Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Análisis de sentimiento de expresiones en Twitter para la predicción del comportamiento del precio del Bitcoin en una organización Fintech

Implementación de técnicas de Text Mining.

AUTOR: CARLOS JOSE LEOTAUD CASTEJON
MENTORA: MG. NATALIA SALABERRY

FEBRERO 2023

Resumen

Las organizaciones *Fintech* tienen entre sus objetivos realizar una distribución eficiente de su capital, invirtiendo en diferentes activos entre los que se encuentran el *Bitcoin*. Para lograr esta meta, conocer las opiniones de los inversores y comprender el efecto que dichas opiniones tienen en el precio del *Bitcoin* resulta de gran importancia. Esto permitiría contribuir al desarrollo de mejores estrategias de adquisición de la moneda digital y conformar estrategias eficientes de inversión.

En este contexto, el tema propuesto a desarrollar en el presente trabajo busca exponer el potencial de la utilización de datos no estructurados en una organización *Fintech* para el diseño eficiente de opciones de inversión. El objetivo buscado es determinar el efecto de las opiniones de los usuarios de *Twitter* en la predicción del precio del *Bitcoin* implementando técnicas de *Text Mining*. Esto constituye un valor agregado para la definición de nuevas estrategias de inversión.

Con el fin de cumplimentar con el objetivo planteado, en primer lugar, se aborda la gestión de datos alternativos en organizaciones *Fintech*. Luego se desarrolla el proceso de obtención de los datos a utilizar, su descripción y se presentan las técnicas a implementar. Posteriormente se implementan estas para clasificar las opiniones de los usuarios. Finalmente, se determina si existe una correlación entre las opiniones recolectadas en la red social *Twitter* y el precio del *Bitcoin*. De esta manera, el uso de datos alternativos constituye un valor agregado para determinar las variaciones en el precio del *Bitcoin*.

Palabras claves: Análisis de sentimientos, *Twitter*, Predicción del precio, *Bitcoin*, *Fintech*.

índice

Introducción

Las organizaciones *Fintech* son un rubro particularmente sensible a los cambios repentinos en la valoración de sus carteras de activos (Vučinić, 2022). Para poder hacer una gestión eficiente de estos, es importante desarrollar herramientas que ayuden a tener previsibilidad sobre los cambios que se pueden producir (Leong & Sung, 2018). El uso de este tipo de mecanismos abre la posibilidad de realizar una gestión defensiva de la inversión realizada y a su vez la implementación de mejores técnicas de distribución del capital por parte de la entidad.

En la actualidad una de las principales fuentes de datos son las redes sociales (Eberendu, 2016). Esto ha sucedido gracias al uso que las organizaciones les dan como canales de comunicación. Estas comunicaciones se pueden dar de diferentes formas. Entre las organizaciones y sus clientes o usuarios de su servicio o entre los miembros que pertenecen a una comunidad (Badea, 2014). A través de este tipo de interacciones, estos espacios generan datos de forma constante y actualizada. Las organizaciones tienen interés en analizar los comentarios que realizan los participantes de las redes sociales en relación con los temas de su interés.

Gracias al uso y tratamiento de los datos obtenidos de las redes sociales, las organizaciones *Fintech* podrán tomar decisiones sustentados en estos. Esto permite así la gestión eficiente del capital invertido en activos como el *Bitcoin*. La aplicación de decisiones de inversión mejor fundamentadas repercutirá en una mayor previsibilidad en el resultado de estas operaciones y en una reducción en las pérdidas totales acumuladas dentro de la organización. De esta forma, se busca responder al siguiente interrogante ¿Cuál es la efectividad del análisis de opiniones de *tweets* en la predicción del comportamiento del precio del *Bitcoin*?

Para responder el interrogante planteado, el objetivo general de este trabajo es aplicar mecanismos de procesamiento y clasificación de datos alternativos en la organización para determinar el efecto de las opiniones de los usuarios de *Twitter* en la predicción del precio del *Bitcoin*. Para resolverlo, en primer lugar, se evalúa la relevancia de los datos alternativos para la toma de decisiones en las organizaciones de tipo *Fintech*. En segundo lugar, se procederá a relevar y procesar los *tweets* para analizar su impacto en el precio del *Bitcoin* mediante el uso de técnicas de *Text Mining y Machine Learning*. Finalmente, en base a los resultados obtenidos se medirá la relación entre las opiniones de los usuarios de *Twitter* y el comportamiento del precio del *Bitcoin*.

Para poder llevar adelante el objetivo planteado, el trabajo se estructura en tres capítulos. En el primer capítulo, se presentará la relevancia de los grandes volúmenes de datos alternativos para la toma de decisiones en organizaciones. Luego se contextualizará dentro de organizaciones *Fintech*. Finalmente, se determinará como una correcta gestión de datos resulta necesaria para la generación de valor agregado en la toma de decisiones

En el segundo capítulo, se realizará el relevamiento de *tweets* y su posterior análisis. Este relevamiento se llevará a cabo extrayendo datos de la red social *Twitter*, específicamente aquellos *tweets* que incluyan la etiqueta *Bitcoin (#Bitcoin)*. Para esto, se usará una API[1] (*Application Programming Interface*) de *Twitter* de uso público y gratuito que se ejecutará sobre un entorno de desarrollo usando el lenguaje *Python*[2]. Mediante el uso de técnicas de *Text Mining* se procesarán los datos obtenidos para posteriormente estructurarlos facilitando así su interpretación y análisis.

Finalmente, en el tercer y último capítulo, se procederá a implementar el modelo de clasificación de los *tweets*. Seguido a esto, se realizará el análisis de los resultados obtenidos al aplicar métodos estadísticos. Como ultimo apartado, se concluirá acerca de la importancia de los datos alternativos para la determinación del precio del *Bitcoin*.

[1] Acerca de API *Twitter*: https://developer.twitter.com/en/docs/twitter-api
[2] Acerca de Python: https://www.python.org/

Capítulo 1: Datos alternativos y Fintech

Las organizaciones financieras en la actualidad se desenvuelven un contexto desafiante y exigente. Debido a esto, se enfrentan constantemente a numerosos obstáculos. Ante este escenario se han visto forzadas a reaccionar de forma rápida con el objetivo de asegurar ventajas competitivas en su mercado. Esta necesidad combinada con la velocidad con que las nuevas tecnologías han sido desarrolladas en las últimas décadas han propiciado un ambiente de constante evolución (Martincevic, Crnjevic, & Klopotan, 2020). En este contexto, han nacido un nuevo tipo de organizaciones financieras especialmente centradas en la aplicación de los avances tecnológicos en el campo de las finanzas. Estas son denominadas *Fintech* (Arner, Barberis, & Buckley, 2015).

Estas organizaciones *Fintech* han encontrado en los grandes volúmenes de datos alternativos y la aplicación de tecnologías emergentes como la *Blockchain* y el *Bitcoin* herramientas para mejorar sus procesos internos. Su implementación también ha ayudado a satisfacer las necesidades de sus clientes. Al mismo tiempo, el uso de los datos y tecnologías crean nuevos desafíos como son la correcta gestión e implementación de estos (Eberendu, 2016).

De esta forma, la gestión de los grandes volúmenes de datos y en particular de los datos alternativos toma un rol principal. Esto se debe a que son elementos centrales que atraviesan a todas las áreas de la organización. De esta manera, resulta especialmente relevante a la hora implementar nuevas tecnologías que ayuden a mejorar la toma de decisiones. No contar con correctos sistemas de gestión de los datos puede causar imprecisiones con sus respectivas consecuencias para el bienestar de la organización (Cleven & Wortmann, 2010).

El objetivo del capitulo es analizar los procesos y metodologías para la gestión de los grandes volúmenes de datos alternativos en el sector *Fintech.* Para esto, en primer lugar, se realizará una aproximación teórica a las organizaciones *Fintech* y el *Bitcoin*. Luego, se planteará una metodología para realizar la gestión de grandes volúmenes de datos. Finalmente, se detallarán los métodos para la gestión de los datos alternativos.

1.1 Las organizaciones Fintech

Fintech es un término que proviene de la unión de dos palabras, finanzas y tecnología. Este nombre, como medio para referirse a la unión de estos dos campos de conocimiento, comenzó a ser utilizado de forma cada vez más extensiva durante el siglo 21 para describir tecnologías que buscan mejorar y automatizar el uso de los servicios financieros. De esta manera, es posible definir *Fintech* como el proceso de digitalización o actualización tecnológica de las soluciones financieras (Arner, Barberis, & Buckley, 2015).

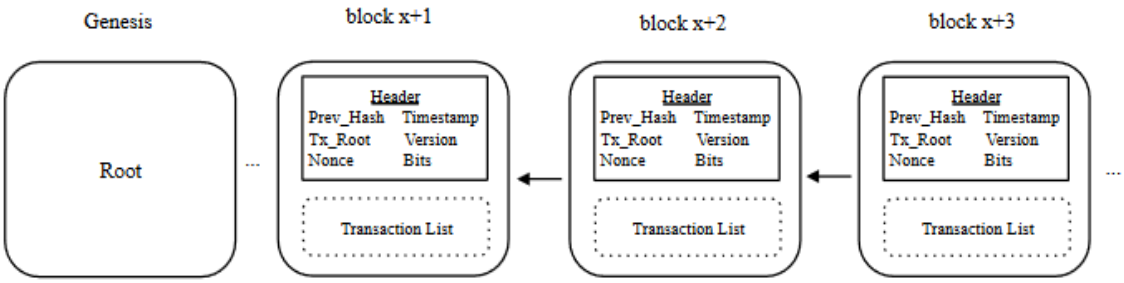
Las organizaciones de *Fintech*, a diferencia de las financieras tradicionales, se caracterizan por la constante aplicación de soluciones innovadoras y de nuevos procesos de negocio. Durante las últimas décadas, la sociedad ha podido disfrutar de numerosos desarrollos que estas han realizado y que son consideradas de uso común hoy día. Algunos de estos avances son, por ejemplo, los cajeros automáticos, las tarjetas de crédito, aplicaciones de finanzas móviles y web entre otros. El objetivo que persiguen las organizaciones con estas innovaciones es la reducción de costos, la creación de ventajas competitivas y la gestión de riesgos. A su vez también brindan beneficios para sus clientes. Estos suelen radicar en la facilidad de uso de servicios ofrecidos a través de la plataforma y la mayor disponibilidad de más funcionalidades (Vives, 2017). Algunos de los ejemplos más conocidos de organizaciones *Fintech* actualmente son los casos de Paypal, Venmo, Wenster Union y Zelle.

En la actualidad la atención de las organizaciones *Fintech* se centran en desarrollos relacionados con la analítica, la inteligencia artificial, la computación cuántica, la realidad virtual entre otros. Sin embargo, dentro de estas áreas de investigación una de las que genera mayor interés es el campo de las tecnologías de cadena de bloques o *Blockchain* (Fernandez, Rosillo, De la Fuente, & Priore, 2019). Esta tecnología se caracteriza por ser un campo en constante desarrollo puesto que su creación se relaciona con la presentación del *Bitcoin* en el año 2008 (Nakamoto, 2008).

En el momento de su creación, el *Blockchain* fue presentada como un sistema de pagos *peer-to-peer* (De un individuo a otro) para transacciones electrónicas. Esta permite a diferentes actores financieros él envío de pagos de uno a otro sin la participación de un agente central y es capaz de prevenir al mismo tiempo las problemáticas relacionadas con el gasto doble. Estas uniones de bloques de información se interconectan por medio de nodos que forman estructuras

similares a cadenas, dando origen al termino con el que se la conoce. (Fernandez, Rosillo, De la Fuente, & Priore, 2019)

Figura 1: Blockchain Network



Fuente: Tomado de Fernandez, Rosillo, De la Fuente, & Priore, 2019

En la figura 1 pueden observarse las principales características de funcionamiento de una red *Blockchain*. Entre estas, la principal encuentra en la de ser de una base de datos transaccional distribuida. Esto implica que cada uno de sus nodos se encarga de verificar la información contenida en dicha base datos (Fernandez, Rosillo, De la Fuente, & Priore, 2019).

El *Bitcoin* opera sobre la estructura tecnológica provista por la cadena de bloques. Este puede ser definido como un software de código libre que es procesado por una red de computadoras compartidas conocidas como nodos. Estos nodos comparten una base de datos donde se almacenan copias exactas de los registros de la cadena de bloques, actuando como un medio de validación para las transacciones dentro de la red del *Bitcoin.* A su vez es lo que permite su funcionamiento de manera descentralizada (Franco, 2014).

