Análisis descriptivo del precio acciones utilizando procesado del lenguaje natural

**ALUMNO 1: David de la Torre Amengual.**

**ALUMNO 2: Henry Ocaña Luna.**

**ALUMNO 3: Óscar Tienda Beteta.**

**ALUMNO 4: Mireya Harillo Gámez.**

**PROGRAMA:**

**MASTER EN BUSSINESS INTELLIGENCE AND DATA SCIENCE**

**NOMBRE DEL PROYECTO:**

Análisis descriptivo del precio de las acciones de GameStop mediante comentarios del subforo de Reddit WallStreetBets.

# **CONTENIDO**

[**CONTENIDO** 2](#_Toc83467357)

[**TABLA DE ILUSTRACIONES** 3](#_Toc83467358)

[**RESUMEN** 4](#_Toc83467359)

[**INTRODUCCIÓN** 4](#_Toc83467360)

[**ESTADO DEL ARTE** 7](#_Toc83467361)

[**OBJETIVOS** 13](#_Toc83467362)

[**SOLUCIÓN PLANTEADA** 13](#_Toc83467363)

[**EVALUACIÓN** 14](#_Toc83467364)

[**RESULTADOS** 14](#_Toc83467365)

[**CONCLUSIONES Y TRABAJOS FUTUROS** 14](#_Toc83467366)

[**REFERENCIAS** 14](#_Toc83467367)

# **TABLA DE ILUSTRACIONES**

[Ilustración 1. Evolución de precios GME. 5](#_Toc82202151)

# **RESUMEN**

El problema planteado para el desarrollo de este proyecto es el análisis del comportamiento del precio de las acciones de la compañía GameStop Corporation, que cotiza bajo el símbolo bursátil GME, tras ser el objetivo de inversores de un conocido sub-foro de la plataforma Reddit conocido como WallStreetBets. