato a ser sometido a análisis sentimental para predecir los movimientos del mercado de criptomonedas, sin embargo, y hasta el conocimiento de los autores, no existen trabajos que exploren esta alternativa para el problema de la predicción del mercado de criptomonedas.

Capítulo 4

Solución

El producto final de la solución se compone de uno o diferentes modelos con los que se pueden realizar predicciones diarias de la dirección del valor de una particular criptomoneda en el mercado. Los modelos son construidos a partir de la relación sentimiento-movimiento de mercado observada durante un período de nueve meses.

Se siguieron las prácticas realizadas en la literatura existente de escoger más de una fuente de información para desarrollar los modelos. Con la variación de que en lugar de usar encabezados de noticias se utilizaron textos completos de noticias, se siguió en parte el modelo presentado en el trabajo de Lamon, Nielsen y Redondo [23] y se escogieron Reddit y noticias relacionadas a las criptomonedas como fuentes de información.

Respecto a las criptomonedas a someter a estudio, se seleccionaron Bitcoin, Litecoin, Ripple y Ethereum por su popularidad a la fecha de este trabajo y por su reiterada selección a lo largo de los diferentes trabajos existentes en la literatura. A lo largo de esta sección se describen las tareas llevadas a cabo para la obtención de dichos modelos. La secuencia en la que se describen las siguientes tareas representa la secuencia seguida en la práctica para desarrollar los modelos.

Los scripts que implementan la totalidad de la solución presentada en este capítulo se pueden encontrar en el repisotorio de Github dedicado a este trabajo. [1]

[1] https://github.com/danielamezcua/crypto-market-predictions

4.1. Selección de fuentes

4.1.1. Subreddits

Los subreddits son el nombre con el que se le conoce a cada foro de discusión dentro de la red social Reddit. En la Figura 4.1 se exponen los subreddits seleccionados para cada criptomoneda, los cuales fueron seleccionados por su relevancia dentro de los diferentes subreddits dedicados a cada criptomoneda y por la actividad diaria que han presentado en los últimos años.

	Bitcoin	**Ripple**	**Litecoin**	**Ethereum**
Subreddits	r/Bitcoin r/btc	r/xrp r/Ripple	r/Litecoin r/LitecoinMarkets	r/ethtrader r/EthFinance

FIGURA 4.1: Subreddits seleccionados por cada criptomoneda

4.1.2. Noticias

Existen diferentes sitios web que se encargan de cubrir noticias relacionadas a las criptomonedas. Se seleccionó el sitio Cointelegraph para la extracción de noticias por la variedad de criptomonedas que cubre y por la frecuencia con la que se publican noticias.

4.2. Extracción de datos

Todos datos recolectados durante esta investigación pertenecen a el período de tiempo comprendido desde el 1 de septiembre de 2019 hasta el 30 de abril de 2020. El número de noticias, comentarios y publicaciones obtenidos para diferentes criptomonedas se muestra en la figura 4.2 A cada extracto de datos le correspondía una fecha de publicación, la cual fue normalizada al estándar de tiempo UTC.

4.2.1. Reddit

Reddit ofrece una API al público que permite operaciones de consulta y de publicación, lo cual facilita enormemente la extracción de los datos. Todas las operaciones relacionadas a esta API que se realizaron para este trabajo cumplen con los términos de uso de la API de Reddit [33].

Criptomoneda	Publicaciones	Comentarios	Noticias
Bitcoin	64,769	678,563	1676
Ripple	5,514	38,394	233
Litecoin	2,585	22,879	132
Ethereum	22,563	200,885	473

FIGURA 4.2: Número de publicaciones, comentarios y noticias obtenidos para cada criptomoneda

Praw [34] es un módulo de código abierto para Python que facilita el acceso a la API para comunicarse con Reddit. Se utilizó este módulo en los scripts encargados de obtener y almacenar todos los comentarios y publicaciones realizados dentro del período de tiempo antes mencionado. Los metadatos asociados a cada publicación y comentario que se almacenaron se encuentran descritos en las figuras 4.3 y 4.4 respectivamente.

Atributo	Descripción
Fecha de publicación	Fecha de publicación de la publicación medida en segundos desde el Epoch
Número de comentarios	Número de comentarios asociados a la publicación
Puntaje	Medida de aprobación o rechazo de la publicación dentro del subreddit asociado.
Selftext	Contenido de la publicación
Id subreddit	Identificador del subreddit al que la publicación está asociado
url	Url donde se puede encontrar la publicación
Discusión diaria	Valor booleano que indica si la publicación es o no una publicación enfocada a la discusión de la criptomoneda.

FIGURA 4.3: Metadatos asociados a cada publicación de Reddit

Atributo	**Descripción**
Fecha de publicación	Fecha de publicación de la publicación medida en segundos desde el Epoch
Id Link	Identificador de la publicación a la que el comentario está asociado
Puntaje	Medida de aprobación o rechazo de la publicación dentro del subreddit asociado.
Selftext	Contenido del comentario
Id subreddit	Identificador del subreddit al que el comentario está asociado

FIGURA 4.4: Metadatos asociados a cada comentario de Reddit

4.2.2. Cointelegraph

Los datos de las noticias del sitio Cointelegraph fueron extraídos utilizando la técnica web crawling, que consiste en analizar la estructura de un sitio web para navegar por él a la vez que se extrae información relevante de forma automática.

Scrapy es un módulo para Python que facilita el desarrollo de *crawlers* [35]. Los scripts desarrollados para este trabajo que se encargan de almacenar los datos de noticias hacen uso de las funciones provistas por este módulo. Los metadatos asociados a cada noticia que se almacenaron se presentan en la figura 4.5.

Atributo	**Descripción**
Título	Título de la noticia
Autor	Nombre del autor o autores de la noticia
Fecha de publicación	Fecha de publicación de la noticia medida en segundos desde el Epoch
Contenido	Contenido de la noticia
Tags	Etiquetas que permiten asociar la noticia a diferentes criptomonedas
Url	Url en donde se puede encontrar la noticia
Análisis	Valor booleano que indica si la noticia es más bien un artículo dedicado al análisis del precio de diferentes criptomonedas

FIGURA 4.5: Metadatos asociados a cada noticia del sitio Cointelegraph

4.2.3. Datos históricos del mercado

Se seleccionó el sitio Cryptocompare [36] para extraer los datos históricos del mercado por la facilidad que ofrece para brindar datos al poner a disposición del público una API. Las funciones que provee esta API en su versión gratuita fueron suficientes para obtener los datos necesarios para el propósito de esta investigación. En la figura 4.6 se muestran los datos del mercado que se almacenaron para las criptomonedas antes mencionadas.