Debido a sus características criptográficas y de descentralización, el *Bitcoin* es parte de un conjunto de monedas digitales conocidas como criptomonedas. En este sentido, *Bitcoin* fue la primera criptomoneda en funcionamiento (Nakamoto, 2008). Su rápido crecimiento en combinación con la pronunciada subida de su precio llevo a la creación de otro gran número activos digitales conocidos con el termino de *altcoin* o monedas alternativas. Estas buscan ganar capitalización de mercado añadiendo o mejorando aspectos tecnológicos que las diferencien del *Bitcoin.* A pesar de estas mejoras, el *Bitcoin* continúa manteniendo su posición de liderazgo debido a su capitalización de mercado que supera el 40% de todo el conjunto de las criptomonedas (Sebastiao, Rupino, & Godinho, 2021).

El *Bitcoin* posee varias características que lo diferencian de las divisas o medios de intercambio tradicionales. Uno de los factores distintivos más evidentes es el hecho de que el *Bitcoin* no es un elemento físico, sino puramente digital y descentralizado (Ammous, 2018). Otra característica importante es que el *Bitcoin* posee su propio sistema económico interno, debido a que el mismo no es producido por una autoridad monetaria centralizada. Por lo tanto, no existe forma de que una entidad de estas características pueda influir en forma directa sobre su precio o la emisión de unidades de este activo en el mercado (Eichengreen, 1996) .

Estas propiedades mencionadas traen consigo una serie de implicaciones sobre la percepción de los individuos con relación al *Bitcoin*, su regulación en los planos legal e impositivo y su aceptación y tratamiento por parte de los inversores. Muchos de estos aspectos continúan siendo tema de estudio y en evolución constantes. Al mismo tiempo, estos puntos explican parcialmente la volatilidad en la cotización del *Bitcoin* y sus fuertes ciclos de subidas y bajadas de precio (Baur & Dimpfl, 2021).

Debido a la creciente popularidad del *Bitcoin* en la sociedad, es cada vez más común ver organizaciones del rubro *Fintech* incluyendo a esta moneda digital entre sus opciones de inversión. Esto presenta la oportunidad y desafío de desarrollar herramientas que permitan a estas organizaciones anteponerse a la volatilidad de este activo. Sin embargo, para que sean capaces de manejar correctamente los grandes volúmenes de datos que resultan de utilidad en el ecosistema, se tienen que desarrollar ciertos métodos para su gestión.

1.2 Gestión de grandes volúmenes de datos

Las organizaciones *Fintech* involucradas en la inversión del *Bitcoin* necesitan herramientas para gestionar los grandes volúmenes de información que poseen y generar valor a partir de estos. Esta problemática no es exclusiva de esta industria. Hoy en día se producen aproximadamente 2.5 quintillones de bytes de datos lo cual, al mismo tiempo, significa que más de 90% de los datos en todo el mundo han sido creados en los últimos dos años (Aguilar, 2013). Estos datos son tanto de tipo estructurado como no estructurado y proceden de varios tipos de fuentes como pueden ser: registros transaccionales, entradas (post) en redes sociales, imágenes, videos, señales GPS entre muchos otros.

Con la creciente cantidad de información disponible, las organizaciones *Fintech* se han abocado a buscar formas de sacar provecho de los datos. Esto tiene varios motivos, principalmente, la ayuda que estos proporcionan para conocer lo que sucede internamente en

la organización y el contexto general de la industria donde esta se desarrolla. El *Big data*, termino con que se conoce a esta tendencia de uso de grandes volúmenes de datos, es extensible también a las soluciones de infraestructura y almacenamiento usadas para manejar estos datos.

El *Big data* tiene tres elementos que lo diferencian sobre el manejo de datos tradicional. Estos son, los volúmenes de datos que se emplean, la velocidad a la que estos datos se producen y la variedad de estos (McAffe, Brynjolfsson, Davenport, Patil, & Barton, 2012). El volumen se relaciona con la gran cantidad de información de la que se dispone. Esta crece exponencialmente y se encuentra en el orden de los miles de terabytes. En la actualidad, la mayoría de estos datos son de tipo no estructurado. Las redes sociales son uno de los generadores de la mayor parte de estos (Miskam & Radin, 2018).

La velocidad, está relacionada con cuan rápido nuevos datos son generados (McAfee, Brynjolfsson, Davenport y Barton, 2012). En el contexto de grandes organizaciones *Fintech*, nuevos datos suelen ser creados segundo a segundo. En organizaciones financieras tradicionales, esta generación de datos era mucho más lenta por lo cual eran más fáciles de manejar. Finalmente, la variedad, es un concepto asociado con la multitud de fuentes de información de las cuales los datos pueden ser extraídos.

En el entorno *Fintech*, así como en la mayoría de las organizaciones, los *stakeholders* o partes interesadas tienden a asumir que la información con la que se cuenta es la correcta para la toma de decisiones (Haug & Stentoft, 2011). De la misma manera, estos asumen que los requerimientos de privacidad y seguridad para salvaguardar los datos del cliente son seguidos (Haneem, et al., 2017). Sin embargo, nada de esto podría ser logrado sin contar con sistemas que permitan poner todos estos elementos bajo análisis. Estos sistemas ayudan a garantizar la integridad de los datos y la constantemente revisión de los requerimientos y procedimientos relacionados a estos.

Frente a la necesidad de estandarizar y debido a las características propias del ambiente del *Big data* mencionadas en los párrafos anteriores, se vuelve imprescindible crear un proceso que funcione como guía. Existen diferentes formas de aproximarse. Algunas organizaciones toman un acercamiento más centrado a los datos o *data-driven* y otras se decantan por una visión más centradas en procesos o *process-driven* (Cleven & Wortmann, 2010).

Indistintamente del acercamiento que la organización decida, se debe crear un proceso que se convierta las entidades centrales de esta. Smith & McKeen (2008) definen las entidades

centrales como un proceso de aplicación independiente que describe, controla y gestiona las entidades centrales del negocio. Esto permite la consistencia y precisión de los datos por medio del uso de pasos definidos para su manejo.

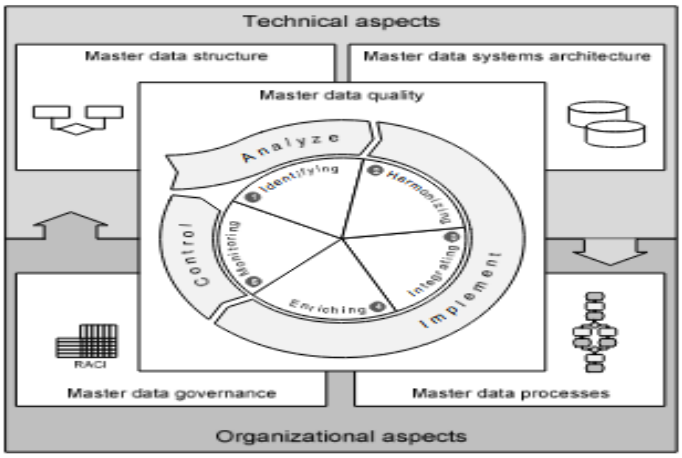
La implementación de este sistema resulta aún más apremiante cuando se considera que la no existencia de un programa apropiado de entidades centrales puede derivar en multitud de inconvenientes. Algunos de estos son: datos redundantes, datos inconsistentes, falta de estandarización, problemas en la validación y limpieza de los datos, desconocimiento de los lugares y sistemas donde los datos deben ser encontrados (Dreibelbis, et al., 2008). Las organizaciones *Fintech* son especialmente susceptibles a estas problemáticas. Esto se debe a que en estas no es inusual tener numerosas fuentes de datos de donde extraer registros. Al mismo tiempo, estos inconvenientes crecen exponencialmente mientras más clientes y productos posee la entidad.

Debido a las problemáticas que ocasiona la no presencia de un sistema de entidades centrales, surge la necesidad de establecer una serie de pasos para su implementación. Diferentes autores proponen metodologías que varían ligeramente entre ellas. Cleven & Wortmann (2010) sugiere que el correcto diseño del sistema de entidades centrales surge de cinco elementos claves que deben ser tenidos en cuenta, estos son: la estructura, la arquitectura, el gobierno de datos, los procesos y la calidad de los datos.

De esta forma Cleven & Wortmann (2010) describen los elementos de este sistema de la siguiente forma. La estructura se refiere a la necesidad de definir cada elemento y la relación entre cada uno de ellos. De esta manera, se evitan confusiones y errores en el manejo de los datos. La arquitectura está relacionada con establecer un diseño del sistema capaz de respaldar el ciclo de vida de las entidades centrales.

También Cleven & Wortmann (2010) explican los últimos dos elementos centrales del sistema. El gobierno de datos se relaciona con la necesidad de establecer claramente la necesidad de estos para las diferentes áreas de la organización y la relación entre cada una de ellas. Finalmente, los procesos, se encuentran formados por los pasos detallados que se indican de como cada una de las partes de la organización deben operar. La correcta combinación de los cuatro aspectos mencionados anteriormente (estructura, arquitectura, gobierno y proceso) trae como resultado la calidad de los datos.

Figura 2: Core elements of Master Data Management



Fuente: Tomado de Cleven & Wortmann, 2010

En la figura 2 se identifican las partes que forman el plan de entidades centrales, también se pueden ver los componentes de la calidad de datos. Esta cuenta con tres elementos que se explicaran a continuación. En la primera etapa de análisis, se identifican los datos relevantes para el sistema de entidades centrales en la organización. Durante la parte de implementación se persigue establecer los conceptos y procesos que serán aplicables a toda la organización. Finalmente, durante la etapa de control se busca corroborar que exista concordancia entre las acciones ejecutadas y lo estipulado en el plan de entidades centrales.

La correcta aplicación de este sistema proporciona varios beneficios a las organizaciones. Los más importantes radican en la precisión, el grado de consistencia y confiabilidad de los datos. Esto cobra aún más importancia cuando se trabaja con organizaciones *Fintech*, debido a que todos estos elementos permiten satisfacer las exigencias regulatorias en materia de información del cliente o *Know Your Customer (KYC)*, prevención de lavado de dinero o *Anti-Money Laundering* (AML) y privacidad (Dreibelbis, et al., 2008). Su cumplimiento resulta imprescindible para tener aprobación de los entes reguladores y operar adecuadamente.

En función de lo expuesto en este apartado, se puede apreciar como la gestión de información en las organizaciones es un proceso cíclico. Esta crece tanto en términos de costos como de dificultad proporcionalmente al crecimiento del volumen de datos. Poner en marcha estos sistemas en las organizaciones, principalmente en el rubro *Fintech*, requiere especial atención en los aspectos de la privacidad y control de los datos. Sin embargo, un plan de

entidades centrales necesita tener en cuenta de forma cuidadosa todas las partes de la organización para resultar efectivo.

Las organizaciones *Fintech*, no tienen únicamente una relación cercana con los grandes volúmenes de datos, pero de forma más específica, han sacado un gran beneficio de los Datos Alternativos o no estructurados. Estas organizaciones han sido capaces de implementar los datos alternativos en sus procesos y obtener importantes beneficios como resultado. Sin embargo, el uso de este tipo de datos requiere de consideraciones especialmente diseñados para su manejo. Estas consideraciones y la forma en que los datos no estructurados se relacionan con esta industria serán tema del siguiente subapartado.

1.3 Gestión de datos alternativos

Las organizaciones que trabajan con datos se encuentran en un contexto de generación constante y exponencial de los mismos. En los últimos dos años se han creado el 90% de los datos existentes y se espera que el volumen de estos continuara creciendo a un ritmo de 40% anual (Agusti, 2018). De este volumen de datos mencionado, más del 60% de los mismos presentes en organizaciones corresponden a la categoría de datos alternativos (Eberendu, 2016). Este tipo de datos provienen de diferentes fuentes dependiendo del tipo de organización. Los más comunes son aquellos provenientes de las redes sociales, noticias y geolocalización entre otros.

