

2023 - 2024

# 

**Máster en Big Data y Data Science**

**05MBID – Minería de datos**

**Nombre: Brahian Andrey Giraldo Alzate**

**Fecha: 06/11/2023**

**Curso 2022 – Ed. Abril**

**ACTIVIDAD GUIADA 1**

Contenido

[1](#_Toc137419356)

[1 Introducción, motivación y objetivo (Entendimiento del dominio) 3](#_Toc137419357)

[1.1 Introducción 3](#_Toc137419358)

[1.2 Motivación 3](#_Toc137419359)

[1.3 Objetivo 3](#_Toc137419360)

[2 Fuente de datos y selección de datos 5](#_Toc137419361)

[2.1 Fuente de datos 5](#_Toc137419362)

[2.2 Selección de los datos 6](#_Toc137419363)

[3 Preparación, limpieza y transformación 9](#_Toc137419364)

[3.1 Preparación 9](#_Toc137419365)

[3.2 Limpieza 11](#_Toc137419366)

[3.3 Transformación 11](#_Toc137419367)

[4 Data mining y método de evaluación. 13](#_Toc137419368)

[4.1 Análisis de sentimiento: 13](#_Toc137419369)

[4.2 Modelos de pronósticos: 13](#_Toc137419370)

[4.3 Método de evaluación: 15](#_Toc137419371)

[5 CONCLUSIONES 15](#_Toc137419372)

[6 ANEXOS 16](#_Toc137419373)

[6.1 Anexo 1: visualización de los campos existentes en la base de datos descargada desde el datalake de twitter. 16](#_Toc137419374)

# Introducción, motivación y objetivo (Entendimiento del dominio)

En esta etapa, se busca comprender el contexto y los objetivos del proyecto de minería de datos. Se exploran las necesidades del negocio, se definen los objetivos del análisis y se establecen las preguntas clave que se quieren responder.

## Introducción

En la fase inicial del proceso de minería de datos de este caso de estudio, se lleva a cabo la selección del objetivo y un análisis preliminar de los datos. En esta etapa, se plantean preguntas clave para definir el problema a resolver y cómo los datos pueden ayudar a los traders.

Se exploran los procesos manuales existentes y se considera su automatización mediante el análisis de datos. También se abordan cuestiones prácticas, como la obtención de datos, evaluación de calidad y procesamiento inicial.

En el contexto de las redes sociales, el análisis de los tweets puede proporcionar información valiosa sobre el mercado y guiar decisiones de inversión al ofrecer perspectivas sobre opiniones, sentimientos y tendencias en tiempo real. A su ves el análisis del mercado realizado debe ir de la mano con un análisis técnico fiananciero del valor de las acciones al cual se le pueden aplicar distintos modelos de machine learning para pronosticar valores o tendencias futuras.

## Motivación

El análisis de los tweets puede complementar otras fuentes de información y herramientas de análisis utilizadas en el mercado financiero, como el análisis fundamental y técnico, que se enfoca en el análisis de los precios en la bolsa de valores para predecir su comportamiento e inferir probabilidades a favor del inversor.

## Objetivo

Con respecto al análisis de la información descargada de esta red social, el objetivo principal es determinar tendencias de inversión en el mercado según la base de datos disponible en mongodb, aplicando técnicas de procesamiento del lenguaje natural, como la detección de temas, análisis de sentimiento o, el análisis de la intención, que permitirán extraer información valiosa sobre los productos, servicios o empresas mencionados en los tweets. A aquellos objetivos de inversión validados por el proceso NLP se les aplicará un análisis de pronóstico con el promedio móvil como subobjetivo del caso de estudio junto con una validación de la tendencia mencionada en le tweet relacionado al a fecha correspondiente.

**¿Qué problema se quiere resolver?**

En este caso de estudio, se busca automatizar el análisis del mercado en redes sociales para ayudar a los traders a tomar decisiones de inversión más fundamentadas en cuanto a qué empresa o nicho elegir para invertir en el mercado de valores.

El objetivo es utilizar la información y tendencias obtenidas de las redes sociales para seleccionar mejores portafolios de inversión de forma más informadas y potencialmente exitosa. Después de ello, se validan los resultados de la tendencia mencionada con análisis de tendencia con promedio móvil o regresión lineal.