Atributo	**Descripción**
Fecha	Fecha asociada a los datos
Max	El valor máximo que alcanzó la criptomoneda durante esta fecha
Min	Fecha de publicación de la noticia medida en segundos desde el Epoch
Apertura	El valor que la criptomoneda tenía al inicio de la fecha
Cierre	El valor que la criptomoneda tenía al momento del fin de la fecha
Moneda	Criptomoneda asociada a los datos anteriores

FIGURA 4.6: Datos del mercado almacenados para las diferentes criptomonedas

4.3. Pre procesamiento de los datos

4.3.1. Análisis Sentimental

4.3.1.1. Comentarios y Publicaciones de Reddit

(Valence Aware Dictionary and sEntiment Reasoner) VADER [37] es una herramienta enfocada al análisis de sentimiento del léxico y que además está afinada para procesar texto extraído de redes sociales. Se utilizó esta herramienta por su popularidad dentro de los trabajos que involucran análisis sentimental en redes sociales. La forma en la que VADER realiza el análisis sentimental es a través de un diccionario de palabras y un conjunto de reglas que permiten calcular un valor agregado de sentimiento. Los contenidos de cada publicación y comentario del conjunto de datos fue analizado para asignarle un valor agregado de sentimiento. Este valor varía dentro del rango inclusivo de -1 a 1, siendo -1 un comentario o publicación altamente negativo, y 1 un comentario o publicación altamente positivo. El criterio para evaluar si un comentario o publicación es positivo o negativo es el siguiente:

$$\text{valor agregado} >= 0{,}05 - > \textit{positivo} \tag{4.1}$$

$$\text{valor agregado} <= -0{,}05 - > \textit{negativo} \tag{4.2}$$

$$-0{,}05 < \text{valor agregado} < 0{,}05 -> neutral \tag{4.3}$$

4.3.1.2. Noticias

Para el análisis sentimental de noticias, de igual forma se utilizó la herramienta VADER, con la variación de que el diccionario que asigna un peso negativo o positivo a cada palabra, fue modificado al agregarle todas las palabras contenidas en el diccionario Loughran-McDonald [38]. A diferencia del diccionario ofrecido por la herramienta VADER, el diccionario Loughran-McDonald fue construido a partir del análisis de textos pertenecientes a noticias financieras. Una limitante de este diccionario es que no le asigna un peso a cada palabra, si no que solo le asigna un sentimiento positivo o negativo, por lo que el peso de estas palabras fue asignado de forma manual, con un valor de -2 a las palabras catalogadas con un sentimiento negativo, y un valor de 2 a las palabras catalogadas con un sentimiento positivo.

A cada noticia se le asignó un valor de sentimiento agregado, de igual forma que se hizo con los comentarios y publicaciones de Reddit, la diferencia es que este valor es un promedio de los valores agregados de sentimiento obtenidos en los diferentes párrafos que componen el texto de cada noticia. Para las noticias que contienen un análisis del precio de las diferentes criptomonedas el proceso de asignación de un sentimiento tuvo una variación: el contenido de la noticia fue dividido y solamente las partes que pertenecían a las criptomonedas de interés eran sujetas al análisis. Por lo tanto, a este tipo de noticias no se les asignó un solo valor sentimental, si no que les asignó un valor sentimental por cada criptomoneda de interés contenida en la noticia.

4.3.2. Síntesis de los datos

Dado que lo que se busca es modelar la relación entre el sentimiento observado en un día en Reddit y noticias con los movimientos del mercado de criptomonedas observados al día siguiente, los datos de comentarios, publicaciones y noticias en su estado singular no sirven para entrenar los modelos, sino que es necesaria una representación que incluya una síntesis de todos los datos asociados a un día en particular y el fenómeno observado al día siguiente en el mercado.

Debido a que las publicaciones y comentarios de Reddit tienen un puntaje asociado, decidimos utilizar esta información para crear además un conjunto de datos aumentado, en el que cada publicación y comentario era considerado de la siguiente forma:

- Si una publicación o un comentario tenían un puntaje positivo, esta publicación o comentario era tomado en cuenta n veces en la síntesis de los datos. Siendo n el puntaje asociado a la publicación o comentario
- Si una publicación o un comentario tenía un puntaje negativo, esta publicación o comentario era desechado para la síntesis de datos.

Además de el conjunto de datos aumentado, se consideró otra forma de sintetizar la información de Reddit: solamente utilizando publicaciones y comentarios pertenecientes a publicaciones enfocados a la discusión de la criptomoneda en cuestión. La razón de crear este conjunto de datos es evaluar el ruido existente dentro de cada foro de discusión de Reddit al eliminar publicaciones que no están enfocadas a discusión.

Considerando las dos alternativas anteriores, al final, se construyeron cuatro conjuntos de datos sintetizados para desarrollar los modelos. Cada conjunto de datos se construyó utilizando diferentes porciones de el conjunto de datos de Reddit, mientras que la porción de datos relativa a noticias e información del mercado fue la misma para todos. Cada conjunto de datos sintetizados, entonces, se distingue por la porción del conjunto de datos de Reddit que se utilizó para construirlo. La figura 4.7 ilustra la porción de datos de Reddit que se utilizó para cada conjunto de datos sintetizado.

	Todos los comentarios y publicaciones de Reddit	Solamente comentarios y publicaciones de discusión de Reddit	Datos sin incluir el puntaje de publicaciones y de comentarios	Datos aumentados con puntaje de publicaciones y comentarios
Sin filtrar y sin aumentar	●		●	
Filtrado y aumentado		●		●
Sin filtrar y aumentado	●			●
Filtrado y sin aumentar		●	●	

Figura 4.7: Porción del conjunto de datos de Reddit utilizada para construir cada conjunto de datos sintetizado

Después de la síntesis del conjunto de datos se obtuvieron, para cada criptomoneda y para cada día, los datos descritos en la figura 4.8.

Atributos		
Fecha	Número de comentarios positivos	Número de noticias publicadas
Valor máximo en el mercado	Nùmero de comentarios negativos	Promedio de sentimiento en las noticias
Valor mínimo en el mercado	Número de comentarios neutrales	Movimiento de precio de esta criptomoneda en el mercado al día siguiente.
Valor de apertura en el mercado	Promedio de sentimiento en los comentarios	
Valor de cierre en el mercado	Promedio de sentimiento de los comentarios positivos	
Criptomoneda asociada	Promedio de sentimiento en los comentarios negativos	

FIGURA 4.8: Datos asociados a cada día y a cada criptomoneda

4.4. Desarrollo de los modelos

Cada experimento que derivó en un modelo predictor consistió en seleccionar las características que servirían como los datos para entrenar el modelo, y la selección de un algoritmo de Machine Learning que regía el procesamiento de los datos y el desarrollo del modelo.

4.4.1. Selección de características

Del conjunto de características mencionadas en la figura 4.8, para cada experimento se seleccionaron un subconjunto de estas características para evaluar el desempeño de estas para predecir el movimiento del mercado de criptomonedas. Los diferentes subconjuntos incluyen características pertenecientes a información de la red social Reddit, a información de noticias, a información del

mercado, o a una combinación de las anteriores. Los diferentes subconjuntos de características utilizados en cada experimento son los siguientes:

1. Número de comentarios totales
2. Promedio de sentimiento en todos los comentarios
3. Número de comentarios totales y promedio de sentimiento en todos los comentarios
4. Promedio de sentimiento en los comentarios positivos
5. Porcentaje de comentarios positivos y porcentaje de comentarios negativos
6. Promedio de sentimiento en las noticias
7. Promedio de sentimiento en los comentarios, promedio de sentimiento en las noticias
8. Número de comentarios, promedio de sentimiento en las noticias, promedio de sentimiento en los comentarios
9. Promedio de sentimiento en los comentarios positivos, promedio de sentimiento en las noticias
10. Porcentaje de comentarios positivos, porcentaje de comentarios negativos, promedio de sentimiento en las noticias
11. Promedio de sentimiento en los comentarios, promedio de sentimiento en las noticias, valor de apertura, valor de cierre, valor máximo, valor mínimo.
12. Promedio de sentimiento en los comentarios positivos, promedio de sentimiento en las noticias, valor de apertura, valor de cierre, valor máximo, valor mínimo.

4.5. Algoritmos de Machine Learning

Se escogieron tres diferentes algoritmos de Machine Learning para desarrollar los modelos: Support Vector Machines, Random Forests y Multilayer Perceptron. Estos algoritmos se seleccionaron por el buen desempeño que han mostrado en trabajos relacionados tanto en el mercado de valores y divisas como en el mercado de criptomonedas.