De forma particular, la industria *Fintech* centra muchos de sus procesos en el uso de datos alternativos para la toma de diferentes decisiones de negocio. Estos datos son tanto generados por la organización como adquiridos por medio de servicios externos para enriquecer sus bases de datos. El uso de esta información ha resultado determinante en muchos procesos recurrentes en este sector como son: la aprobación de préstamos, emisión de productos financieros, *Profiling* de clientes entre otros (Di Maggio, Ratmadowakara, & Carmichael, 2022)

La aplicación de datos alternativos en las organizaciones *Fintech* ha tenido efectos positivos. Algunos de estos pueden ser la generación de historiales financieros, permitiendo a los usuarios tener acceso a productos y servicios a los que no podrían sin los mismos (Johnson, 2019). Por otro lado, también se ha demostrado que tiene efectos negativos en aspectos como la privacidad de datos, abuso de campañas de Marketing y en algunos casos los efectos

desfavorables que estos datos alternativos pueden tener en la clasificación de usuarios de bajos ingresos (Johnson, 2019).

En el aspecto técnico, el uso de este tipo de datos ha exigido una transformación en las soluciones de almacenamiento y métodos usados para su tratamiento. Tradicionalmente los datos se sometían un proceso de extracción, transformación y carga (ETL por sus siglas en ingles). Esto tiene varias razones, principalmente, debido a que la velocidad y el tipo de datos que eran manejados en las finanzas tradicionales permitían el uso de este método sin mayores limitaciones. Esto también está relacionado a que anteriormente el uso de datos alternativos no era tan extendido e importante como lo es hoy día.

Sin embargo, frente al desafío que plantean los datos alternativos muchas organizaciones *Fintech* han optado por aplicar metodologías de extracción, carga y transformación (ELT por sus siglas en ingles). Como indica Broby & Hopper (2019) Esto permite mayor flexibilidad e integración de los datos, sin las limitaciones de la transformación previa. Esto aumenta las opciones de análisis a través de métodos más novedosos con datos de origines más variados.

Otra razón para este cambio de métodos se encuentra en que una de las características más destacables de los datos alternativos es que mayormente son producidos en tiempo real. Ejemplos de estos casos pueden ser observados típicamente en redes sociales y datos provenientes de aplicaciones (Eberendu, 2016). Este escenario requiere por parte de las organizaciones la creación de métodos para captar estos datos de forma más veloz que lo que se lograba con la aplicación del método ETL.

Es importante tener en cuenta como indica Salaberry (2019) que para mantener la integridad y calidad de los datos es fundamental contar con un programa de entidades centrales. Esto es, un programa con características similares a las planteadas en el subapartado anterior. Salaberry (2019) también menciona que esto gana aún más relevancia cuando las organizaciones conviven con diferentes tipos de datos provenientes de varias fuentes. Esto es usual en las organizaciones *Fintech*. Como unas de las principales ventajas resultantes de aplicar el programa de entidades centrales, Salaberry indica que se posibilita la integración de todas las fuentes de información en un mismo repositorio sin incongruencias de los datos.

Al mismo tiempo, dentro del sector *Fintech* resulta particularmente determinante contar con procesos de control de la información. Esto se debe a la necesidad de precisión en los datos manejados. Como resultado de aplicar estos procesos de control de entidades centrales durante

la extracción y carga de datos será posible detectar valores anómalos que requieran modificación previa para completar la integración (Salaberry, 2019).

Por último, la correcta integración y gestión de datos alternativos en las organizaciones *Fintech* resulta vital para el funcionamiento de estas. De esta manera, a pesar de la planificación y costos requeridos para creación y mantenimiento del programa de gestión de datos, es innegable que este resulta indispensable para la calidad y seguridad de estos. Los datos correctamente gestionados, se transforman en una fuente de información con mucho potencial para alimentar el desarrollo de la organización. Estos se pueden aplicar en la creación de nuevos productos, servicios y metodologías para la toma de decisiones como la que se plantean en este trabajo.

De esta manera la integración del *Bitcoin* como producto o como parte del portafolio de inversión de las organizaciones *Fintech* demanda el desarrollo de arquitecturas y metodologías de gestión adecuadas. Adicionalmente, se debe tener noción sobre las implicaciones legales y técnicas que se deben implementar. Esto afecta principalmente aspectos como la privacidad de los datos y la necesidad de incorporar herramientas adaptadas a la gestión de datos alternativos provenientes de fuentes novedosas como redes sociales y servicios de la web 2.0.

Capítulo 2: Procesamiento y análisis de datos alternativos

El análisis y predicción de activos bursátiles se ha convertido en uno de los principales usos de los datos alternativos en los últimos años (Yan,Zhou, Zhao, Tian, & Yang, 2016). Estos permiten realizar diferentes estudios sobre la intención de compra y volumen de interacción de los usuarios. De forma particular la red social *Twitter* se ha posicionada como una fuente predilecta de datos para este tipo de investigación debido a su extendido uso por parte de los inversores.

En el campo de la predicción del precio de un activo bursátil, como lo es el *Bitcoin*, existen dos corrientes principales. Una es el análisis técnico y la otra es el análisis fundamental (Petrusheva & Jordanoski, 2016). Los autores Petrusheva & Jordanoski (2016) indican que el análisis técnico tiene en consideración los patrones de precio y volumen en periodos anteriores, así como la acción del precio en el presente. Con el uso de estos elementos se trata de determinar cómo podría evolucionar la cotización del activo en el futuro. Los mismos autores indican que el análisis fundamental, por otra parte, se apoya en el estudio de acontecimientos en el mundo real y el análisis de información financiera para intentar hallar oportunidades de compra y venta que puedan resultar beneficiosas.

Los avances en la minería de datos y la creación de nuevos algoritmos han permitido la gestión de grandes cantidades de información en tiempo real. La disponibilidad de esta cantidad de información no era posible dentro del análisis fundamental hasta hace pocos años. Estos métodos han permitido utilizar las expresiones textuales de los usuarios en redes sociales como *Twitter* para analizar su influencia en los activos bursátiles

El análisis del lenguaje natural o NLP (Por sus siglas en inglés) aprovecha las capacidades computacionales actuales para obtener patrones contenidos en las expresiones de lenguaje (Khedr, Salama, & Yaseen, 2017). Estos pueden ser semánticos (basado en el significado del mensaje expresado) o sintácticos (basado en el orden dentro de la estructura gramatical) (Salaberry, 2019). Estas técnicas se aplican en numerosos campos de conocimiento, permitiendo la toma de decisiones de forma oportuna y resulta especialmente atractivo en un ambiente volátil como lo es la valoración de criptoactivos.

La aplicación del análisis del lenguaje natural requiere de una fase de procesamiento que permita convertir los datos no estructurados en una entrada adecuada para el modelo a utilizar

(Kalyani Joshi, Bharathi, & Jyothi, 2016). Durante esta etapa se eliminan elementos innecesarios del texto como URLs, nombres de usuario y símbolos. De la misma forma, se usan métodos para sustraer aquellas palabras que no agregan valor al texto, estas son conocidas como *stopwords* (Khedr, Salama, & Yaseen, 2017).

Sumado a los pasos planteadas anteriormente, también se transforman palabras para mejorar los resultados del algoritmo. Esto se logra por medio de procesos como el *steeming.* A través de este, se convierten las palabras a su estructura raíz (Jiva, 2011). Una vez que se cuenta con los textos procesados, se procede a tokenizarlos. Haciendo esto se convierten las palabras en una lista sobre la que se pueden aplicar las técnicas de lenguaje natural (Salaberry, 2019).

El estudio de los datos procesados permite obtener información fundamental para la toma de decisiones. En el ámbito de la valuación de activos uno de los métodos que ha captado mayor atención es el Análisis de sentimientos (Narendra, et al., 2016). Por medio de herramientas de análisis de lenguaje natural especializadas en este campo como VADER[3] (*Valence Aware Dictionary and Sentiment Reasoner*) se puede asignar una polaridad a las palabras contenidas en textos y determinar el tipo de sentimiento que este representa. Los resultados provenientes de la aplicación de este método sobre los *tweets* se pueden combinar con las variaciones en el precio de un activo como el *Bitcoin* para conocer si existe una influencia de las opiniones analizadas sobre el precio de este.

El capítulo desarrollado a continuación se estructura de la siguiente forma. En el primer apartado, se presenta el proceso por el cual se obtuvieron los datos para conformar el conjunto de datos sobre el cual se trabajará. Posteriormente, sobre este conjunto de datos se aplica el procesamiento y limpieza. En el siguiente apartado, se ponen en práctica técnicas de *Text Mining* con especial énfasis en el análisis de sentimientos y la interpretación de sus resultados. En el tercer apartado, se plantea un modelo de clasificación de *tweets* basado en *Machine Learning*. Finalmente, en el último apartado, se presentan las técnicas de Causalidad de Granger y correlación de series de tiempo que se usarán en el último capítulo para determinar si existe un efecto de las opiniones en los *tweets* sobre el precio del *Bitcoin*.

[3] Acerca de VADER https://github.com/cjhutto/vaderSentiment

2.1 Obtención y procesamiento de tweets

Con el objetivo de crear un conjunto de datos que contenga opiniones de inversores e individuos que interactúen con el *Bitcoin*, se decidió utilizar la red social *Twitter* como fuente de estos. La razón principal de su elección fue su extendido uso por usuarios de criptoactivos (Sattarov, Jeon, Oh, & Lee, 2020). Para poder interactuar con la plataforma y extraer los datos, se empleó la API de *Twitter* de uso público y gratuito. El acceso a esta fue solicitado con fines de estudio para el presente trabajo. La conexión con esta para la extracción de *tweets* se realizó mediante el lenguaje Python a través de la librería Tweepy de J. Roesslein (2020).

El acceso proporcionado por *Twitter* permite extraer los *tweets* perteniences a los ultimos 10 dias que contengan palabras determinadas elegidas por el investigador. Para este trabajo se extranjeron aquellos *tweets* que hacian referencia a "#BTC" como unico termino de busqueda. Claramente esto responde a recabar cualquier mencion sobre el activo bajo estudio. Al mismo tiempo, busca evitar la contaminacion de los *tweets* al incluir mas terminos que puediesen crear un ruido innecesario durante el analisis.

Los *tweets* analizados fueron extraidos durante un periodo comprendido entre el 15 de Junio del 2022 hasta el 15 de Julio del mismo año. El total de *tweets* recabados durante ese periodo fue de 71.646. Cada dia se extraian 2500 *tweets* con excepción de aquellos dias en los cuales la API no lograba recopilar dicha numero y arrojaba un resultado inferior. Los *tweets* reunidos no tenian ningun filtro geografico,sin embargo, el idioma seleccionado por medio de la API fue el ingles debido a la preponderancia de este idioma en el entorno de inversores del *Bitcoin*.

Los datos crudos extraidos de *Twitter* contienen una gran cantidad de ruido y elementos no deseados para el analisis. Este ruido esta compuesto por simbologia, direcciones webs y palabras que no agregan valor al proceso de analisis y clasificacion que se busca realizar. Con el fin de reducir lo mas posible los elementos mencionados, se realiza una etapa de procesamiento de datos. Este proceso esta compuesto por una serie de pasos entre los que se encuentran: La eliminacion de caracteres innecesarios como *hashtags* y otros simbolos, eliminacion de menciones a otros usuarios y la eliminacion de *stopwords*. Para realizar estos pasos se utilizo Regex (*Regular expressions)* (Frenz, 2008) y la librería de python NLKT[4] (*Natural Language Toolkit*).

[4] Acerca de NLKT https://www.nltk.org/

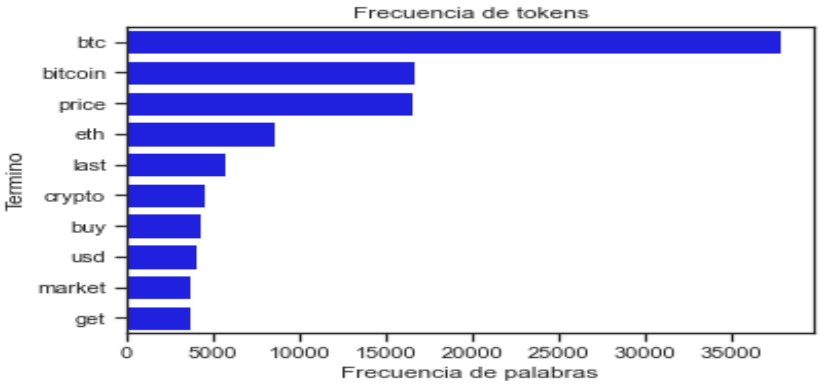
Adicionalmente, los *tweets* fueron sometidos a un proceso de *Steeming* el cual consiste en la reduccion de las palabras a su estructura raiz, por ejemplo jugando a jugar. Finalmente, con el objetivo de estructurar las palabras en listas que puedan ser procesadas por los metodos a utilizar, se procedio a tokenizar unicamente las palabras que se hayan mencionado un minimo de dos veces en todos los *tweets*. Este requerimiento en cuanto a la frecuencia de las palabras tiene el objetivo de reducir aun mas aquellas expresiones que pudiesen no agregar ningun sentido relevante. En la Figura 3 se aprecia un resumen de los pasos ejecutados.