**¿Cómo se puede ayudar al trader a generar conocimiento?**

Con la automatización del análisis de redes sociales de trading para la selección de portafolios de inversión. Los traders que se enfocan en el análisis técnico suelen utilizar herramientas y gráficos para identificar patrones, tendencias y señales de compra o venta. Estas herramientas pueden ayudar a automatizar **parte del análisis** y a generar alertas cuando se cumplan ciertos criterios predefinidos.

**¿Qué proceso manual se está realizando que se podría realizar de manera automática a través de los datos?**

Algo muy beneficioso del trading algorítmico es que se pueden automatizar las cuestiones referentes a los sentimientos, evitando diversas tareas como la **técnicas de análisis de sentimiento para determinar de forma automática el comportamiento del mercado teniendo en cuenta tendencias bajistas o alcistas.**

**Análisis de sentimiento:** Utiliza técnicas de NLP para analizar el sentimiento expresado en los tweets o noticias relacionadas con una empresa o mercado en particular. Por ejemplo, si la mayoría de los tweets o noticias sobre una empresa son positivos, esto puede indicar un sentimiento alcista, mientras que un sentimiento negativo podría señalar un sentimiento bajista.

Suele ser mejor invertir en empresas que están creciendo, aunque no sean muy grandes, ya que lo que genera ganancias a los inversores es el crecimiento de las empresas.

**Detección de temas:** Utiliza técnicas de NLP para identificar los temas y las discusiones más relevantes en los tweets relacionados con una empresa o un mercado. Esto te permitirá comprender mejor los temas de interés para los inversores y seguir de cerca los eventos o noticias que podrían afectar el mercado.

Un ejemplo del valor agregado que se puede generar en el caso de análisis de sentimiento es el siguiente tweet de Jim Cramer:

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Donde aparentemente muestra que tiene intenciones de invertir en apple. Se puede hacer un análisis de sentimiento para validar si es el caso de una tendencia alcista.

De la misma forma se pueden añadir distintas bases de datos de diferentes entidades con yahoo finance, que cumplan los criterios de análisis en twitter para realizar análisis de tendencia y transformación de variables con respecto a los precios de yahoo finance.

# Fuente de datos y selección de datos

En esta etapa, se identifican las fuentes de datos relevantes para el análisis. Se determina qué datos se necesitan y se recopilan de diferentes fuentes, como bases de datos, archivos, APIs, etc.

## Fuente de datos

Se eligieron 10 cuentas de Twitter, de personas influyentes en el área del trading para realizar un estudio de los datos y analizar el mercado en diversos nichos de inversión. Estos podrían ayudar a los traders a analizar el mercado en la bolsa de valores:

@Schuldensuehner - Holger Zschaepitz: Reportero financiero y editor de mercados de Die Welt.

@elerianm - Mohamed A. El-Erian: Asesor económico, autor y exdirector ejecutivo de PIMCO.

@tracyalloway - Tracy Alloway: Reportera financiera de Bloomberg News.

@SJosephBurns - Steve Burns: Trader a tiempo completo y autor de varios libros sobre trading.

@jimcramer - Jim Cramer: Presentador de "Mad Money" en CNBC, inversionista y autor de varios libros sobre inversión.

@muddywatersre - Muddy Waters Research: Una firma de investigación independiente que investiga empresas y proporciona análisis de inversión.

@SamanthaLaDuc - Samantha LaDuc: Analista de mercado y estratega de inversión.

@SJosephBurns - Steve Burns: Trader a tiempo completo y autor de varios libros sobre trading.

@RayDalio - Ray Dalio: Fundador de Bridgewater Associates, una de las firmas de inversión más grandes del mundo.

@NorthmanTrader - Sven Henrich: Analista de mercado, estratega y trader independiente.

@Traderstewie es una cuenta de Twitter que comparte análisis técnico del mercado de valores y criptomonedas, así como estrategias de trading, noticias y opiniones sobre el mercado financiero en general.

Los datos se sacaron del datalake de twitter con la APPI.

**¿Es suficiente con una fuente de datos?**

Para el objetivo principal, que es determinar las tendencias alcistas o bajistas a través del análisis de los tweets es suficiente con la fuente de datos de twitter. Sin embargo es conveniente analizar la tendencia de los precios con métodos de regresión con una fuente de datos más precisa, en este caso, la fuente de datos de yahoo finance ayudará a validar dicha tendencia para determinar si el modelo está funcionando correctamente (este sería un objetivo secundario).