La implementación de todos los algoritmos se hizo utilizando el módulo de Python scikit-learn [39], que provee las funciones necesarias para entrenar y evaluar los modelos requeridos. Cada algoritmo puede ser ajustado a través de una serie de parámetros que varía con cada algoritmo. A pesar de que scikit-learn provee por defecto los parámetros más estables para cada algoritmo, algunos de estos parámetros fueron ajustados para lograr un mejor desempeño en nuestros modelos.

4.5.1. Support Vector Machines

Este algoritmo de aprendizaje supervisado construye un hiperplano o una serie de hiperplanos que intenta separar de mejor forma posible los puntos de datos de acuerdo a la clasificación que estos tienen. Este hiperplano está representado por un vector que separa los puntos más cercanos de dos diferentes clases, el vector de soporte. Los datos son representados en el espacio a través de una función kernel [40] .La función kernel que se utilizó para construir los modelos es la función de base radial, la cual es la función kernel por defecto que ofrece el módulo scikit-learn.

4.5.2. Random Forests

Los Random Forests son un algoritmo de clasificación que combina múltiples árboles predictores de tal forma que cada árbol depende de los valores de un vector aleatorio y que tiene la misma distribución para todos los árboles predictores incluidos en el algoritmo. Se construyen diferentes árboles predictores con diferentes subconjuntos de datos para después obtener una predicción basada en las predicciones hechas por diferentes árboles [41]. Se utilizó un número de 100 estimadores para entrenar el algoritmo, pues después de experimentar con diferentes números de estimadores, aumentar a más de 100 estimadores no produjo cambios significativos en los resultados.

4.5.3. Multilayer Perceptron

Los Multilayer Perceptron son un tipo de redes neuronales que se utilizan en problemas tanto de clasificación como de regresión. Este tipo de redes neuronales se caracterizan por tener al menos 3 capas de nodos: una capa de entrada, una o más capas intermedias, y una capa de salida. A excepción de los nodos de la capa de entrada, cada nodo tiene asociada una función de activación no lineal [42].

Los parámetros que se utilizaron para afinar las redes neuronales utilizadas en este trabajo fueron seleccionados de acuerdo a las heurísticas recomendadas en el trabajo [43]. Además, se experimentaron con diferentes valores para los parámetros con el fin obtener aquellos que mejoran el desempeño de los modelos desarrollados con este algoritmo. Los parámetros resultantes de este proceso de selección de parámetros se describen en la figura 4.9. Los parámetros no mencionados en la figura 4.9 toman los valores provistos por defecto por el módulo scikit-learn.

Parámetro	**Valor**
Función de activación	Función sigmoidal tanh
Penalidad L2	0.002
Número de capas intermedias	1
Nodos en cada capa intermedia	10
Solucionador	Descenso de gradiente estocástico

FIGURA 4.9: Parámetros utilizados para el Multilayer Perceptron

La arquitectura de los Multilayer Perceptron utilizados para este trabajo consiste en la siguiente: un número de nodos equivalente al número de características en el subconjunto de características seleccionado para cada experimento; 10 nodos en la singular capa intermedia; y 1 nodo de salida, el cual representa en su modo activado una predicción de aumento de precio, y una predicción de decremento de precio en su modo no activado.

Capítulo 5

Metodolgía de Evaluación

Se entrenaron un total de 576 modelos. Cada uno de estos modelos fue el resultado de un experimento distinguido por la criptomoneda evaluada, el conjunto de datos sintetizado utilizado, el subconjunto de características con las que se entrenó el modelo, y el algoritmo utilizado para entrenarlo. Es decir, cada uno de los 12 subconjuntos de características obtenidos de cada uno de los 4 diferentes conjuntos de datos sintetizados fue utilizado para entrenar un modelo con cada uno de los 3 algoritmos de Machine Learning, lo que resulta en total de 144 modelos para comparar por cada una de las 4 criptomonedas.

Esta variedad de experimentos nos permite observar aquellos algoritmos que tienen un mejor desempeño para la tarea y aquellas características que resultan mejores predictoras de el movimiento en el mercado para cada una de las criptomonedas.

5.1. Validación Cruzada

Para garantizar que el desempeño de los modelos obtenidos no es dependiente de el conjunto de datos con el que se entrenó y fue puesto a prueba, en cada experimento se entrenaron 5 modelos utilizando un subconjunto de datos diferente para realizar el entrenamiento y un subconjunto diferente de datos de prueba.

Debido a la naturaleza de series temporales de la tarea, el subconjunto de datos de prueba siempre fue cronológicamente posterior al conjunto de datos utilizado para entrenamiento. La figura 5.1 ilustra la metodología seguida para realizar la validación cruzada en cada experimento.

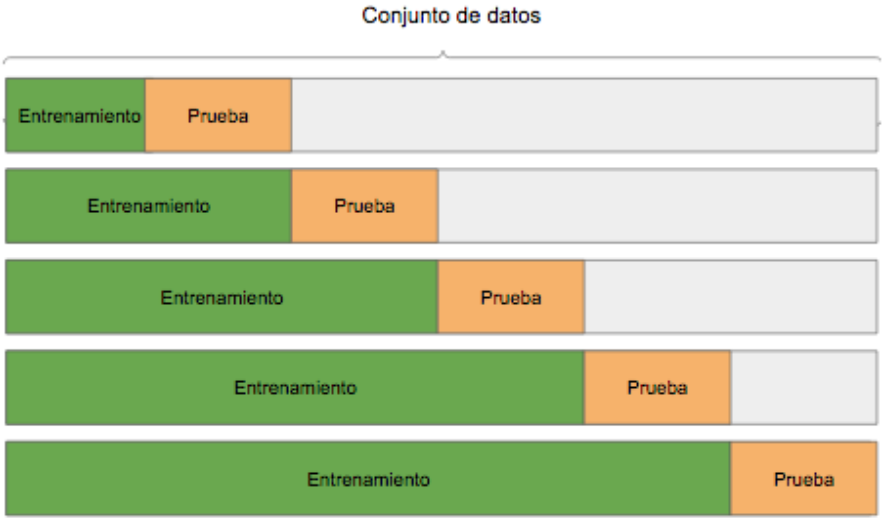


FIGURA 5.1: Validación cruzada

Esta forma de validación cruzada nos permite observar el desempeño de los modelos conforme aumenta el número de datos que se usan para entrenarlos, a costo de que en las primeras pruebas se sacrifica una gran porción de los datos.

5.2. Métricas

Las métricas utilizadas para la evaluación del desempeño de cada modelo se muestran en la figura 5.2.

Métrica	**Descripción**
Precisión	$\frac{Positivos\ verdaderos}{Positivos\ verdaderos + Falsos\ positivos}$
Exactitud	$\frac{Positivos\ verdaderos + Negativos\ verdader}{Total\ de\ predicciones}$
Valor F	$\frac{Precisión \bullet Exhaustividad}{Precisión + Exhaustividad}$
Exhaustividad	$\frac{Positivos\ verdaderos}{Positivos\ verdaderos + Falsos\ negativos}$

FIGURA 5.2: Métricas para evaluar el desempeño de un modelo clasificador

Estas métricas son utilizadas con frecuencia en la literatura para evaluar modelos clasificadores y por tanto fueron utilizadas en este trabajo. La validación cruzada permite obtener con un intervalo de confianza del 95 % el puntaje que cada modelo logra para cada una de las métricas.

Capítulo 6

Resultados

Debido a la cantidad de experimentos que se llevaron a cabo, sólo los experimentos que dieron los modelos que son capaces de vencer al modelo aleatorio con certeza para cada criptomoneda son presentados. Los resultados en su completa extensión pueden encontrarse en el Apéndice A del presente documento. Además de la descripción de los modelos, se incluye la métrica de exactitud lograda en el trabajo [1] como referencia a lo logrado en el estado del arte.