Figura 3: transformación de tweets

Fuente: Elaboración propia

Como resultado de la tokenizacion se obtuvieron un total de 52.742 tokens en los 71.646 registros que componen el conjunto de datos a analizar. En la figura 4 se puede observar cuales son aquellas palabras que tienen mayor frecuencia dentro del conjunto de datos. Algunas de ellas como “btc”, “bitcoin” y “price” resultan entre las más comunes. Esto resulta esperable dentro del contexto de trabajo. Otras como “buy” o compra, “like” o gustar y “alert” o alerta denotan opiniones de usuarios que resultan recurrentes en los *tweets* analizados.

Figura 4: Frecuencia de los tokens



Fuente: Elaboración propia

El procesamiento de los *tweets* extraídos resulta una etapa crucial. Esto se debe a que la ausencia de esta puede condicionar considerablmente los resultados obtenidos por medio de técnicas y modelos a aplicar en las etapas posteriores. Al mismo tiempo, se puede apreciar de forma prematura ciertos elementos que caracterizan al conjunto de datos, como son las palabras más frecuentes. Sin embargo, con el objetivo de obtener mas informacion de los datos es necesario aplicar técnicas de minería de textos que ayuden a cumplir este cometido. El siguiente apartado se centrara en la aplicación de una técnica de *Text Mining* como es el Analisis de Sentimientos y su correspondiente interpretacion de resultados.

2.2 Análisis de tweets a través de técnicas de *Text Mining*

El *Text Mining* o minería de texto se refiere al proceso de extraer patrones de interés o conocimiento potencialmente útil desde documentos de texto (Hotho, Nurnberger, & Gerhard, 2015). Esta área resulta fundamental en las organizaciones debido a que se estima que el 80% de los registros en estas se encuentran compuestos de datos no estructurados siendo el formato de texto la mayoría de este porcentaje (Eberendu, 2016). El volumen de datos disponibles para analizar ha atraído la atención de las organizaciones hacia él *Text Mining*. En consecuencia, se han desarrollado durante los últimos años un gran número de usos comerciales por medio de los cuales obtener beneficios de estos (Ah-Hwee, 2000).

El *Text Mining* cuenta con diferentes técnicas a ser aplicadas dependiendo del tipo de análisis que se busque realizar. Alguna de las áreas en las cuales estos algoritmos se aplican son la extracción de información, el seguimiento de tópicos, la categorización, el *Clustering* o aglomeración de términos y el *concept linkage* (Vishal & Gurpreet Lehai, 2009). Estas técnicas mencionadas nos permiten visualizar los datos y responder preguntas de la organización con ellos.

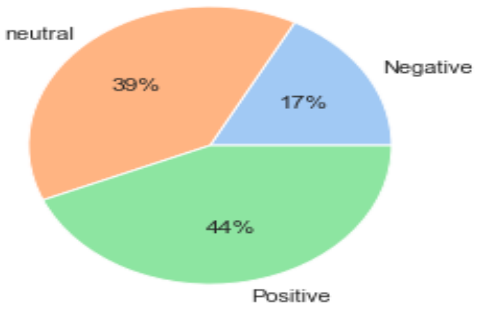
La cantidad de información textual compartida por usuarios en las redes sociales con relación a intereses, preferencias y opiniones sobre productos, eventos y circunstancias ha crecido exponencialmente en los últimos años (Smith & O'Hare, 2022). Esto se debe principalmente al auge de la web 2.0 y la creación continua de nuevas redes sociales, blogs y sitios de *streaming*. Estos datos, al igual que los de origen organizacional, son de gran interés para la aplicación de técnicas de *Text Mining*.

Uno de los algoritmos más útiles para analizar este tipo de datos provenientes de redes sociales es el Análisis de sentimientos. Este método se enfoca en la clasificación automática de textos dependiendo de su contenido, asignando una categoría de positivo, negativo o neutral (Narendra, y otros, 2016). Existen diferentes formas de aplicar este método, pero en el caso de este trabajo se utiliza un modelo basado en Bag of Words (BOW) que enfatiza las palabras usadas en lugar del contexto en el que se usan (Smith & O'Hare, 2022). La herramienta usada para el Análisis de sentimientos VADER, contiene referencias de la polaridad de cada palabra, asignando pesos a cada una de ellas. La valoración del sentimiento en un texto determinado es el resultado de la suma de los pesos de cada palabra contenida en él.

El uso de la herramienta VADER sobre otras opciones disponibles se fundamenta en dos aspectos importantes. Como indican Smith & O'Hare (2022), en primer lugar, es rápida en su procesamiento y no quiere ningún entrenamiento previo. La segunda ventaja radica en que se especializa en datos provenientes de redes sociales como los usados para este trabajo provenientes de *Twitter*.

Al aplicar el Análisis de sentimientos sobre el conjunto de datos procesados y tokenizado en el apartado anterior, se pueden visualizar en la figura 5 a continuación la distribución de los sentimientos de los *tweets* extraídos. De esta manera, el 17% fue clasificado como negativo, 44% como positivo y 39% como neutral. Gracias a esto, podemos intuir que la percepción de la comunidad de *Twitter* con respecto al *Bitcoin* tendió a ser positiva durante la mayor parte del periodo en estudio.

Figura 5: Distribución de los sentimientos asociados a los tweets extraídos.

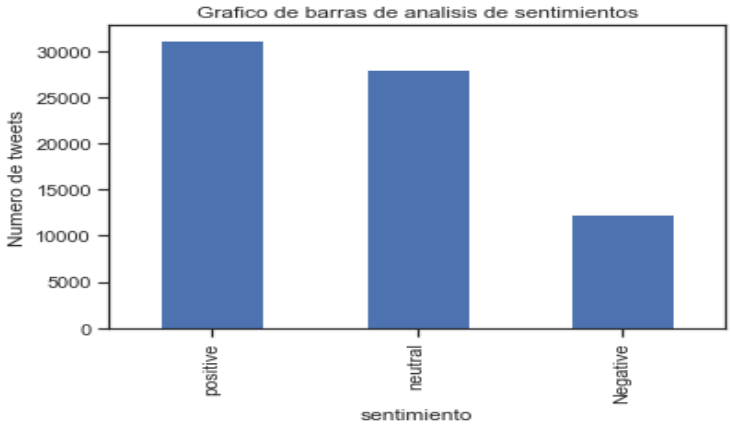


Fuente: Elaboración propia

De la misma forma, en la figura 6 se puede apreciar una representación similar de la frecuencia de cada categoría. Esta vez visualizando el número total de los registros clasificados en cada sentimiento. Esta figura nos permite observar como el Análisis de sentimientos resulto

en 31.217 registros clasificados como positivos, 28.045 como neutrales y 12.384 como negativos. Esto, confirma en números absolutos que la mayoría de las opiniones fueron positivas durante el intervalo analizado en el trabajo.

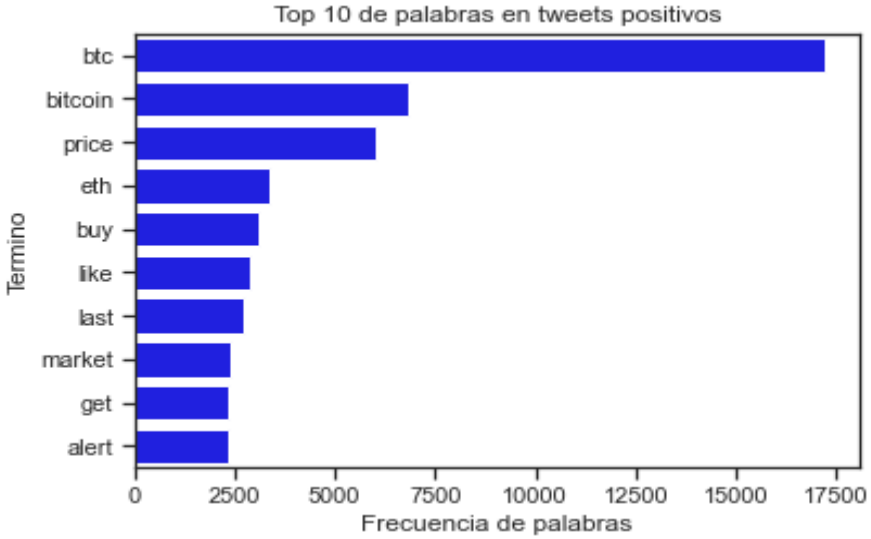
Figura 6: Numero de tweets asociados a cada sentimiento



Fuente: Elaboración propia

Debido a la significativa mayoría de palabras clasificadas como positivas, resulta interesante apreciar el top 10 de palabras contenidas en esta clase. De esta manera en la figura 7 se puede ver como en esta categoría existen varias palabras que denotan intenciones de compra o que pueden resultar alcistas en la cotización del *Bitcoin*. Las más relevantes en ese sentido serian "buy" o compra, "like" o gustar y "get" u obtener.

Figura 7: Top 10 de palabras más frecuentes en tweets positivos



Fuente: Elaboración propia

Por medio de la aplicación del Análisis de sentimientos se pudo constatar a priori la preponderancia de un sentimiento positivo en los *tweets* extraídos para análisis. Existe una frecuencia considerable de palabras que se relacionan con operaciones de compra o expresiones de interés sobre el *Bitcoin*. Al mismo tiempo, es importante mencionar que la extracción de opiniones y su análisis es una tarea desafiante, esto se debe a que a pesar de las técnicas aplicadas durante la etapa de procesamiento y limpieza siempre existirá ruido en los datos. Igualmente, la clasificación de datos en dominios especializados como lo es el mercado bursátil supone otro desafío debido a la existencia de términos y expresiones específicos a esta área. En el próximo apartado se planteará como usar los datos extraídos y etiquetados como entrada para un modelo de clasificación basado en *Naive Bayes* que pueda ser de utilidad para las organizaciones *Fintech* a la hora de categorizar el estado de ánimo expresado en textos.

2.3 Modelos de clasificación para Tweets en organizaciones Fintech

En la actualidad, las redes sociales han tomado un papel de relevancia en la comunicación y la formación de la opinión pública (Miskam & Radin, 2018). Como resultado, el análisis de los datos generados en estas puede proporcionar una valiosa fuente de información. De forma particular, en la industria *Fintech*, se han utilizado los datos provenientes de estas redes en el análisis de tendencias, la medición del sentimiento de los usuarios sobre una empresa o producto particular, el análisis de préstamos y el estudio de la reputación de las organizaciones (Utami Handika, Ade Purnama, & Nizar Hidayanto, 2022; Di Maggio, Ratmadowakara, & Carmichael, 2022). En estos ejemplos, modelos de clasificación son entrenados con los datos extraídos y aplicados para automatizar estos procesos de categorización.

Existen diferentes métodos de clasificación en función del modo en que son entrenados. El tipo de método también tendera a poseer ciertas ventajas y desventajas que indicaran cual es más apropiado para cada tipo de análisis. Estos algoritmos se dividen principalmente en tres grupos siendo estos: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje semi-supervisado (Narendra, et al., 2016).

Este trabajo se centrará en los modelos de aprendizaje supervisado, debido a la naturaleza de la pregunta de investigación planteada y a la disponibilidad de datos para entrenar el modelo. En el aprendizaje supervisado, se utilizan datos previamente etiquetados para entrenar el algoritmo y permitirle hacer predicciones sobre nuevos registros (Sarker, 2021). Los datos de

entrenamiento se pueden etiquetar de varias maneras, siendo las opciones más comunes el etiquetado automática o manual de los registros. Matemáticamente, la clasificación se puede definir como una función que, dado un conjunto de entradas X, un conjunto de etiquetas Y y un conjunto de entrenamiento T, asocia cada entrada con su etiqueta correspondiente, permitiendo predecir la etiqueta de una nueva entrada.