Entonces se tiene una tabla de datos de yahoo finance con las columnas Date, Open, High, Low, Close, Adj Close y finalmente Volume.

Otra tabla, o colección en este caso es la proveniende del datalake de twitter. El datalake de twitter, permitió descargar 158 campos (atributos o columnas). Dichos campos se describen y se da un ejemplo de cada uno de ellos en el anexo 1, debido a su gran extensión, no es viable mencionarlos acá.

La base de datos de mongodb, tiene 8.000 documentos (filas),

**Analizar si los datos iniciales se encuentran procesados de alguna manera**

Los datos extraídos de la API de Twitter generalmente no están procesados de forma predeterminada. La API de Twitter proporciona acceso a datos en tiempo real y a una amplia gama de información disponible en la plataforma, como tweets, usuarios, tendencias, menciones y más. Sin embargo, los datos en bruto obtenidos de la API no están procesados ni analizados de manera automática.

**¿Cuántos registros hay?**

Son 8.000 registros (filas o documentos) en la colección descargada del datalake de twitter y almacenada en mongodb. 2.000 registros (filas), por cada empresa o acción a analizar. Para esto se identificarán las empresas que se mencionan en los textos de twitter, y cada mención de una empresa diferente, implicaría una base de datos nueva a agregar en el análisis. Se estima que en el análisis se pueden incluir al menos 10.000 registros (5 empresas) para añadir al portafolio de inversión y realizar sus respectivos análisis.

En este análisis solo se incluyen aquellas empresas que se mencionaron en el año 2023.

**– ¿Son demasiado pocos?¿hay capacidad suficiente para el procesamiento?¿hay valores núlos?**

Para el objetivo del caso de estudio es más que suficiente, porque lo que se necesita en la vida real es que se puedan descargar los datos de forma rápida para tomar decisiones en tiempo real, por lo cual, los datos útiles siempre serán los de los últimos días. Se tiene suficiente capacidad para el procesamiento

Normalmente los valores null, se encuentran en campos que no se están utilizando en el análisis, por lo cual, cuando se encuentra un campo con valor null, la información sigue siendo útil.

## Selección de los datos

**¿Qué datos son discretos y cuáles continuos?**

A continuación, se clasifica los tipos de datos en la base de datos actual descargada de twitter.

Son 158 campos diferentes, de los cuales no se utilizarán todos, por lo cual se clasifican a continuación los más útiles que son tipo int (valores discretos) y el resto de campos se pueden visualizar en el anexo 1.

{

followers\_count: 450887,

friends\_count: 72,

favourites\_count: 40595,

retweet\_count: 6,

favorite\_count: 49,

}

followers\_count: Indica la cantidad de seguidores que tiene la cuenta.

friends\_count: Representa el número de amigos o cuentas a las que sigue la cuenta.

favourites\_count: Muestra la cantidad de tweets que la cuenta ha marcado como favoritos.

retweet\_count: Indica el número de veces que los tweets de la cuenta han sido retuiteados por otros usuarios.

favorite\_count: Representa la cantidad de veces que los tweets de la cuenta han sido marcados como favoritos por otros usuarios.

Los anteriores **campos son necesarios para identificar el nivel de credibilidad que tiene la cuenta, asegurando así, un nivel de credibilidad determinado** de los tweets.

No se observaron valores tipo float o double útiles (valores continuos) en la base de datos.

Los valores **tipo texto o string** que son útiles para el caso de estudio son los siguientes:

{

created\_at: 'Tue May 09 16:46:57 +0000 2023',

text: '$34.4 trillion index price discovery. https://t.co/6PefA9mjWz',

name: 'Sven Henrich',

screen\_name: 'NorthmanTrader',

location: 'England, United Kingdom',

description: 'Founder: NorthmanTrader. Financial Market Strategist. Macro & Technical Analysis. Keeping it real. Subscribe: https://t.co/rmRFFvMRaH…',

url: 'https://t.co/2rOR1Gp0G6',

created\_at: 'Tue Jul 24 10:12:44 +0000 2012',

}

created\_at: Indica la fecha y hora en que se creó el tweet. Proporciona información sobre el momento exacto en que se publicó el tweet. Permite alinear las fechas de registro del valor de la acción de la fuente de datos de yahoo con el datalake de twitter.