6.1. Bitcoin

Para la criptomoneda Bitcoin, solamente uno de los modelos tuvo un mejor desempeño que el modelo aleatorio. Este modelo se encuentra descrito en la figura 6.1.

Algoritmo utilizado	Random Forest
Conjunto de datos utilizado	Filtrado y aumentado
Subconjunto de características	Promedio de sentimiento en los comentarios positivos
Exactitud	0.55 (+/- 0.03)
Precisión	0.55 (+/- 0.03)
Exhaustividad	0.55 (+/- 0.03)
Valor F	0.55 (+/- 0.03)
Exactitud de referencia	0.76 +/- 0.03

FIGURA 6.1: Descripción y resultados del modelo para Bitcoin

6.2. Ripple

Para la criptomoneda Ripple, fueron tres los modelos que lograron vencer al modelo aleatorio con certeza, estos modelos se presentan en las figuras 6.2, 6.3 y 6.4, respectivamente

Algoritmo utilizado	Random Forest
Conjunto de datos utilizado	Filtrado y sin aumentar
Subconjunto de características	Promedio de sentimiento en los comentarios
Exactitud	0.56 (+/- 0.06)
Precisión	0.56 (+/- 0.06)
Exhaustividad	0.56 (+/- 0.06)
Valor F	0.56 (+/- 0.06)
Exactitud de referencia	0.64 (+/- 0.04)

FIGURA 6.2: Descripción y resultados del modelo para Ripple

Algoritmo utilizado	Random Forest
Conjunto de datos utilizado	N/A
Subconjunto de características	Promedio de sentimiento en las noticias
Exactitud	0.58 (+/- 0.08)
Precisión	0.59 (+/- 0.08)
Exhaustividad	0.58 (+/- 0.08)
Valor F	0.56 (+/- 0.09)
Exactitud de referencia	0.64 (+/- 0.04)

FIGURA 6.3: Descripción y resultados del modelo para Ripple

Algoritmo utilizado	Random Forest
Conjunto de datos utilizado	Filtrado y aumentado
Subconjunto de características	Promedio de sentimiento en los comentarios, promedio de sentimiento en las noticias
Exactitud	0.59 (+/- 0.06)
Precisión	0.59 (+/- 0.06)
Exhaustividad	0.59 (+/- 0.06)
Valor F	0.59 (+/- 0.06)
Exactitud de referencia	0.64 (+/- 0.04)

FIGURA 6.4: Descripción y resultados del modelo para Ripple

6.3. Litecoin

Para la criptomoneda Ripple, fueron dos los modelos que lograron vencer al modelo aleatorio con certeza, estos modelos se presentan en las figuras 6.5 y 6.6, respectivamente.

Algoritmo utilizado	Random Forest
Conjunto de datos utilizado	Filtrado y aumentado
Subconjunto de características	Promedio de sentimiento en los comentarios positivos
Exactitud	0.55 (+/- 0.04)
Precisión	0.55 (+/- 0.04)
Exhaustividad	0.55 (+/- 0.03)
Valor F	0.54 (+/- 0.03)
Exactitud de referencia	0.66 (+/- 0.04)

FIGURA 6.5: Descripción y resultados del modelo para Litecoin

Algoritmo utilizado	Random Forest
Conjunto de datos utilizado	Filtrado y sin aumentar
Subconjunto de características	Promedio de sentimiento en los comentarios positivos y promedio de sentimiento en noticias
Exactitud	0.58 (+/- 0.08)
Precisión	0.58 (+/- 0.08)
Exhaustividad	0.58 (+/- 0.08)
Valor F	0.57 (+/- 0.08)
Exactitud de referencia	0.66 (+/- 0.04)

FIGURA 6.6: Descripción y resultados del modelo para Litecoin

6.4. Ethereum

En el caso de la criptomoneda Ethereum, al igual que con la criptomoneda Bitcoin, solamente uno de los modelos logró tener un mejor desempeño que el del modelo aleatorio.Este modelo se presenta en la figura 6.7

Algoritmo utilizado	Random Forest
Conjunto de datos utilizado	Filtrado y sin aumentar
Subconjunto de características	Promedio de sentimiento en los comentarios positivos
Exactitud	0.59 (+/- 0.07)
Precisión	0.59 (+/- 0.08)
Exhaustividad	0.59 (+/- 0.08)
Valor F	0.59 (+/- 0.07)
Exactitud de referencia	0.44 (+/- 0.02)

FIGURA 6.7: Descripción y resultados del modelo para Ethereum

Capítulo 7

Discusión y Conclusiones

Para las criptomonedas Bitcoin,Litecoin y Ripple no se lograron modelos que igualaran el desempeño obtenido en trabajos previos. Por otro lado, para la criptomoneda Ethereum se logró un modelo que vence a los modelos obtenidos con anterioridad en la literatura.

Random Forest resultó ser el mejor algoritmo para realizar modelos clasificadores, pues todos los modelos que vencen al aleatorio fueron logrados utilizando este algoritmo. Con ningún otro algoritmo se consiguieron modelos que vencen con certeza al aleatorio, en cambio, estos modelos presentaron un desempeño bajo y una desviación estándar alta, lo que significa que los modelos variaron en gran medida al entrenarlos con diferentes subconjuntos de datos. Esta observación reafirma lo que se ha observado en otros estudios tanto del mercado de divisas y valores como el mercado de criptomonedas, en donde los Random Forests se posicionan como uno de los algoritmos que mejor encajan para esta tarea.

Los conjunto de datos sintetizados limitados a incluir solamente las publicaciones enfocadas a discusión fueron los conjuntos de datos con los que se desarrollaron todos los modelos que vencen al aleatorio. Esta observación nos permite concluir que las publicaciones no orientadas a discusión actúan como ruido y deben excluirse para el análisis sentimental en trabajos futuros.

Los datos sentimentales de noticias no fueron suficientes por sí solos para obtener un buen desempeño, sin embargo, en dos de las cuatro criptomonedas los modelos con el mejor desempeño incluyen el sentimiento en noticias, por lo que los autores sugieren la inclusión de más fuentes de noticias en trabajos futuros.

Por otro lado, y contrario a los trabajos realizados previos en el área, la inclusión de los datos históricos del mercado de valor de cierre y de apertura, así como el valor máximo y mínimo, resultaron en modelos con un bajo desempeño. Este fenómeno se atribuye a la alta volatilidad que presentaron los valores de las diferentes criptomonedas durante el final del período de tiempo analizado.

Una observación sobresaliente es que el promedio de sentimiento en comentarios positivos de Reddit resultó ser una característica que predice de mejor forma los movimientos del mercado de criptomonedas que el promedio de sentimiento en todos los comentarios. El hecho de que qué tan positiva se muestra la gente en lugar de cuál es el sentimiento general sea un indicador más confiable del movimiento del mercado sugiere que las opiniones dentro de la red social Reddit tienden a ser optimistas a lo largo de los diferentes foros dedicados a las criptomonedas.

Por los resultados obtenidos en esta investigación, y a pesar de haber obtenido un modelo para cada criptomoneda que vence con certeza al aleatorio, concluimos que la red social Reddit no es una fuente de información que se deba preferir a la red social Twitter cuando se quiera trabajar con el análisis de sentimiento para predecir el movimiento de precio en el mercado de criptomonedas.

Una posible limitante de esta investigación se encuentra en la forma en la que se pre procesaron los datos de la red social Reddit. Una forma más acertada de ponderar el sentimiento mostrado en los comentarios y publicaciones es representar estos datos con una estructura de datos de árbol, pues un comentario clasificado como negativo puede ser respondido con comentarios positivos. Mientras que en este trabajo se trataron como un comentario negativo y muchos positivos, es posible que en realidad se traten de muchos comentarios negativos. Tratar de esta forma los datos reduciría sustancialmente el número de puntos de datos en pro de tener promedio de sentimiento más aproximado a la realidad, por lo que trabajos futuros deben explorar esta alternativa.