El uso de modelos de clasificación para ayudar a organizaciones *Fintech* a reducir el riesgo en el proceso de toma de decisiones es ampliamente estudiado (Ghaith , Alkubaisi, Sakira, & Husni, 2018). A raíz de esto, existe un repertorio amplio de métodos a usar. En este sentido cada uno de estos cuenta con diferentes ventajas y desventajas. Usualmente, el mayor inconveniente tiende a ser la baja precisión de las predicciones que repercute negativamente en la confiabilidad de los resultados. Es parte del proceso de entrenamiento elegir los parámetros y el algoritmo que de un mejor resultado para este parámetro.

Existen diferentes elementos que afectan la precisión de los modelos. Los mas relevantes son la calidad de los datos, la subjetividad del lenguaje y la limitación del vocabulario (Shabunina, 2013). En el trabajo realizado se intento reducir la influencia negativa de estos por medio del procesamiento de los datos y la aplicación de una herramienta para el análisis de sentimiento en textos especializada en las redes sociales como es VADER.

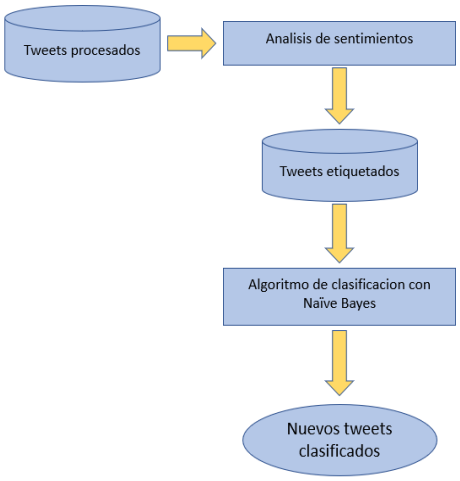
VADER fue aplicado para el etiquetado automático de los *tweets* por medio del Análisis de sentimientos. Estos *tweets* ya etiquetados componen el set de entrenamiento y prueba para el modelo predictivo. En este sentido, se eligió un algoritmo de clasificación basado en *Naive Bayes*. Este método se fundamenta en el teorema de Bayes, que asume la independencia entre las variables (Berrar, 2019). *Naive Bayes* funciona particularmente bien en la clasificación de documentos de texto, al mismo tiempo, no necesita una gran cantidad de datos de entrenamiento para ser efectivo (Sarker, 2021). Este algoritmo se expresa matemáticamente con la siguiente ecuación 1.

$$p(C_K|x) = \frac{p(C_K)p(x|C_K)}{p(x)} \quad (1)$$

donde, P(C_K) representa la probabilidad previa de la clase C_K, $p(C_K|x)$ representa la probabilidad de la clase condicional y $p(x|C_K)$ es la probabilidad de x de pertenecer a la clase C_K . Existen diferentes variantes de Naive Bayes como son la Multimodal y Bernoulli.

La figura 8 muestra un diagrama de flujo que ilustra las etapas del proceso utilizado para preparar y entrenar el modelo. Este proceso comienza con los datos procesados y finaliza con la predicción del modelo de clasificación. En el medio de este diagrama se encuentra el etiquetado automático usado para la creación del conjunto de datos de entrenamiento por medio de VADER.

Figura 8: Diagrama de flujo de clasificador de tweets



Fuente: Elaboración propia

Como cierre de este apartado, es posible apreciar como los modelos de clasificación a pesar de sus limitaciones resultan en herramientas poderosas para analizar y entender de forma rápida la gran cantidad de datos generados en redes sociales como *Twitter*. Por medio del uso de un algoritmo de aprendizaje supervisado, como lo es *Naive Bayes*, se puede entrenar un modelo de clasificación que determine si un *tweet* es positivo, negativo o neutral basado en las palabras que componen el cuerpo del tweet. Esto ayuda a analizar rápidamente las tendencias y responder interrogantes sobre la opinión pública de un activo bursátil como lo es el *Bitcoin*. Igualmente, con la ayuda de estos patrones se pueden descubrir nuevas oportunidades de negocios o tendencias al relacionar el Análisis de sentimientos con otras variables. En el siguiente apartado se profundizará sobre las técnicas usadas para determinar si existe una relación o causalidad entre el resultado del Análisis de sentimientos y la cotización del *Bitcoin*

2.4 Causalidad de Granger y correlación de series de tiempo

El concepto de causalidad es fundamental en la investigación y el análisis de datos. En el contexto de las series de tiempo, la causalidad se refiere a la relación entre dos variables, donde una variable es considerada la causa y la otra es el efecto (Granger, 1969). Una forma comúnmente utilizada para evaluar la causalidad en las series de tiempo dentro de un ámbito econométrico, como el planteado en el presente trabajo, es a través del concepto de causalidad de Granger. A su vez, existe otra forma que se explorará en este apartado como es el análisis de correlación utilizando el método de Pearson.

La causalidad de Granger es una técnica aplicada en el análisis de series de tiempo para determinar si una serie temporal es capaz de predecir el comportamiento de otra. En estos casos, se dice que si una serie X_1 "Granger-causa" una serie X_2, los valores anteriores de X_1 deben contener valores que ayuden a predecir X_2 (Granger, 1969). Este método es frecuentemente utilizado para encontrar factores que afecten los precios de activos bursátiles (Wang, Ho, Liu, & Wang, 2013). Matemáticamente, la causalidad de Granger se expresa con las siguientes ecuaciones 2 y 3.

$$X_t = \sum_{j=1}^{m} a_j X_{t-j} + \sum_{j=1}^{m} b_j Y_{y-j} + \varepsilon_t \quad (2)$$

$$Y_t = \sum_{j=1}^{m} C_j X_{t-j} + \sum_{j=1}^{m} d_j Y_{y-j} + n_t \quad (3)$$

Donde ε_t, n_t son dos series no relacionadas de ruido blanco.

El ruido blanco como indican (Moffat & Akpan, 2019) puede ser definido como una secuencia de variables aleatorias no correlacionadas. Este tiene una media de cero y varianza constante. Debido a esto, un proceso de ruido blanco por definición es estacionario.

Es fundamental para la aplicación de la causalidad de Granger que las series de tiempo a analizar sean estacionarias (Wanbin Wan, Kin-yip, Wai-Man, & Wang, 2013). Esto significa que su comportamiento a lo largo del tiempo no cambie significativamente. Estadísticamente se interpreta como una media, varianza y distribución de probabilidad que no varían considerablemente a lo largo de la serie temporal. Si esto no se cumple, es necesario realizar una transformación de las series. En el presente trabajo, se utilizó la prueba de Dickey Fuller para determinar si se cumplía este supuesto.

Otro procedimiento utilizado para medir la similitud existente entre dos series temporales es el análisis de correlación de series de tiempo. De esta forma, si dos series temporales tienen una alta correlación, su comportamiento a lo largo del tiempo tendera a ser similar, en caso contrario, su comportamiento difiere (Smith & O'Hare, 2022) . Para este trabajo se utilizó la correlación de Pearson. Esta proporciona un rango entre +1 a -1 siendo +1 una correlación perfecta y el -1 representar dos series completamente opuestas. La correlación de Pearson se expresa con la siguiente ecuación 4.

$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \quad (4)$$

donde σ es la desviación estándar en ambas variables y ρ es el coeficiente de correlación de las series. En el presente trabajo, este último representaría la correlación entre la serie de cambios de precio y la serie que contiene la sumatoria de los valores del análisis de sentimiento en cada día. Para la ejecución de este análisis se utilizó el método corr de la librería pandas[5] en Python.

En resumen, la causalidad de Granger y el análisis de correlación son herramientas valiosas para el análisis de series de tiempo. Ambas se utilizan para examinar las relaciones entre diferentes series temporales y entender en qué medida estas son compatibles. En el próximo capítulo, se presentarán los resultados obtenidos mediante los métodos propuestos.

[5] Acerca de pandas https://pandas.pydata.org/docs/

Capítulo 3: Relación de los tweets con las variaciones del precio del Bitcoin

En este capítulo, se aplicarán las técnicas de análisis de causalidad y correlación además del método de clasificación de *tweets* presentados en el segundo capítulo. Adicionalmente, se expondrá la importancia de los datos alternativos dentro de las organizaciones *Fintech*. Esto busca responder la pregunta central sobre la capacidad del análisis de sentimientos para determinar el precio del *Bitcoin* y la efectividad de los métodos de clasificación aplicados a este tipo de tareas en las organizaciones. Los resultados obtenidos se presentarán en tres secciones.

En el primer apartado, se implementará el modelo de clasificación propuesto basado en *Naive Bayes* y se evaluará su rendimiento utilizando diferentes parámetros como *precisión*, *recall* y *F1-score*. Con esto, se busca determinar qué grado de fiabilidad se le puede atribuir al modelo en el contexto de las organizaciones *Fintech* para predecir la opinión contenida en un tweet de un inversor con respecto al *Bitcoin*.

Posteriormente, en el segundo apartado, se procede a utilizar las pruebas de causalidad y correlación planteadas en el capitulo previo. Con esto se buscará, dentro del periodo temporal estudiado, determinar si existe un efecto estadísticamente significativo de las opiniones expresadas en los *tweets* sobre el precio del *Bitcoin*. En el marco del presente trabajo, esto se describe como una situación en la que el porcentaje de cambio del precio del *Bitcoin* fue positivo, negativo o neutro dependiendo de la animosidad dominante en las opiniones expresadas en *Twitter* durante el día previo. De esta forma, se analizan los efectos de las opiniones de *Twitter* 24 horas luego de que estas fueron realizadas.

Finalmente, en base a los resultados obtenidos, en el tercer apartado se analizará la importancia de los datos alternativos para determinar la variación del precio del *Bitcoin*. Con esto, se busca responder a la pregunta de investigación e indicar la utilidad que este tipo de datos puede representar para las organizaciones *Fintech*, de forma particular, en la predicción del precio de criptoactivos.

3.1 Implementación y evaluación de resultados del modelo de clasificación de tweets

Una vez explicados los conceptos teóricos del modelo de clasificación basado en *Naive Bayes* en el capítulo anterior. En este apartado, se presentará la implementación y evaluación del modelo. Finalizada la etapa de entrenamiento, el modelo será evaluado para determinar su precisión y rendimiento utilizando diferentes parámetros.

En primer lugar, para elegir la variante del modelo y parámetros que dieran mejores resultados, se realizó una serie de pruebas. En estas se evalúa el rendimiento del algoritmo utilizando unigramas y bigramas. De igual forma, se probaron diferentes algoritmos de *Naive Bayes* como son el multinomial, Bernoulli y una variación del multinomial conocida como *ComplementNB.* Adicionalmente, estos algoritmos fueron evaluados con y sin la aplicación de tf-idf (*Term frequency-inverse document frequency*) método que permite ponderar la importancia de una palabra en relación con su uso dentro del texto (Kalyani Joshi, Bharathi, & Jyothi, 2016). Finalmente, es importante resaltar que para evaluar los valores de *Accuracy (*indicador que mide la proporción del total de muestras clasificadas correctamente*)* se realizó una validación cruzada utilizando un *K-Fold validation* de 10 *folds*.

El *K-Fold validation* es un método de validación cruzada. Esto quiere decir, que el conjunto de datos es dividido aleatoriamente en un numero n de partes. De esta forma, $\frac{1}{n}$ es utilizada como el conjunto de datos de prueba y el resto se usa para el entrenamiento del modelo (Shabunina, 2013). Este proceso es repetido un numero n de veces usando cada sección del conjunto de datos una vez.

Previo a explicar los resultados obtenidos por medio de las pruebas realizadas, es importante tener en consideración que un resultado de *Accuracy* o precisión de un modelo de clasificación no puede ser considerado aisladamente. Esto significa, sin saber cual sería el resultado mínimo que supondría una mejoría con respecto al modelo base o *baseline* (Devasena, 2014). Este modelo base varia en relación con el método que la organización haya usado previamente para clasificar. Por lo tanto, el objetivo es que los nuevos modelos de clasificación superen este valor. Para la implementación realizada se tomó como *baseline* el método *Zero-R Classifier,* que calcula el modelo base prediciendo siempre la clase mayoritaria sobre el total de la muestra (Devasena, 2014). Como resultado de su aplicación, en base a un total de 71.646 registros de los cuales 31.217 fueron clasificados como positivos, el *Zero-R Classifier* dio un resultado de 43.57%.