text: Contiene el texto del tweet en sí. Este atributo es importante porque es donde se aplica el proceso NLP de análisis de sentimiento, tema o intención.

name: Representa el nombre asociado a la cuenta de Twitter. Es importante para identificar al sujeto que proporciona la información para evaluar mejor la credibilidad.

screen\_name: Es el nombre de usuario único asignado a la cuenta en Twitter. Es importante para ubicar la cuenta en twitter y validar distintas entradas de datos.

location: Indica la ubicación especificada por el propietario de la cuenta.

description: Proporciona una breve descripción o biografía del propietario de la cuenta. Permite fundamentar la credibilidad de la cuenta basada en el historial de conocimiento del propietario.

url: Representa una URL asociada a la cuenta. Permite ubicar la cuenta en la web.

created\_at: Indica la fecha y hora en que se creó la cuenta de Twitter. La fecha en la que se creó la cuenta es un importante indicador de credibilidad de la misma.

A continuación, con 7 campos diferentes, se clasifican los tipos de datos que se pueden encontrar en la base de datos descargada de yahoo finance:

Valores discretos:

* Volume: 122207100

Valores continuos (tipo float o double):

* Open: 132.869995
* High: 135.199997
* Low: 131.440002
* Close: 131.880005
* Adj Close: 131.097198

Valores tipo texto:

* Date: 2022-06-13

Date: Esta columna representa la fecha en la que se registraron los datos de precios. Permite alinear la fecha de inversión con la fecha de sugerencia de tendencia alcista en twitter.

Open: muestra el precio de apertura de las acciones en el inicio del día de negociación.

High: indica el precio más alto alcanzado por las acciones de la empresa durante el día.

Low: La columna "Low" muestra el precio más bajo registrado por las acciones de la empresa durante el día.

Close: representa el precio de cierre de las acciones de la empresa al final del día de negociación.

Adj Close: se refiere al precio de cierre ajustado. A veces, los precios de cierre se ajustan para tener en cuenta eventos como dividendos, desdoblamientos de acciones u otras acciones corporativas. Permite calcular los retornos históricos con mayor precisión. Permite determinar los valores históricos a partir de los cuales se crea la nueva variable predictiva.

Volume: indica el volumen de acciones negociadas durante el día. Representa la cantidad total de acciones compradas y vendidas.

Todos los datos se usan para desarrollar un diagrama de velas de la cotización de los valores de la bolsa que permita ver en qué precio se inició la primera venta en el periodo determinado, en que precio se cerró la última venta, precios máximos y mínimos y también volumen de datos, todo en una misma gráfica, cuyo ejemplo de visualización sería el siguiente:

# Preparación, limpieza y transformación

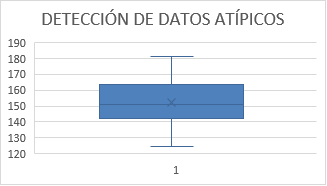
## Preparación

**¿Cuáles parecen ser features importantes? ¿Cuáles podemos descartar? ¿Por qué crees que tienes que descartar estas características?**

Los campos señalados anteriormente son los features importantes, porque son los que en principio generan valor agregado al caso de estudio. El resto de campos que no coinciden con los campos resaltados y que se ubican en el anexo 1 se pueden descartar, porque no generan valor agregado al trabajo actual. Esto permite simplificar y mejorar el análisis, aumentar la precisión de los modelos y facilitar la interpretación de los resultados.

¿**Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y “contaminan” ó desvían las distribuciones) – Podemos eliminarlos? ¿es importante conservarlos? – son errores de carga o son reales?**

En el siguiente gráfico, con respecto al rango máximo y mínimo de los datos, podemos observar que de momento no hay ningún punto, valor de las acciones o dato fuera de los cuartiles esperados.



Se hizo el análisis 3-sigma, para descartar valores atípicos entre todos los datos, pero el análisis indica que no hay ningún dato a una distancia de 3 desviaciones de la media. En caso de encontrar un dato atípico se reemplazaría por el promedio entre el valor de la acción el día anterior y el día siguiente.

No se han observado valores faltantes en los campos o atributos elegidos para cumplir con el objetivo de investigación.

**Integración de datos:** se deben hacer un join entre los datos que coincidan con las fechas mencionadas. Aquellas fechas de los valores de las acciones registrados en yahoo finance que no tengan una mención en las cuentas de twitter que indiquen intención de inversión, se tomarán como objetivo de inversión no válido según el criterio de análisis de tendencias alcistas en el mercado.