Tras observar que utilizar solamente las publicaciones orientadas a discusión dentro de los diferentes foros de criptomonedas mejoran significativamente el desempeño de los modelos, los autores sugieren explorar los algoritmos de Inteligencia Artificial de Inteligencia de Enjambre en trabajos futuros en el tema que involucren la red social Reddit.

Este trabajo, por lo tanto, ofrece una visión más clara de las formas de proceder en los trabajos futuros que pretendan realizar estrategias de inversión serias en el mercado de criptomonedas

incluyendo el sentimiento manifestado en la red social Reddit.

Apéndice A

Lista de Resultados

Resultados obtenidos para cada uno de los 576 modelos desarrollados. Los resultados se dividen por criptomoneda y para cada modelo se muestra la exactitud obtenida.

Cada una de las últimas cuatro columnas enlista los resultados obtenidos utilizando los diferentes conjuntos de datos sintetizados, mientras que cada fila enlista los resultados obtenidos para cada subconjunto de características. RF representa Random Forest, SVM Support Vector Machine, y MLP Multilayer Perceptron.

Bitcoin

Subconjunto de características	Algoritmo	Sin filtrar y sin aumentar	Sin filtrar y aumentado	Filtrado y sin aumentar	Filtrado y aumentado
Número de comentarios totales	RF	0.45 (+/- 0.10)	0.50 (+/- 0.15)	0.50 (+/- 0.12)	0.44 (+/- 0.15)
	SVM	0.53 (+/- 0.12)	0.49 (+/- 0.10)	0.49 (+/- 0.07)	0.50 (+/- 0.07)
	MLP	0.51 (+/- 0.02)	0.51 (+/- 0.02)	0.50 (+/- 0.11)	0.51 (+/- 0.11
Promedio de sentimiento en todos los comentarios	RF	0.45 (+/- 0.16)	0.53 (+/- 0.05)	0.50 (+/- 0.09)	0.49 (+/- 0.09)
	SVM	0.50 (+/- 0.19)	0.48 (+/- 0.09)	0.50 (+/- 0.20)	0.51 (+/- 0.15)
	MLP	0.51 (+/- 0.02)	0.53 (+/- 0.05)	0.51 (+/- 0.02)	0.51 (+/- 0.02)
Número de comentarios totales y promedio de sentimiento en todos los comentarios	RF	0.45 (+/- 0.12)	0.44 (+/- 0.13)	0.47 (+/- 0.18)	0.48 (+/- 0.06)
	SVM	0.46 (+/- 0.15)	0.45 (+/- 0.19)	0.46 (+/- 0.13)	0.44 (+/- 0.05)
	MLP	0.54 (+/- 0.07)	0.51 (+/- 0.09)	0.52 (+/- 0.15)	0.51 (+/- 0.13)
Promedio de sentimiento en los comentarios positivos	RF	0.39 (+/- 0.07)	0.47 (+/- 0.13)	0.49 (+/- 0.06)	0.55 (+/- 0.03)
	SVM	0.41 (+/- 0.11)	0.51 (+/- 0.11)	0.45 (+/- 0.08)	0.42 (+/- 0.08)
	MLP	0.51 (+/- 0.03)	0.52 (+/- 0.07)	0.49 (+/- 0.04)	0.51 (+/- 0.02)
Porcentaje de comentarios positivos y porcentaje de comentarios negativos	RF	0.54 (+/- 0.10)	0.52 (+/- 0.12)	0.53 (+/- 0.10)	0.48 (+/- 0.05)
	SVM	0.50 (+/- 0.17)	0.50 (+/- 0.07)	0.49 (+/- 0.12)	0.54 (+/- 0.08)
	MLP	0.48 (+/- 0.15)	0.49 (+/- 0.07)	0.48 (+/- 0.20)	0.52 (+/- 0.16)
Promedio de sentimiento en las noticias	RF	0.45 (+/- 0.07)	0.45 (+/- 0.07)	0.44 (+/- 0.05)	0.44 (+/- 0.05)
	SVM	0.53 (+/- 0.09)	0.53 (+/- 0.09)	0.53 (+/- 0.11)	0.53 (+/- 0.11)
	MLP	0.51 (+/- 0.02)	0.51 (+/- 0.02)	0.51 (+/- 0.02)	0.51 (+/- 0.02)
Promedio de sentimiento en los comentarios, promedio de sentimiento en las noticias	RF	0.50 (+/- 0.11)	0.49 (+/- 0.11)	0.52 (+/- 0.10)	0.52 (+/- 0.13)
	SVM	0.55 (+/- 0.14)	0.52 (+/- 0.12)	0.58 (+/- 0.23)	0.55 (+/- 0.18)
	MLP	0.55 (+/- 0.15)	0.54 (+/- 0.16)	0.53 (+/- 0.12)	0.55 (+/- 0.14)

Número de comentarios, promedio de sentimiento en las noticias, promedio de sentimiento en los comentarios	RF	0.46 (+/- 0.26)	0.44 (+/- 0.13)	0.51 (+/- 0.14)	0.54 (+/- 0.11)
	SVM	0.57 (+/- 0.16)	0.49 (+/- 0.23)	0.51 (+/- 0.20)	0.51 (+/- 0.18)
	MLP	0.52 (+/- 0.04)	0.50 (+/- 0.03)	0.52 (+/- 0.08)	0.54 (+/- 0.07)
Promedio de sentimiento en los comentarios positivos, promedio de sentimiento en las noticias	RF	0.50 (+/- 0.11)	.47 (+/- 0.11)	0.52 (+/- 0.14)	0.48 (+/- 0.06)
	SVM	0.52 (+/- 0.11)	0.52 (+/- 0.09)	0.55 (+/- 0.14)	0.57 (+/- 0.12)
	MLP	0.53 (+/- 0.09)	0.53 (+/- 0.06)	0.53 (+/- 0.17)	0.52 (+/- 0.13)
Porcentaje de comentarios positivos, porcentaje de comentarios negativos, promedio de	RF	0.51 (+/- 0.02)	0.54 (+/- 0.14)	0.50 (+/- 0.16)	0.51 (+/- 0.07)
	SVM	0.52 (+/- 0.09)	0.50 (+/- 0.11)	0.52 (+/- 0.20)	0.54 (+/- 0.11)
	MLP	0.54 (+/- 0.05)	0.50 (+/- 0.10)	0.54 (+/- 0.12)	0.57 (+/- 0.09)