Este resultado significo que, para justificar la implementación de un nuevo modelo de clasificación en la organización, este debería tener un *Accuracy* mayor a 43.57%. De lo contrario, el proceso de seleccionar la clase con mayor frecuencia seria más efectivo. De esta manera, se facilita la posibilidad de direccionar los recursos de la empresa a métodos que realmente presenten resultados prometedores.

Adicionalmente, existen otras métricas de desempeño que serán utilizadas para evaluar el método. Como indican Ghaith , Alkubaisi, Sakira, & Husni (2018) las metricas mas usadas en sistemas de clasificacion son *: Recall, Precision, F1-Measure y support*. El *Recall* es la proporcion de casos que fueron identificados como positivos. *Precision* es el total de casos positivos divididos por el total de casos positivos mas los falsos positivos. Finalmente, el *F1-Measure* es un calculo que combina los resultados de *Precision* con el *Recall* y el *Support* esta definido como el numero de ocurrencias de cada clase.

Otra metrica que sera utilizada para presentar los resultados es el area bajo la curva ROC. Esta metrica muestra el ratio de verdaderos positivos (Beltrame, 2020). En este caso corresponderia con aquellos *tweets* correctamente clasificados como positivo, negativo o neutral cuando estos lo son tambien en el conjunto de datos de prueba.

Durante las pruebas de las variantes del modelo de clasificacion, fue evidente que la aplicación de bi-gramas asi como el uso de *Tf-Idf* no genero resultados superiores en comparacion al uso de monogramas sin el uso de *Tf-idf*. Tambien se evidencio que el tipo de *Naive Bayes* que usa el metodo de Bernoulli resulto superior a los otros algoritmos evaludos con un *accuracy* del 83.65%. Este valor fue obtenido al promediar los resultados arrojados por el *K-Fold Cross Validation*. Es importante destacar que este resultado supera ampliamente el obtenido por el metodo de *Zero-R Classifier* siendo de esta forma superador al modelo base.

En la Tabla 1 a continuacion, se puede observar el reporte de clasificacion para el modelo con mejores resultados. En este se encuentran contenidos los valores para *Precision, Recall, F1-Score* y *Support*. Cada uno de estos esta dividio en las categorias disponibles en el conjunto de datos. En la ultima fila se puede observa el valor promedio generalizado de todas las categorias.

Tabla 1

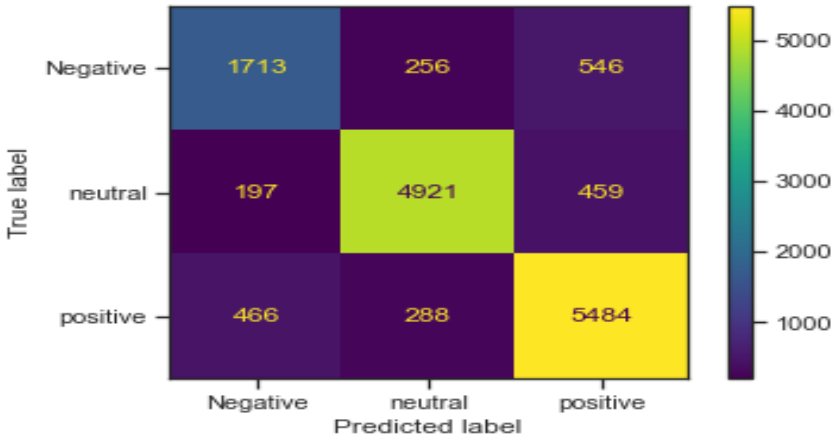
Reporte de clasificación para modelo de mayor precisión.

	Precision	**Recall**	**F1-score**	**Support**
Negative	0.72	0.68	0.70	2515
Neutral	0.90	0.88	0.89	5577
Positive	0.85	0.88	0.86	6238
Accuracy			0.85	14330

Fuente: Elaboración propia

En la Figura 9 se presenta la matriz de confusión. Esta divide las clasificaciones en función de su categoría verdadera y categoría predicha por el modelo seleccionado. De esta forma, se puede conocer rápidamente aquellos resultados que son positivos, neutros o negativos verdaderos o falsos y por lo tanto deducir que sesgo tiene el modelo empleado.

Figura 9: Matriz de confusión



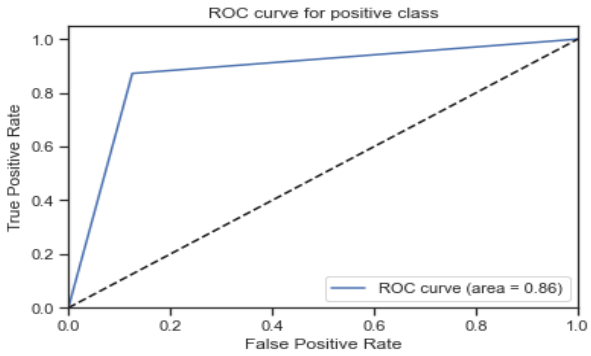
Fuente: Elaboración propia

La matriz de confusión muestra como el modelo tiene una precisión significativamente mayor detectando *tweets* verdaderos positivos por sobre los verdaderos negativos. De esta forma, hay mayor tendencia a los falsos negativos que falsos positivos. En términos del uso que la organización podría darle al modelo, esta tendencia a tener una mayor proporción de falsos negativos no debería resultar particularmente perjudicial siempre que la precisión del modelo en conjunto sea elevada.

Para finalizar el análisis de gráficos de rendimiento. En la figura 10 se presenta la curva ROC. El área bajo la curva da como resultado 0.8602. Este valor contrasta con la línea

transversal que indicaría un área bajo la curva de 0.50 en la cual la selección aleatoria seria igualmente efectiva al modelo.

Figura 10: Curva ROC



Fuente: Elaboración propia

La evaluación de los resultados arrojados por el reporte de clasificación y otros indicados asociados al modelo sugieren que este resulta eficaz en las tareas de clasificación de sentimientos. Esto se constata en la precisión con la que este modelo cuenta y la amplia mejoría que supone al compararse con el modelo base. El modelo tiene una leve tendencia a clasificar *tweets* negativos como positivos, sin embargo, sigue arrojando una mejora con respecto a la alternativa con que se compara. Es importante tener en consideración que el etiquetamiento de la base de entrenamiento fue realizado de forma automática por herramientas de análisis de sentimiento de uso general. A pesar de esto, siendo esta una solución de rápida implementación y bajo costo se puede evidenciar como resulta de utilidad para las organizaciones *Fintech* que enfrentan este tipo de problemáticas.

El siguiente apartado aborda el análisis de la causalidad y correlación de opiniones de *Twitter* sobre el precio del *Bitcoin*. Para determinar esta relación se usarán dos métodos como son la Causalidad de Granger y correlación de series de tiempo. Esto resultara en una respuesta a la pregunta central del presente trabajo sobre si es posible predecir los cambios del precio del *Bitcoin*, con las condiciones y características propias de este, utilizando los sentimientos en las expresiones provenientes de *Twitter*.

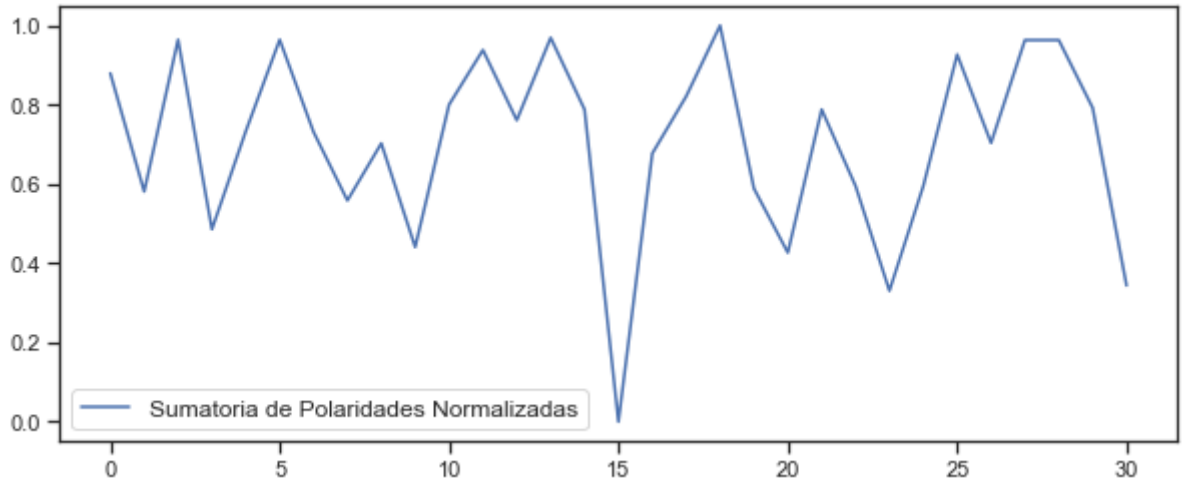
3.2 Análisis de la influencia de opiniones contenidas en tweets sobre el precio del Bitcoin usando Causalidad de Granger y correlación de series de tiempo.

Al contar con *tweets* correctamente procesados, analizados y etiquetados en función a las emociones que usuarios de *Twitter* expresan sobre el *Bitcoin*, se pueden desarrollar diferentes metodologías que permitan comparar esta variable con otras. Esto servirá para generar nuevas oportunidades de estudio potencialmente beneficiosas para la organización que cuente con los datos. En el caso del presente trabajo, la variable de comparación elegida es la evolución del precio durante el periodo en que los *tweets* fueron extraídos. Por medio del uso de dos métodos estadísticos como son la causalidad de Granger y la correlación de series de tiempo se busca conocer si existe una relación entre estas series temporales.

Para responder este interrogante, es importante estructurar los datos disponibles de forma tal que ambas variables puedan ser comparadas y analizadas. En este sentido, como se explico en apartados anteriores, el análisis de sentimiento devuelve un valor de polaridad que se encuentra entre -1 a 1 para cada tweet. Adicionalmente, conocemos que se extrajeron un total de 71.646 *tweets* con un promedio de 2500 *tweets* durante cada día. Por medio de la sumatoria de los valores de polaridad de cada uno de los *tweets* que fue escrito en un día determinado, podemos obtener un valor total para dicho día. De esta forma, un día en el cual el sentimiento general hubiese sido positivo tendrá un valor más alto que un día en que el sentimiento hubiese sido negativo, en este caso la sumatoria será negativa o tendera a ser un valor positivo bajo en relación al sentimiento positivo.

El resultado de este proceso puede ser apreciado a continuación en la figura 11. En esta se observa el grafico de la sumatoria de los valores para cada día contra los días transcurridos en el estudio. Es importante notar que ningún valor fue negativo, sin embargo, si existen día en los cuales la sumatoria fue considerablemente más baja que otros.

Figura 11: Sumatorio de polaridad de análisis de sentimiento por día normalizado.



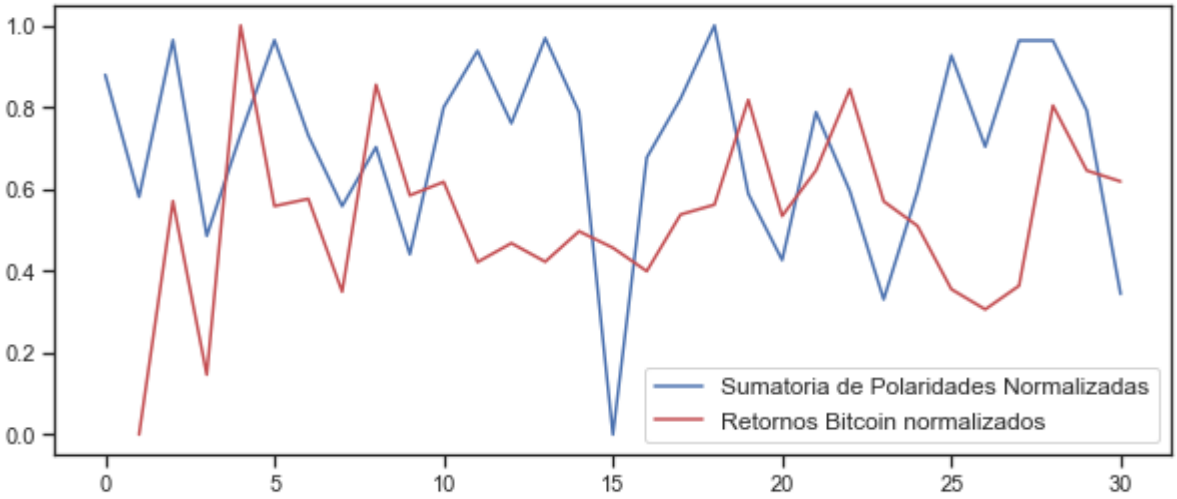
Fuente: Elaboración propia

Por otra parte, para calcular las variaciones del precio del *Bitcoin* se procedió a determinar el porcentaje de cambio que tenia la moneda de un día al siguiente usando el precio de cierre de esta. Los datos de cotización para el cierre de cada día fueron obtenidos por medio de la plataforma de inversiones *Yahoo Finance* [6]. De esta forma, la sumatoria de las polaridades de un día, causan los cambios de precio del día siguiente. Si el porcentaje de cambio del precio del *Bitcoin* fue positivo y la sumatoria de las polaridades del Análisis de sentimientos en el día anterior también, se presenta una correlación entre ellas. Caso contrario, no existe una correlación o casualidad entre las variables para ese día.