**¿Hay correlación entre features (características)?**

La idea del caso de estudio con respecto a la combinación de la base de datos de twitter con la de yahoo finance, es precisamente, determinar en qué casos existe esta correlación. Por lo cual se pueden analizar casos determinados y ver si cuando se afirma que es recomendable invertir en una acción, tiene coherencia con las tendencias que presenta el valor de las acciones en yahoo finance.

**¿Estamos ante un problema dependiente del tiempo? Es decir, un TimeSeries**

Si, el valor de las acciones que registra yahoo finance, depende del tiempo y de otros factores, pero por lo regular, el valor siempre depende del día, e inclusive de la hora en la que se definen los precios. Si se realiza un análisis más profundo, se puede identificar que los análisis pueden ser más volátiles, buscando así variaciones en periodos de tiempo en minutos o segundos inclusive.

**¿Tenemos posible sesgo de datos? (por ejemplo, perjudicar a clases minoritarias por no incluirlas y que el modelo de ML discrimine)**

## Limpieza

Se eliminarán de la colección, aquellos datos que no correspondan a los datos campos mencionados anteriormente.

Si se llegase a encontrar un dato atípico en el valor de las acciones, se reemplazaría por el promedio entre el valor de la acción el día anterior y el día después.

Aquellos valores de las acciones que no tengan mención en una fecha específica correspondiente en twitter, obtendrán un nuevo campo que indique que la inversión en ese periodo no debería ser aconsejable según el criterio de análisis de tendencias alcistas o bajistas. El hecho de que una acción determinada no sea recomendada en una fecha determinada, no indica que sea mala opción de inversión, simplemente indica que

## Transformación

**¿Puedo transformar alguna variable para generar información adicional?**

Las fechas de la colección de mongodb están en formato texto, por lo cual se deberán pasar a formato fecha para operar con los tiempos en caso de ser necesario.

Los valores de la base de datos de yahoo finance, se pueden transformar para generar información adicional. Para este caso de estudio, se utilizará el precio del campo Adj close, puesto que es el precio destinado a calcular retornos históricos con mayor precisión, puesto que ha recibido ajustes correspondientes a diferentes eventos que no se tienen en cuenta en los demás campos, como dividendos, desdoblamiento de acciones, etc.

Este valor se puede transformar a una variable predictiva a través de una granvariedad de modelos de pronóstico. Cada modelo de pronóstico puede arrojar un resultado distinto, por lo cual a partir de la columna Adj close se pueden derivar numerosas variables objetivo para hacer forecasting.

Se pueden llevar a cabo procesos de binning o discretización para identificar patrones no lineales o para simplificar el análisis en ciertos casos, por ejemplo, con patrones mariposa, que son muy conocidos en el trading.

**¿Conozco conocimiento experto para incorporarlo en el dataset? ¿Cómo lo puedo modelar?**

En el caso de los modelos de predicción con respecto a los datos históricos de la columna Adj close, tengo el conocimiento para aplicar diferentes modelos de predicción. Entre ellos el análisis de promedio móvil es una opción de análisis muy frecuente en el trading, y el patrón de mariposa es de alto interés para mi, aprender como desarrollarlo.

En el caso del promedio móvil se deben realizar los siguientes pasos para desarrollar el modelo:

* **Selección de un período de tiempo:** Se debe determinar el período de tiempo que se utilizará para el análisis. Por ejemplo, se puede elegir utilizar datos diarios, semanales o mensuales, dependiendo de las necesidades y el horizonte de pronóstico. Para este caso de estudio aplicaríamos un periodo diario.
* **Cálculo del promedio móvil:** Se aplica el método del promedio móvil para suavizar los datos y identificar tendencias. El promedio móvil se calcula sumando los valores de un número determinado de períodos y dividiendo el resultado entre ese mismo número. Por ejemplo, para un promedio móvil de 5 días, se sumarían los valores de los últimos 5 días y se dividirían entre 5.
* **Representación gráfica del promedio móvil:** Se debe graficar el promedio móvil junto con los datos de precios de las acciones. Esto ayudará a visualizar las tendencias y patrones en los datos.
  + Visualmente, la gráfica tendría un aspecto parecido al siguiente:
    - Para el caso de predicción del valor de las acciones, en vez se usaría el valor de las acciones, y se añadiría una tercera gráfica correspondiente al cálculo de los valores futuros que permitirán al trader tomar decisiones de acuerdo a esta estrategia. La gráfica roja, indica los valores del promedio móvil, y la gráfica azul lo valores reales.
* **Interpretación de los resultados:** Se analiza la relación entre los valores de los precios de las acciones y el promedio móvil. Si el precio de las acciones cruza por encima del promedio móvil, puede ser una señal alcista, mientras que si cruza por debajo, puede ser una señal bajista.
* **Realización de pronósticos:** Se utiliza el promedio móvil para hacer pronósticos futuros del valor de las acciones. Se puede extrapolar la tendencia identificada y estimar posibles valores futuros.