Ripple

Subconjunto de características	Algoritmo	Sin filtrar y sin aumentar	Sin filtrar y aumentado	Filtrado y sin aumentar	Filtrado y aumentado
Número de comentarios totales	RF	0.53 (+/- 0.13)	0.51 (+/- 0.17)	0.60 (+/- 0.15)	0.48 (+/- 0.12)
	SVM	0.58 (+/- 0.16)	0.55 (+/- 0.09)	0.52 (+/- 0.05)	0.57 (+/- 0.15)
	MLP	0.50 (+/- 0.01)	0.50 (+/- 0.01)	0.50 (+/- 0.02)	0.50 (+/- 0.02)
Promedio de sentimiento en todos los comentarios	RF	0.51 (+/- 0.19)	0.48 (+/- 0.06)	0.56 (+/- 0.06)	0.51 (+/- 0.15)
	SVM	0.46 (+/- 0.05)	0.50 (+/- 0.08)	0.52 (+/- 0.05)	0.59 (+/- 0.12)
	MLP	0.50 (+/- 0.01)	0.49 (+/- 0.05)	0.50 (+/- 0.04)	0.53 (+/- 0.07)
Número de comentarios totales y promedio de sentimiento en todos los comentarios	RF	0.45 (+/- 0.10)	0.53 (+/- 0.09)	0.56 (+/- 0.13)	0.55 (+/- 0.13)
	SVM	0.51 (+/- 0.02)	0.51 (+/- 0.16)	0.51 (+/- 0.10)	0.59 (+/- 0.13)
	MLP	0.52 (+/- 0.14)	0.53 (+/- 0.12)	0.51 (+/- 0.09)	0.50 (+/- 0.08)
Promedio de sentimiento en los comentarios positivos	RF	0.45 (+/- 0.08)	0.45 (+/- 0.13)	0.51 (+/- 0.07)	0.46 (+/- 0.07)
	SVM	0.42 (+/- 0.15)	0.49 (+/- 0.08)	0.47 (+/- 0.08)	0.46 (+/- 0.11)
	MLP	0.48 (+/- 0.06)	0.53 (+/- 0.12)	0.49 (+/- 0.10)	0.49 (+/- 0.03)
Porcentaje de comentarios positivos y porcentaje de comentarios negativos	RF	0.57 (+/- 0.11)	0.49 (+/- 0.17)	0.51 (+/- 0.15)	0.48 (+/- 0.12)
	SVM	0.51 (+/- 0.27)	0.47 (+/- 0.11)	0.48 (+/- 0.09)	0.54 (+/- 0.15)
	MLP	0.50 (+/- 0.14)	0.45 (+/- 0.05)	0.47 (+/- 0.15)	0.44 (+/- 0.10)
Promedio de sentimiento en las noticias	RF	0.59 (+/- 0.09)	0.59 (+/- 0.09)	0.58 (+/- 0.08)	0.58 (+/- 0.08)
	SVM	0.57 (+/- 0.10)	0.57 (+/- 0.10)	0.57 (+/- 0.13)	0.57 (+/- 0.13)
	MLP	0.48 (+/- 0.06)	0.48 (+/- 0.06)	0.48 (+/- 0.06)	0.48 (+/- 0.06)
Promedio de sentimiento en los comentarios, promedio de sentimiento en las noticias	RF	0.58 (+/- 0.10)	0.55 (+/- 0.12)	0.55 (+/- 0.17)	0.59 (+/- 0.06)
	SVM	0.51 (+/- 0.06)	0.56 (+/- 0.10)	0.50 (+/- 0.05)	0.51 (+/- 0.07)
	MLP	0.52 (+/- 0.07)	0.51 (+/- 0.10)	0.51 (+/- 0.11)	0.48 (+/- 0.09)

Número de comentarios, promedio de sentimiento en las noticias, promedio de sentimiento en los comentarios	RF	0.52 (+/- 0.10)	0.51 (+/- 0.13)	0.55 (+/- 0.11)	0.60 (+/- 0.14)
	SVM	0.51 (+/- 0.09)	0.52 (+/- 0.07)	0.50 (+/- 0.07)	0.56 (+/- 0.09)
	MLP	0.45 (+/- 0.05)	0.45 (+/- 0.04)	0.47 (+/- 0.09)	0.48 (+/- 0.10)
Promedio de sentimiento en los comentarios positivos, promedio de sentimiento en las noticias	RF	0.54 (+/- 0.07)	0.52 (+/- 0.07)	0.48 (+/- 0.13)	0.52 (+/- 0.08)
	SVM	0.53 (+/- 0.12)	0.55 (+/- 0.11)	0.51 (+/- 0.07)	0.54 (+/- 0.13)
	MLP	0.49 (+/- 0.06)	0.51 (+/- 0.07)	0.52 (+/- 0.13)	0.54 (+/- 0.06)
Porcentaje de comentarios positivos, porcentaje de comentarios negativos, promedio de sentimiento en las noticias	RF	0.55 (+/- 0.16)	0.44 (+/- 0.14)	0.49 (+/- 0.12)	0.54 (+/- 0.13)
	SVM	0.56 (+/- 0.21)	0.47 (+/- 0.11)	0.50 (+/- 0.11)	0.52 (+/- 0.08)
	MLP	0.52 (+/- 0.08)	0.54 (+/- 0.11)	0.50 (+/- 0.04)	0.51 (+/- 0.06)

Litecoin

Subconjunto de características	Algoritmo	Sin filtrar y sin aumentar	Sin filtrar y aumentado	Filtrado y sin aumentar	Filtrado y aumentado
Número de comentarios totales	RF	0.52 (+/- 0.13)	0.49 (+/- 0.10)	0.45 (+/- 0.09)	0.49 (+/- 0.11)
	SVM	0.46 (+/- 0.07)	0.48 (+/- 0.07)	0.44 (+/- 0.07)	0.47 (+/- 0.09)
	MLP	0.51 (+/- 0.05)	0.51 (+/- 0.04)	0.52 (+/- 0.05)	0.52 (+/- 0.06)
Promedio de sentimiento en todos los comentarios	RF	0.54 (+/- 0.16)	0.48 (+/- 0.11)	0.45 (+/- 0.12)	0.51 (+/- 0.19)
	SVM	0.42 (+/- 0.03)	0.48 (+/- 0.07)	0.45 (+/- 0.14)	0.49 (+/- 0.08)
	MLP	0.49 (+/- 0.08)	0.49 (+/- 0.07)	0.49 (+/- 0.07)	0.51 (+/- 0.07)
Número de comentarios totales y promedio de sentimiento en todos los comentarios	RF	0.50 (+/- 0.18)	0.53 (+/- 0.06)	0.49 (+/- 0.12)	0.46 (+/- 0.08)
	SVM	0.44 (+/- 0.12)	0.46 (+/- 0.09)	0.47 (+/- 0.10)	0.47 (+/- 0.06)
	MLP	0.48 (+/- 0.19)	0.50 (+/- 0.14)	0.49 (+/- 0.16)	0.50 (+/- 0.11)
Promedio de sentimiento en los comentarios positivos	RF	0.48 (+/- 0.20)	0.52 (+/- 0.17)	0.58 (+/- 0.13)	0.55 (+/- 0.04)
	SVM	0.48 (+/- 0.10)	0.48 (+/- 0.13)	0.51 (+/- 0.11)	0.52 (+/- 0.14)
	MLP	0.52 (+/- 0.09)	0.54 (+/- 0.07)	0.50 (+/- 0.03)	0.51 (+/- 0.10)
Porcentaje de comentarios positivos y porcentaje de comentarios negativos	RF	0.47 (+/- 0.21)	0.51 (+/- 0.08)	0.44 (+/- 0.15)	0.52 (+/- 0.14)
	SVM	0.49 (+/- 0.05)	0.53 (+/- 0.04)	0.53 (+/- 0.11)	0.47 (+/- 0.13)
	MLP	0.52 (+/- 0.20)	0.53 (+/- 0.22)	0.53 (+/- 0.11)	0.51 (+/- 0.09)
Promedio de sentimiento en las noticias	RF	0.52 (+/- 0.10)	0.52 (+/- 0.10)	0.49 (+/- 0.11)	0.49 (+/- 0.12)
	SVM	0.49 (+/- 0.16)	0.49 (+/- 0.16)	0.50 (+/- 0.17)	0.50 (+/- 0.17)
	MLP	0.51 (+/- 0.02)	0.51 (+/- 0.02)	0.51 (+/- 0.01)	0.51 (+/- 0.01)
Promedio de sentimiento en los comentarios, promedio de sentimiento en las noticias	RF	0.51 (+/- 0.20)	0.51 (+/- 0.08)	0.46 (+/- 0.15)	0.54 (+/- 0.13)
	SVM	0.46 (+/- 0.10)	0.53 (+/- 0.08)	0.51 (+/- 0.16)	0.53 (+/- 0.18)
	MLP	0.51 (+/- 0.07)	0.50 (+/- 0.08)	0.52 (+/- 0.13)	0.49 (+/- 0.09)