A continuación, en la figura 12 se procede a presentar ambas variables normalizadas para ver su evolución. Dicho proceso de normalización se realizo para reducir el impacto que tendría en el grafico y en otros análisis posteriores el comparar series con rangos de magnitudes considerablemente diferentes. De esta forma, se puede apreciar gráficamente como existen áreas en las que hay similitud en la evolución de las series temporales, así como otras en donde reaccionan de forma divergente.

[6]Acerca de Yahoo Finance https://finance.yahoo.com/

Figura 12: Evolución de sumatoria de polaridades y porcentaje de cambio de precio del Bitcoin normalizados.

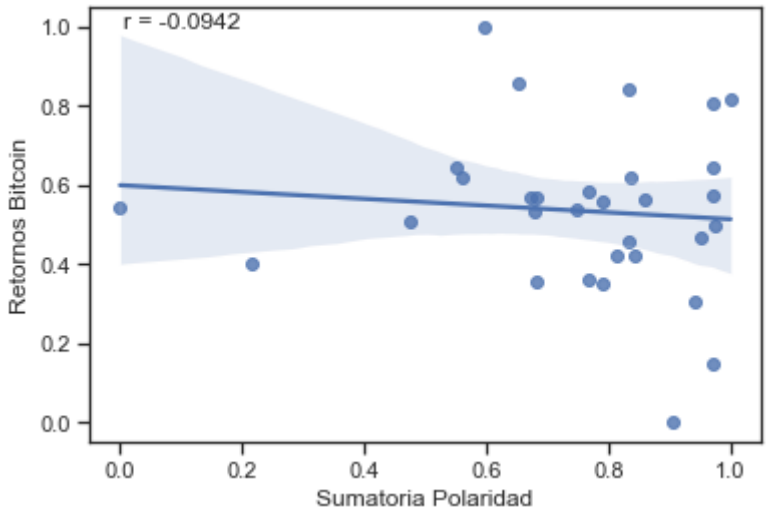


Fuente: Elaboración propia

La representación gráfica de ambas series temporales no resulta suficiente para determinar con precisión si existe o no causalidad. Para esto se procedió a aplicar los métodos de correlación de series temporales y causalidad de Granger explicados en el capítulo previo. Es importante resaltar que para la aplicación de la casualidad de Granger es necesario probar primero el supuesto de estacionalidad de las series temporales. Al aplicar la prueba de Dickey Fuller se pudo constatar que ambas series temporales son estacionarias con una confianza del 99%.

El análisis de correlación de series temporales usando el método de Pearson dio como resultado un coeficiente de -0.0942. Esto, como fue explicado en apartados anteriores, indica que hay una correlación ligeramente negativa. Sin embargo, esta es extremadamente débil como para ser considera significativa y concluyente. A continuación, en la figura 13 se presenta el grafico que contiene la correlación entre ambas series temporales.

Figura 13: Análisis de correlación de series temporales



Fuente: Elaboración propia

Por otra parte, el análisis de causalidad de Granger fue ejecutado para evaluar hasta 3 *lags* que representan 3 retrasos. Es decir, se analiza si la serie temporal de sumatoria de la polaridad de sentimientos tiene algún efecto sobre los retornos del precio del *Bitcoin* haciendo uso de hasta 3 periodos anteriores al valor actual de la serie de sumatoria de polaridad con la que se compara. De esta forma, no solo se puede conocer si existe causalidad de Granger utilizando el valor un día antes de su efecto en el precio del *Bitcoin*, pero también dos y tres días antes.

A continuación, en la Tabla 2 se pueden observar los resultados de la prueba de causalidad de Granger. De esta forma, utilizando un nivel de significancia del 95% podemos observar que no es posible asegurar que una serie Granger causa a la otra en ninguno de los retrasos evaluados. Esta conclusión se replica sin importar el tipo de prueba de hipótesis aplicada en el análisis.

Tabla 2

Resultados prueba de causalidad de Granger

Numero de retrasos = 1		
Prueba F	F= 0.1090	p= 0.7439
Prueba Chi2	Chi2= 0.1216	p= 0.7273
Parámetro prueba Chi2	Chi2= 0.1213	p= 0.7276
Parámetro prueba F	F=0.1090	p= 0.7439
Numero de retrasos = 2		
Prueba F	F= 1.8323	p= 0.1826
Prueba Chi2	Chi2=4.4613	p= 0.1075
Parámetro prueba Chi2	Chi2= 4.1396	p= 0.1262
Parámetro prueba F	F= 1.8323	p= 0.1826
Numero de retrasos = 3		
Prueba F	F= 1.0716	p= 0.3835
Prueba Chi2	Chi2= 4.3400	p= 0.2270
Parámetro prueba Chi2	Chi2= 4.0246	p= 0.2588
Parámetro prueba F	F= 1.0716	p= 0.3835

Fuente: Elaboración propia

Por medio de las pruebas realizadas es posible deducir que no se puede probar una correlación o causalidad entre las dos variables estudiadas con los parámetros establecidos para este estudio. Sin embargo, es importante tener en consideración que el análisis de las series temporales de activos bursátiles esta afectado por numerosas variables como son la tendencia del activo, la macroeconomía entre otros. En el siguiente y ultimo apartado de este trabajo se presenta un análisis de la importancia de los datos alternativos para determinar las variaciones de precio del *Bitcoin*.

3.3 Importancia de los datos alternativos para la determinación de la variación del precio del Bitcoin

El uso de los datos para determinar el precio de activos bursátiles ha sido sujeto de estudio durante décadas. Sin embargo, las fuentes de información tradicionalmente utilizadas para

estos análisis se limitaban a registros de ingresos, proyecciones de rendimiento y factores macroeconómicos propios de la organización o activo a considerar (Petrit & Kuql, 2019). A pesar de la utilidad de estos reportes, presentan limitaciones importantes como la imposibilidad de ser generados en tiempo real y el no contar con grandes volúmenes de datos.

El avance de las tecnologías para procesar, almacenar y analizar datos de diferentes fuentes como son las redes sociales y comunidades online posibilito la creación de nuevos tipos de análisis alternativos. Estos permitieron solventar las limitaciones de los métodos tradicionales gracias a poder ser extraídos en tiempo real y con volúmenes de datos muy superiores (McAtter, 2014). Estos dos aspectos son de vital relevancia cuando se aplican a determinar las variaciones de precio en activos como el *Bitcoin*. Esto se debe a su alto nivel de volatilidad y la multitud de eventos que pueden afectar su valor (Baur & Dimpfl, 2021).

Gracias a las ventajas mencionadas, los datos alternativos han sido aplicados en algunos casos para intentar identificar signos prematuros de cambios en el mercado. Sin embargo, en este trabajo para el periodo considerado y para el activo bajo análisis no se lograron obtener resultados significativos que avalen esta práctica. Es importante tener en consideración que existen numerosos factores macroeconómicos, de calidad de datos y de técnicas aplicadas que influyen en esta conclusión.

Otro uso potencial, es la aplicación de los datos alternativos para mejorar el manejo del riesgo en las inversiones. Parámetros como el volumen de comentarios o tráfico en redes sociales pueden ser indicadores para contrastar con el volumen de operaciones de un activo. De esta forma, se puede visualizar si una tendencia de precio es espuria o débil.

A pesar de los resultados obtenidos en este trabajo. El uso de los datos alternativos ya es aplicado por grandes grupos de inversión y considerado en algunos casos como un posible requerimiento para mantenerse competitivos (Dannemiller & Kataria, 2017). A pesar de esto, Dannemiller & Kataria tambien comentan que al evaluar el riesgo y beneficio de estos modelos es importance tener en cuenta varios factores como: la proveniencia de los datos, la atencion a la privacidad de los mismos, las posibles fallas en el modelo aplicado, los riesgos regulatorios entre otros. Sin embargo, dichos riesgos no limitan el interes y el estudio sobre como continuar aplicando los datos alternativos en los analisis bursatiles.

De acuerdo con lo expresado en los párrafos anteriores y a los resultados obtenidos en este trabajo. Resulta posible afirmar que los datos alternativos y los métodos que los utilizan representan un área de interés para determinar las variaciones en el precio del *Bitcoin*. Sin embargo, estos no resultan infalibles para todas las tareas en los que son usados. El uso de estos datos para tal fin presenta desafíos en diferentes campos como la selección del método de análisis correcto y la calidad de datos usados. Además, este se encuentra influido por factores externos como el contexto macroeconómico que puede resultar en tendencias de precio erráticas. En referencia a lo mencionado, diferentes recomendaciones serán realizadas en la conclusión del estudio a continuación

Conclusión

El presente trabajo permitió determinar la relación existente entre las opiniones expresadas por los usuarios de *Twitter* sobre el *Bitcoin* y la variación de precio de este criptoactivo. Esto se logró a través de la aplicación de técnicas de *Text Mining* sobre *tweets* que fueron posteriormente etiquetados a través de herramientas de análisis de sentimiento. El producto de este proceso de clasificación fue estudiado posteriormente utilizando el análisis de la causalidad de Granger y la correlación de las series de tiempo analizadas. La aplicación de estos métodos indico que no pudo probarse significativamente la existencia de una causalidad ni correlación fuerte entre ambas variables para el periodo en estudio.

En el primer capítulo se analizaron las organizaciones *Fintech*, el *Bitcoin* y su relación los datos alternativos. Para esto, se explicaron las características de las organizaciones *Fintech* y los fundamentos teóricos del *Bitcoin.* A continuación, se describió en qué consisten los grandes volúmenes de datos caracterizados por su alto volumen, variedad y velocidad. Las organizaciones *Fintech* tienden a ser agiles a la hora de incorporar nuevas tecnologías. De esta forma, se definieron los componentes de gestión y arquitectura necesarios para utilizar grandes volúmenes de datos alternativos en estas. También, se ahondo en los elementos éticos y de privacidad que derivan de la utilización de estos datos para la toma de decisiones.

En el segundo capítulo se describió el proceso para la extracción y procesamiento de los datos, luego se explicaron los métodos que harán uso de estos. Luego de extraer y procesar los datos, se aplicaron técnicas de *Text Mining* que permitieron etiquetar los *tweets* en base al sentimiento expresado por estos. A continuación, se explicaron las bases teóricas para el desarrollo de un modelo de clasificación de *tweets* que pueda ser utilizado en las organizaciones *Fintech* para cumplir esta función. Finalmente, se desarrollaron los aspectos teóricos de los métodos estadísticos a aplicar para determinar la correlación y causalidad entre las variables en estudio.

Por último, en el tercer capitulo se implementó y evaluó el modelo de clasificación de *tweets*. Este dio resultados positivos superando ampliamente al modelo *baseline* con un *accuracy* del 83.65%. A continuación, se evaluaron los resultados obtenidos de la aplicación del método de correlación de series de tiempo y causalidad de Granger. Esto permitió concluir

que en el periodo estudiado no se puede probar que exista una relación significativa entre el sentimiento expresado en los *tweets* y la evolución del precio del *Bitcoin*. Finalmente, en base al estudio realizado se indico la importancia de los datos alternativos para la determinación de la variación del precio del *Bitcoin*.

En relación con el alcance del trabajo, los datos utilizados fueron extraídos directamente de fuentes públicas como *Twitter* lo que permitió trabajar con datos crudos y desarrollar una metodología para el análisis de este tipo de datos en organizaciones. Adicionalmente, se desarrollo un método de clasificación que permite su rápida implementación de forma de relacionar los resultados de análisis de sentimientos con cualquiera variable que la organización considere. En relación con la capacidad del análisis de sentimientos para predecir el comportamiento del precio, los resultados indican que no resulta fiable para la toma de decisiones en base al periodo y métodos aplicados.