**¿Siguen alguna distribución?**

Si, La distribución de los precios de las acciones puede variar según la empresa y el período de tiempo considerado. Sin embargo, es común observar una distribución asimétrica con colas pesadas, lo que refleja la naturaleza volátil y dinámica del mercado de valores.

# Data mining y método de evaluación.

## Análisis de sentimiento:

El análisis de sentimiento de un texto es una técnica utilizada para determinar la actitud emocional expresada en un texto, generalmente en términos de positividad, negatividad o neutralidad. Para realizar dicho análisis se definieron los siguientes pasos:

1. Preprocesamiento de texto: Limpiar y procesar el texto para eliminar ruido y elementos irrelevantes, como signos de puntuación, números, stopwords (palabras comunes sin significado específico) y convertir el texto en minúsculas.
2. Tokenización: Dividir el texto en unidades más pequeñas llamadas tokens, como palabras individuales o frases cortas.
3. Construcción de un diccionario de sentimientos: Crear un diccionario que asocie palabras o frases con un valor de sentimiento predefinido, como positivo, negativo o neutral. Esto se conoce como un lexico de sentimientos.
4. Asignación de puntuaciones de sentimiento: Asignar puntuaciones de sentimiento a cada token del texto en función de las palabras o frases presentes en el diccionario de sentimientos. Estas puntuaciones pueden ser valores numéricos, binarios o categóricos.
5. Agregación de puntuaciones: Calcular una puntuación agregada para todo el texto sumando o promediando las puntuaciones de sentimiento de los tokens individuales. Esto proporciona una medida general del sentimiento expresado en el texto.
6. Análisis e interpretación: Interpretar la puntuación agregada para determinar si el texto tiene un sentimiento positivo, negativo o neutral. También se pueden realizar análisis adicionales, como identificar las palabras o frases más influyentes en el sentimiento expresado.

Cálculos matemáticos asociados:

Puntuación de sentimiento promedio: Se calcula sumando todas las puntuaciones de sentimiento de los tokens y dividiendo entre el número total de tokens.

Porcentaje de sentimiento positivo/negativo: Se calcula dividiendo el número de tokens con sentimiento positivo/negativo por el número total de tokens y multiplicando por 100.

Frecuencia de palabras positivas/negativas: Se calcula contando el número de palabras con sentimiento positivo/negativo presentes en el texto.

Puntuación de sentimiento ponderada: Se puede asignar un peso a las puntuaciones de sentimiento de acuerdo con la importancia relativa de las palabras en el texto. La puntuación ponderada se calcula multiplicando cada puntuación de sentimiento por su peso correspondiente y sumando los productos resultantes.

## Modelos de pronósticos:

Se define el modelo de promedio móvil como predictor que permitirá validar la tendencia alcista o bajista del mercado mencionada en las cuentas de twitter. Adicionalmente, para asegurar que se cumplen los requisitos del modelo seleccionado se escoge un modelo propuesto en el pdf guía de la actividad.

Los modelos de pronósticos suelen hacer parte de un sistema que permite el análisis de los datos, en este caso, se ve de la siguiente forma:

Diagrama, Escala de tiempo

Descripción generada automáticamente

De lo cual se puede observar que de las fuentes de datos se extraen los datos necesarios. De yahoo finance los valores de las acciones de distintas entidades u activos financieros y de twitter appi, se extraen los valores de los campos que permiten validar la credibilidad del tweet y analizar el texto con un análisis de sentimiento para determinar si existe intención de compra o inversión en la acción a analizar.