Número de comentarios, promedio de sentimiento en las noticias, promedio de sentimiento en los comentarios	RF	0.43 (+/- 0.12)	0.46 (+/- 0.14)	0.49 (+/- 0.09)	0.47 (+/- 0.14)
	SVM	0.45 (+/- 0.15)	0.51 (+/- 0.08)	0.47 (+/- 0.15)	0.47 (+/- 0.19)
	MLP	0.51 (+/- 0.07)	0.51 (+/- 0.04)	0.50 (+/- 0.06)	0.51 (+/- 0.04)
Promedio de sentimiento en los comentarios positivos, promedio de sentimiento en las noticias	RF	0.53 (+/- 0.23)	0.56 (+/- 0.07)	0.58 (+/- 0.08)	0.52 (+/- 0.15)
	SVM	0.48 (+/- 0.14)	0.51 (+/- 0.11)	0.49 (+/- 0.19)	0.50 (+/- 0.19)
	MLP	0.50 (+/- 0.10)	0.47 (+/- 0.13)	0.53 (+/- 0.07)	0.47 (+/- 0.11)
Porcentaje de comentarios positivos, porcentaje de comentarios negativos, promedio de sentimiento en las noticias	RF	0.48 (+/- 0.06)	0.57 (+/- 0.11)	0.50 (+/- 0.15)	0.48 (+/- 0.07)
	SVM	0.48 (+/- 0.14)	0.53 (+/- 0.18)	0.54 (+/- 0.14)	0.51 (+/- 0.14)
	MLP	0.50 (+/- 0.11)	0.49 (+/- 0.14)	0.51 (+/- 0.15)	0.51 (+/- 0.15)

Ethereum

Subconjunto de características	Algoritmo	Sin filtrar y sin aumentar	Sin filtrar y aumentado	Filtrado y sin aumentar	Filtrado y aumentado
Número de comentarios totales	RF	0.48 (+/- 0.10)	0.47 (+/- 0.10)	0.47 (+/- 0.10)	0.47 (+/- 0.20)
	SVM	0.45 (+/- 0.12)	0.48 (+/- 0.10)	0.48 (+/- 0.08)	0.46 (+/- 0.09)
	MLP	0.47 (+/- 0.10)	0.49 (+/- 0.07)	0.47 (+/- 0.15)	0.50 (+/- 0.11)
Promedio de sentimiento en todos los comentarios	RF	0.43 (+/- 0.18)	0.49 (+/- 0.15)	0.48 (+/- 0.15)	0.54 (+/- 0.17)
	SVM	0.43 (+/- 0.11)	0.50 (+/- 0.11)	0.53 (+/- 0.21)	0.51 (+/- 0.18)
	MLP	0.48 (+/- 0.06)	0.53 (+/- 0.16)	0.50 (+/- 0.18)	0.52 (+/- 0.16)
Número de comentarios totales y promedio de sentimiento en todos los comentarios	RF	0.48 (+/- 0.06)	0.53 (+/- 0.12)	0.46 (+/- 0.12)	0.49 (+/- 0.19)
	SVM	0.44 (+/- 0.12)	0.47 (+/- 0.10)	0.50 (+/- 0.14)	0.53 (+/- 0.10)
	MLP	0.47 (+/- 0.11)	0.40 (+/- 0.11)	0.43 (+/- 0.11)	0.45 (+/- 0.15)
Promedio de sentimiento en los comentarios positivos	RF	0.49 (+/- 0.18)	0.45 (+/- 0.08)	0.59 (+/- 0.07)	0.51 (+/- 0.05)
	SVM	0.48 (+/- 0.05)	0.54 (+/- 0.09)	0.45 (+/- 0.10)	0.52 (+/- 0.10)
	MLP	0.48 (+/- 0.04)	0.50 (+/- 0.06)	0.51 (+/- 0.20)	0.51 (+/- 0.10)
Porcentaje de comentarios positivos y porcentaje de comentarios negativos	RF	0.52 (+/- 0.12)	0.51 (+/- 0.11)	0.43 (+/- 0.11)	0.54 (+/- 0.18)
	SVM	0.46 (+/- 0.12)	0.57 (+/- 0.12)	0.49 (+/- 0.17)	0.54 (+/- 0.11)
	MLP	0.48 (+/- 0.12)	0.45 (+/- 0.09)	0.48 (+/- 0.15)	0.49 (+/- 0.05)
Promedio de sentimiento en las noticias	RF	0.48 (+/- 0.19)	0.48 (+/- 0.19)	0.48 (+/- 0.21)	0.48 (+/- 0.21)
	SVM	0.45 (+/- 0.15)	0.45 (+/- 0.15)	0.46 (+/- 0.13)	0.46 (+/- 0.13)
	MLP	0.45 (+/- 0.09)	0.45 (+/- 0.09)	0.46 (+/- 0.08)	0.46 (+/- 0.08)
Promedio de sentimiento en los comentarios, promedio de sentimiento en las noticias	RF	0.43 (+/- 0.06)	0.43 (+/- 0.13)	0.46 (+/- 0.15)	0.49 (+/- 0.14)
	SVM	0.45 (+/- 0.09)	0.47 (+/- 0.09)	0.49 (+/- 0.15)	0.54 (+/- 0.16)
	MLP	0.48 (+/- 0.18)	0.49 (+/- 0.13)	0.46 (+/- 0.11)	0.49 (+/- 0.17)

Número de comentarios, promedio de sentimiento en las noticias, promedio de sentimiento en los comentarios	RF	0.45 (+/- 0.04)	0.47 (+/- 0.20)	0.44 (+/- 0.19)	0.41 (+/- 0.12)
	SVM	0.48 (+/- 0.13)	0.46 (+/- 0.13)	0.50 (+/- 0.16)	0.54 (+/- 0.12)
	MLP	0.46 (+/- 0.03)	0.48 (+/- 0.05)	0.46 (+/- 0.06)	0.48 (+/- 0.06)
Promedio de sentimiento en los comentarios positivos, promedio de sentimiento en las noticias	RF	0.38 (+/- 0.07)	0.43 (+/- 0.09)	0.41 (+/- 0.17)	0.41 (+/- 0.17)
	SVM	0.41 (+/- 0.12)	0.47 (+/- 0.08)	0.47 (+/- 0.08)	0.52 (+/- 0.11)
	MLP	0.53 (+/- 0.16)	0.52 (+/- 0.11)	0.49 (+/- 0.12)	0.52 (+/- 0.17)
Porcentaje de comentarios positivos, porcentaje de comentarios negativos, promedio de sentimiento en las noticias	RF	0.45 (+/- 0.13)	0.46 (+/- 0.13)	0.45 (+/- 0.18)	0.42 (+/- 0.11)
	SVM	0.46 (+/- 0.12)	0.50 (+/- 0.15)	0.51 (+/- 0.15)	0.52 (+/- 0.12)
	MLP	0.49 (+/- 0.14)	0.48 (+/- 0.10)	0.47 (+/- 0.10)	0.48 (+/- 0.10)

Bibliografía

[1] F. Valencia, A. Gómez-Espinosa, y B. Valdés-Aguirre, "Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning," *Entropy* , vol. 21, no. 6, p. 589, Jun. 2019.

[2] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System." Nov. 20, 2019, Accessed: May 07, 2020. [Online]. Disponible: https://git.dhimmel.com/bitcoin-whitepaper/.