Como futuras líneas de investigación, resulta de interés la extensión del horizonte temporal analizado. También, la recopilación de un volumen mayor de *tweets* y la aplicación de otras técnicas para el etiquetado de estos. Estas técnicas podrían ser la categorización manual de los *tweets*. Esto ayudara a mejorar el desempeño de los algoritmos al contar con formas de interpretar expresiones propias de la comunidad del *Bitcoin* que pueden no estar contempladas en las herramientas de etiquetado automático.

Referencias bibliográficas

Aguilar, L. J. (2013). *Big Data, Analisis de grandes volumenes de datos en organizaciones.* Mexico: Alfaomega.

Agusti, C. M. (2018). Datos masivos y datos abiertos para una gobernanza inteligente. *El profesional de la informacion*, pp. 1128-1135.

Ah-Hwee, T. (2000). *Text Mining: The state of the art and the challenges.* Singapore: Kent Ridge Digital Labs.

Ammous, S. (2018). *El Patrón Bitcoin.* Barcelona, España: Editorial Planeta, S.A.

Arner, D., Barberis, J. N., & Buckley, R. (2015). *The Evolution of Fintech : A new Post-Crisis Paradigm?* Hong Kong: University Of Hong Kong Faculty of Law Research.

Badea, M. (2014). Social Media and Organizational Communication. *Procedia - Social and Behavioral Sciences*, pp. 70-75.

Baur, D., & Dimpfl, T. (2021). *The volatility of Bitcoin and its role as a medium of exchange and a store of value.* Empirical Economics.

Beltrame, S. N. (2020). *Grandes volumenes de datos y riesgo de credito: Tecnica de machine learning para el default en tarjetas de credito.* Universidad de Buenos Aires, Capital Federal.

Berrar, D. (2019). *Bayes´ Theorem and Naive Bayes Classifier.* Tokio Institute of Technology, Ookayama.

Broby, D., & Hopper, H. (2019). *Creating a Financial Data Lake for Academic Fintech Research.* Department of Acc, finance & Economics, Ulster University Business School.

Cleven, A., & Wortmann, F. (2010). Unconvering four strategies to approach master data management. In 2010 43rd Hawaii International Conference on System Sciences. Hawaii International Conference on System Sciences.

Dannemiller, D., & Kataria, R. (2017). Alternative data for investment decisions: Today's innovation could be tomorrow's requirement. *Deloitte Center for Financial Services*.

Devasena, L. (2014). Effectiveness Analysis of ZeroR, RIDOR and PART Classifiers for Credit Risk Appraisal. *Iternational Journal od Advances in Computer Science and Technology (IJACST)*, 6-11.

Di Maggio, M., Ratmadowakara, D., & Carmichael, D. (2022). *INVISIBLE PRIMES : FINTECH LENDING WITH ALTERNATIVE DATA.* Massachusetts: National Bureau of economic research.

Dreibelbis, A., Milman, I., Van Run, P., Hechler, E., Oberhofer, M., & Wolfson, D. (2008). *Enterprise Master Data Management : An SOA Approach to Managing Core Information.* IBM Press.

Eberendu, A. C. (2016). Unstructured Data : an overview of the data of Big Data. *Journal of emerging Trends & Technology in Computer Science*.

Eichengreen, B. J. (1996). *Globalizing capital: a history of the international monetary system.* Princeton University Press.

Fernandez, S., Rosillo, R., De la Fuente, D., & Priore, P. (2019). *BlockChain in FinTech : A Mapping Study.* Oviedo: Business Management Department, University of Oviedo.

Franco, P. (2014). *Understanding Bitcoin : Cryptography, Engineering and economics.* John Wiley & Sons.

Frenz, C. M. (2008). *Introduction to Searching with Regular Expressions.* New York: Proceedings of the 2008 Trenton. Computer Festival.

Ghaith , A., Alkubaisi, J., Sakira, S., & Husni, H. (2018). Stock Market Classification Modelo Using Sentiment Analysis on Twitter Based on Hybrid Naive Bayes Classifiers. *Computer and Information Science*, Vol. 11, No. 1.

Granger, C. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica,Volume 37*, 424-438.

Guizani, S., & Kahloul, I. (2019). The determinants of Bitcoin Price Volatility: An investigation with ARDL Model. *Procedia Computer Science*, 233-238.

Haneem, F., Kama, N., Azmi, A., Azizan, A., Mohd, S., Yusop, O., & Abas, H. (2017). *Master Data Definition and the Privacy Classification in Government Agencies : Case Studies of Local Government.* American Scientific Publishers.

Haug, A., & Stentoft, J. (2011). Barriers to master data quality. *Journal Of Enterprise Information Management*.

Hotho, A., Nurnberger, A., & Gerhard, P. (2015). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, 19-62.

Jiva, M. A. (2011). *A Comparative Study of Stemming Algorithms.* Department Of Computer Science & Engineering.

Johnson, K. (2019). *The future of finance : Alternative data in Credit Underwriting* .

Kalyani Joshi, Bharathi, H., & Jyothi, R. (2016). *Stock Trend Prediction Using News Sentiment Analysis.* ArXiv Prepint.

Khedr, A., Salama, S., & Yaseen, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *I.J Intelligent Systems and Applications*, 22-30. doi:10.5815/ijisa.2017.07.0

Leong, K., & Sung, A. (2018). Fintech:What is it and how to use Technologies to create business Value in Fintech Way? *International Journal Of Innovation, Management and Technology*, pp 74-78.

Martincevic, I., Crnjevic, S., & Klopotan, I. (2020). *Fintech Revolution in the Financial Industry*.

McAffe, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). *Big data : The management Revolution.* Harvard business review.

McAtter, C. (2014). *Twitter Sentiment Analysis to Predict Bitcoin Exchange Rate.* University of Dublin, Dublin.

Miskam , S., & Radin, E. (2018). *Big Data and Fintech in Islamic Finance: Prospects and Challenges.* Kuala Lumpur.

Moffat, I., & Akpan, E. (2019). White Noise Analysis: A measure of Time Series Model Adequacy. *Applied Mathematics 10*, 983-1003.

Nakamoto, S. (2008). *Bitcoin.org.* Obtenido de Bitcoin : A Peer-to-Peer Electronic Cash System: https://bitcoin.org/bitcoin.pdf

Narendra, B., Sai, K., Rajesh, G., Hermanth, K., Teja, M., & Kumar, K. (2016). Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies. *International Journal of Intelligent Systems and Applications*, 66-70.

Petrit, S., & Kuql, B. (2019). Analysis of Financial Statements: The importance of Financial Indicators in Enterprise. *Humanities and Social Science Research*. doi:https://doi.org/10.30560/hssr.v2n2p17

Petrusheva, N., & Jordanoski, I. (2016). Comparative Analysis Between the fundamentel and technical analysis of stocks. *Journal of Process Management – New Technologies, International*.

Roesslein, J. (2020). *Tweepy: Twitter for Python!* Obtenido de https://www.tweepy.org/

Russom, P. (2015). Modernizacion de la integracion de datos para dar cabida a los nuevos requisitos de negocio y Big Data. *TDWI*.

Salaberry, N. R. (2019). *Deteccion de problematicas en el uso de la tarjeta SUBE. Un analisis y clasificacion de tweets.* Buenos Aires.

Sarker, I. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI*, 160.

Sattarov, O., Jeon, O., Oh, R., & Lee, J. (2020). "Forecasting Bitcoin Price Fluctuation by Twitter Sentiment Analysis. *2020 International Conference on Information Science and Communications Technologies (ICISCT)*, pp. 1-4.

Schmarzo, B. (2013). *Big Data: Understanding how data powers big business.* John Willey & Sons.

Sebastiao, H., Rupino, P., & Godinho, P. (2021). *Cryptocurrencies and blockchain. Overview and future perspectives.* Inderscience Publishers.

Shabunina, E. (2013). CRF to find stock price correlation with company-related Twitter Sentiment. *(Master Of Science in Computer Engineering).* POLITECNICO DI MILANO, Milano.

Smith, H. A., & McKeen, J. D. (2007). *Developments in practice XXIV: information management: the nexus of business and IT*. Communications of the Association for Information Systems.

Smith, H., & McKeen, J. (2008). Developments in Practice XXX: Master Data Management Salvation or Snake oil? *Communications of the association for information systems*, Vol.23 pp 63-72.

Smith, S., & O'Hare, A. (2022). Comparing traditional news and social media with stock price movements: Which comes first, the news of the price change? *Smith and O'Hare journal of Big Data*. doi:https://doi.org/10.1186/s40537-022-00591-6

Utami Handika, S., Ade Purnama, A., & Nizar Hidayanto, A. (2022). Fintech Lending in Indonesia: A Sentiment Analysis, Topic Modeloling, and Social Network Analysis using twitter Data. *International Journal of Applied Engineering & Technology*, Vol4 No1.

Vishal, G., & Gurpreet Lehai. (2009). A survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence* .

Vives, X. (2017). *The Impact of Fintech on Banking.* Madrid: IESE Business School.

Vučinić, M. (2022). Fintech And Financial Stability Potential Influence of Fintech on Financial Stability, Risks and Benefits. *Journal of Central Bnaking Theory and Practice*, pp. 43-66.

Wanbin Wan, W., Kin-yip, H., Wai-Man, R., & Wang, K. (2013). *The relation between news events and stock price jump: an analysis based on neural network.*

Wang, W., Ho, K.-Y., Liu, W.-M., & Wang, K. (2013). *The relation between news events and stock price jump: an analysis based on neural network.* Adelaide: 20th international Congress on Modelling and Simulation.

EVALUACION DE MENTORA

Trabajo Final de Especialización de Carlos Leotaud:
"Análisis de sentimiento de expresiones en Twitter para la predicción del comportamiento del precio del Bitcoin en una organización Fintech. Implementación de técnicas de Text Mining."
Mentora: Natalia Salaberry

Fecha: febrero 2023

El alumno Carlos Leotaud en su versión final de trabajo final de especialización establece de manera clara y precisa, la pregunta de investigación, el objetivo y la hipótesis general que guiaran el desarrollo de este. Plantea de manera clara la existencia de una necesidad en un tipo de organización que luego resolverá con las herramientas aprendidas en el campo diciplinar de la Especialización en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones. En este sentido, el problema propuesto a resolver se vincula con determinar el valor que aporta el procesamiento y análisis de datos alternativos para la determinación del comportamiento del precio del *Bitcoin* en una organización *Fintech*.

Para cumplimentar con el objetivo planteado, en un primer capítulo desarrolla una clara contextualización de las implicancias de la gestión de datos en organizaciones *Fintech*. A partir de este punto, describe la situacionalidad en base a las oportunidades y desafíos que presenta la disponibilidad de grandes volúmenes de datos alternativos. Pero también advierte sobre la necesidad e importancia de una gestión adecuada de estos para lograr la construcción de valor agregado para la toma de decisiones.

En un segundo capítulo, se centra en exponer de manera concisa los datos con los cuales trabajará, así como las herramientas metodológicas a utilizar y que fueron aprendidas en diversas materias de la especialización. Desde el aspecto técnico de implementación de modelos con lenguaje de programación, realiza un trabajo perfecto, respetando todas las etapas de aplicación, análisis y evaluación de resultados. Supo seleccionar adecuadamente los modelos acordes que le permitieron cumplimentar con el objetivo buscado.

Finalmente, en un tercer capítulo expone como los resultados obtenidos permiten generar valor agregado para la toma de decisiones a partir de desarrollar un análisis de manera completa. La evaluación y análisis de resultados obtenidos es lo que le permite identificar la necesidad de tener en cuenta otros factores y no ya solo el sentimiento de los usuarios para poder explicar la volatilidad del precio del *Bitcoin*. De este modo, se aproxima parcialmente al cumplimiento de su objetivo principal, estableciendo la necesidad de ampliar el alcance de trabajo. Esto, es realizado en completo detalle.

De esta manera, el alumno logra cumplimentar con las etapas necesarias para desarrollar el objetivo principal y los específicos de forma ordenada y clara, realizando un trabajo completamente estructurado en el marco de las pautas establecidas. Existe una coherencia entre la problemática planteada, el título y las palabras claves lo que le permitió poder elaborar un planteo y desarrollo adecuado.