Podría ser interesante agregar un análisis de demanda de las acciones basado en el volumen de los datos proporcionado por yahoo finance, ya que esto nos permitirá predecir a partir de la demanda de acciones cuando subirán y cuando bajarán los precios. Con esto se lograría la integración de dos enfoques distintos de pronóstico al mismo tiempo en las fechas correspondientes al análisis.

**Promedio móvil**

El uso del promedio móvil en el entorno de los algoritmos en Big Data puede ayudar a identificar patrones, realizar pronósticos y tomar decisiones basadas en el comportamiento histórico de la serie temporal analizada.

*=*

**La regresión lineal**

Es un algoritmo de aprendizaje supervisado que se utiliza en Machine Learning y en estadística. En su versión más sencilla, lo que haremos es “dibujar una recta” que nos indicará la tendencia de un conjunto de datos continuos (si fueran discretos, utilizaríamos Regresión Logística).

El modelo matemático para este caso sería el siguiente:

*t =*

Los parámetros a y b se estiman mediante el método de mínimos cuadrados. Consiste en encontrar los valores para los parámetros a y b, tal que la suma de los errores al cuadrado sea mínima.

Para el cálculo de los errores del pronóstico se suelen utilizar los siguientes indicadores:

ERROR PROM. (Error Promedio): Es la diferencia promedio entre el pronóstico y el valor real de la demanda. Se calcula sumando todos los errores individuales y dividiéndolos por el número total de observaciones. Proporciona una medida general de la dirección y magnitud del sesgo en los pronósticos.

MAD (Desviación Absoluta Media): Es la media de los valores absolutos de los errores entre el pronóstico y el valor real de la demanda. Calcula la magnitud promedio de los errores, sin tener en cuenta su dirección. Cuanto menor sea el valor de MAD, mejor será la precisión del pronóstico.

ECM (Error Cuadrático Medio): Es la raíz cuadrada de la media de los errores al cuadrado entre el pronóstico y el valor real de la demanda. El ECM tiene en cuenta tanto la dirección como la magnitud de los errores, y penaliza los errores grandes. Cuanto menor sea el valor de ECM, mejor será la precisión del pronóstico.

MAPE (Error Porcentual Absoluto Medio): Es el promedio de los errores porcentuales absolutos entre el pronóstico y el valor real de la demanda. Se calcula como el promedio de los valores absolutos de los errores porcentuales, dividido por el número total de observaciones, y se expresa como un porcentaje. El MAPE permite evaluar la precisión del pronóstico en términos de porcentaje y es útil para comparar la precisión entre diferentes series temporales.

MAPE' (Error Porcentual Absoluto Medio Ajustado): Es similar al MAPE, pero ajusta el cálculo para evitar divisiones por cero cuando los valores reales son cercanos a cero. El MAPE' es especialmente útil cuando se trabaja con series temporales que tienen valores cercanos a cero o valores negativos.

## Método de evaluación:

Para evaluar si el modelo del análisis de sentimiento obtuvo resultados confiables, se seleccionan los datos correspondientes a las fechas elegidas en la mención del tweet caso de estudio, y se valida que la tendencia mencionada corresponda con la tendencia de los datos mostrada en el análisis de promedio móvil o de regresión lineal.

# CONCLUSIONES

Revisando las posibilidades con respecto al aporte del big data en la vida diaria de los traders, se puede observar la posibilidad de automatizar muchas tareas diarias que permite a los traders enfocarse en aquello que realmente les genera ganancias sin dejar de lado el hecho que siempre deben tomar decisiones informadas con respecto al mercado, el análisis técnico enfocado al crecimiento de las empresas.

Es muy importante asegurar la calidad de los datos, para obtener resultados confiables.

La combinación de varias fuentes de datos puede generar un valor agregado inesperado, puesto que incluye en los análisis, perspectivas diferentes sobre el mismo problema.

Las redes sociales a pesar de tener información cuestionable en algunos casos, puede ser útil y confiable si se valida la calidad de la información generada a través de análisis de credibilidad del texto, por ejemplo.

# ANEXOS

## Anexo 1: visualización de los campos existentes en la base de datos descargada desde el datalake de twitter.

Los siguientes campos corresponden a un documento de la colección de tweetsfinancieros descargados con la APPI de twitter.