[3] M. Alvarez, "A Comparative Analysis of Cryptocurrency Regulation in the United States, Nigeria, and China: The Potential Influence of Illicit Activities on Regulatory Evolution," *ILSA J. Int'l Comp. L.*, vol. 25, p. 33, 2018.

[4] Global Legal Research Directorate Staff, "Regulation of Cryptocurrency Around the World," Jun. 2018. https://www.loc.gov/law/help/cryptocurrency/world-survey.php (visitado en Jun. 02, 2020).

[5] G. Hileman y M. Rauchs, "Global cryptocurrency benchmarking study. 2017." 2017.

[6] A. ElBahrawy, L. Alessandretti, A. Kandler, R. Pastor-Satorras, y A. Baronchelli, "Evolutionary dynamics of the cryptocurrency market," *R Soc Open Sci*, vol. 4, no. 11, p. 170623, Nov. 2017.

[7] "Cryptocurrency Market Capitalizations — CoinMarketCap," *CoinMarketCap*. https://coinmarketcap.com/1/ (visitado en May 07, 2020).

[8] J. Liang, L. Li, W. Chen, y D. Zeng, "Towards an Understanding of Cryptocurrency: A Comparative Analysis of Cryptocurrency, Foreign Exchange, and Stock," en *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, Shenzhen, China, Jul. 2019, pp. 137–139.

[9] D. Stosic, D. Stosic, T. B. Ludermir, y T. Stosic, "Collective behavior of cryptocurrency price changes," Physica A: Statistical Mechanics and its Applications, vol. 507, pp. 499–509, Oct. 2018.

[10] S. Yi, Z. Xu, y G.-J. Wang, "Volatility connectedness in the cryptocurrency market: Is Bitcoin a dominant cryptocurrency?," *International Review of Financial Analysis*, vol. 60, pp. 98–114, Nov. 2018.

[11] N. Antonakakis, I. Chatziantoniou, y D. Gabauer, "Cryptocurrency market contagion: Market uncertainty, market complexity, and dynamic portfolios," *Journal of International Financial Markets, Institutions and Money*, vol. 61, pp. 37–51, Jul. 2019.

[12] "Website." Aggarwal, Gourang, Vimal Patel, Gaurav Varshney, y Kimberly Oostman. 2019. "Understanding the Social Factors Affecting the Cryptocurrency Market." arXiv [cs.CY]. arXiv. http://arxiv.org/abs/1901.06245. (accessed May 07, 2020).

[13] P. M. Krafft, N. Della Penna, y A. S. Pentland, "An Experimental Study of Cryptocurrency Market Dynamics," en *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Montreal QC, Canada, 2018, pp. 1–13.

[14] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012.

[15] X. Fan, X. Li, F. Du, X. Li, y M. Wei, "Apply word vectors for sentiment analysis of APP reviews," in *2016 3rd International Conference on Systems and Informatics (ICSAI)*, Nov. 2016, pp. 1062–1066.

[16] T. H. Nguyen, K. Shirai, y J. Velcin, "Sentiment analysis on social media for stock movement prediction," *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9603–9611, Dec. 2015.

[17] R. Feldman, "Techniques and applications for sentiment analysis," *Commun. ACM*, vol. 56, no. 4, pp. 82–89, Apr. 2013.

[18] D. Shah, H. Isah, y F. Zulkernine, "Stock Market Analysis: A Review and Taxonomy of Prediction Techniques," *IJFS*, vol. 7, no. 2, p. 26, May 2019.

[19] M. Ballings, D. Van den Poel, N. Hespeels, y R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7046–7056, Nov. 2015.

[20] R. Chowdhury, M. A. Rahman, M. S. Rahman, y M. R. C. Mahdy, "An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning," *Physica A: Statistical Mechanics and its Applications*, vol. 551, p. 124569, Aug. 2020.

[21] A. Regal et al., "Proyección del precio de criptomonedas basado en Tweets empleando LSTM," *Ingeniare. Rev. chil. ing.*, vol. 27, no. 4, pp. 696–706, Dec. 2019.

[22] A.-D. Vo, Q.-P. Nguyen, y C.-Y. Ock, "Sentiment Analysis of News for Effective Cryptocurrency Price Prediction," *International Journal of Knowledge Engineering*, vol. 5, pp. 47–52, Dec. 2019.

[23] C. Lamon, E. Nielsen, y E. Redondo, "Cryptocurrency Price Prediction Using News And Social Media Sentiment," *SMU Data Sci. Rev*, vol. 1, no. 3, pp. 1–22, 2017, Visitado: May 07, 2020. [Online].

[24] K. Wołk, "Advanced social media sentiment analysis for short-term cryptocurrency price prediction," *Expert Syst.*, vol. 37, no. 2, p. 1, Apr. 2020.

[25] J. Abraham, D. Higdon, J. Nelson, y J. Ibarra, "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," *SMU Data Science Review*, vol. 1, no. 3, p. Article 1, 2018, Visitado: May 07, 2020. [Online].

[26] N. A. Hitam, A. R. Ismail, y F. Saeed, "An Optimized Support Vector Machine (SVM) based on Particle Swarm Optimization (PSO) for Cryptocurrency Forecasting," *Procedia Comput. Sci.*, vol. 163, pp. 427–433, 2019.

[27] V. S. Subrahmanian et al., "The DARPA Twitter Bot Challenge," *Computer* , vol. 49, no. 6, pp. 38–46, Jun. 2016.

[28] "API reference index." https://developer.twitter.com/en/docs/api-reference-index (accessed Jun. 02, 2020).

[29] S. Priya, R. Sequeira, J. Chandra, y S. K. Dandapat, "Where should one get news updates: Twitter or Reddit," *Online Social Networks and Media*, vol. 9, pp. 17–29, Jan. 2019.

[30] "reddit.com: api documentation." https://www.reddit.com/dev/api/ (accessed Jun. 02, 2020).

[31] "Bitcoin," Twitter. https://www.twitter.com/Bitcoin/ (accessed Jun. 02, 2020) .

[32] "r/Bitcoin," reddit. https://www.reddit.com/r/Bitcoin/ (accessed Jun. 02, 2020).

[33] "Reddit API Terms of Use," *Google Docs*. https://docs.google.com/forms/d/e/1FAIpQLSezNdDNK1-P8mspSbmtC2r86Ee9ZRbC66u929cG2GX0T9UMyw/viewform?usp=embed_facebook (accessed Jun. 02, 2020).

[34] B. Boe, *praw*. Github.

[35] "Scrapy — A Fast and Powerful Scraping and Web Crawling Framework." https://scrapy.org/ (visitado Jun. 02, 2020).

[36] "The Best Free CryptoCurrency Price and Historical Data API for Developers — CryptoCompare API (trades, news, streaming and toplists also available)," *CryptoCompare*. https://min-api.cryptocompare.com/ (visitado Jun. 02, 2020).

[37] C. J. Hutto y E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in Eighth international AAAI conference on weblogs and social media, 2014, [Online]. Disponible: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109.

[38] T. Loughran y B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," J. Finance, vol. 66, no. 1, pp. 35–65, 2011.

[39] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, Visitado: Jun. 02, 2020. [Online].

[40] C. Cortes y V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[41] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[42] S. Sathyanarayana, "A gentle introduction to backpropagation," *Numer. Insights*, 2014, [Online]. Disponible: https://www.researchgate.net/profile/Shashi_Sathyanarayana/publication/266396438_A_Gentle_Introduction_to_Backpropagation/links/577d124808aeaa6988aba0bc/A-Gentle-Introduction-to-Backpropagation.pdf.

[43] Y. A. LeCun, L. Bottou, G. B. Orr, y K.-R. Müller, "Efficient BackProp," in *Neural Networks: Tricks of the Trade: Second Edition*, G. Montavon, G. B. Orr, y K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48.