{

\_id: ObjectId("645a7dc7f55dd5c3370166e6"),

created\_at: 'Tue May 09 16:46:57 +0000 2023',

id: 1655977667215097900,

id\_str: '1655977667215097862',

text: '$34.4 trillion index price discovery. https://t.co/6PefA9mjWz',

truncated: false,

entities: {

hashtags: [],

symbols: [],

user\_mentions: [],

urls: [],

media: [

{

id: 1655977590501277700,

id\_str: '1655977590501277708',

indices: [

38,

61

],

media\_url: 'http://pbs.twimg.com/media/Fvs22EXWAAwNh8e.jpg',

media\_url\_https: 'https://pbs.twimg.com/media/Fvs22EXWAAwNh8e.jpg',

url: 'https://t.co/6PefA9mjWz',

display\_url: 'pic.twitter.com/6PefA9mjWz',

expanded\_url: 'https://twitter.com/NorthmanTrader/status/1655977667215097862/photo/1',

type: 'photo',

sizes: {

thumb: {

w: 150,

h: 150,

resize: 'crop'

},

medium: {

w: 1200,

h: 640,

resize: 'fit'

},

large: {

w: 2048,

h: 1092,

resize: 'fit'

},

small: {

w: 680,

h: 363,

resize: 'fit'

}

}

}

]

},

extended\_entities: {

media: [

{

id: 1655977590501277700,

id\_str: '1655977590501277708',

indices: [

38,

61

],

media\_url: 'http://pbs.twimg.com/media/Fvs22EXWAAwNh8e.jpg',

media\_url\_https: 'https://pbs.twimg.com/media/Fvs22EXWAAwNh8e.jpg',

url: 'https://t.co/6PefA9mjWz',

display\_url: 'pic.twitter.com/6PefA9mjWz',

expanded\_url: 'https://twitter.com/NorthmanTrader/status/1655977667215097862/photo/1',

type: 'photo',

sizes: {

thumb: {

w: 150,

h: 150,

resize: 'crop'

},

medium: {

w: 1200,

h: 640,

resize: 'fit'

},

large: {

w: 2048,

h: 1092,

resize: 'fit'

},

small: {

w: 680,

h: 363,

resize: 'fit'

}

}

}

]

},

source: '<a href="https://mobile.twitter.com" rel="nofollow">Twitter Web App</a>',

in\_reply\_to\_status\_id: null,

in\_reply\_to\_status\_id\_str: null,

in\_reply\_to\_user\_id: null,

in\_reply\_to\_user\_id\_str: null,

in\_reply\_to\_screen\_name: null,

user: {

id: 714051110,

id\_str: '714051110',

name: 'Sven Henrich',

screen\_name: 'NorthmanTrader',

location: 'England, United Kingdom',

description: 'Founder: NorthmanTrader. Financial Market Strategist. Macro & Technical Analysis. Keeping it real. Subscribe: https://t.co/rmRFFvMRaH…',

url: 'https://t.co/2rOR1Gp0G6',

entities: {

url: {

urls: [

{

url: 'https://t.co/2rOR1Gp0G6',

expanded\_url: 'https://northmantrader.com',

display\_url: 'northmantrader.com',

indices: [

0,

23

]

}

]

},

description: {

urls: [

{

url: 'https://t.co/rmRFFvMRaH',

expanded\_url: 'http://northmantrader.com/market-service',

display\_url: 'northmantrader.com/market-service',

indices: [

110,

133

]

}

]

}

},

protected: false,

followers\_count: 450887,

friends\_count: 72,

listed\_count: 7731,

created\_at: 'Tue Jul 24 10:12:44 +0000 2012',

favourites\_count: 40595,

utc\_offset: null,

time\_zone: null,

geo\_enabled: false,

verified: false,

statuses\_count: 5366,

lang: null,

contributors\_enabled: false,

is\_translator: false,

is\_translation\_enabled: false,

profile\_background\_color: '1A1B1F',

profile\_background\_image\_url: 'http://abs.twimg.com/images/themes/theme9/bg.gif',

profile\_background\_image\_url\_https: 'https://abs.twimg.com/images/themes/theme9/bg.gif',

profile\_background\_tile: false,

profile\_image\_url: 'http://pbs.twimg.com/profile\_images/1534496843186552832/IFhiLmiE\_normal.jpg',

profile\_image\_url\_https: 'https://pbs.twimg.com/profile\_images/1534496843186552832/IFhiLmiE\_normal.jpg',

profile\_banner\_url: 'https://pbs.twimg.com/profile\_banners/714051110/1557695427',

}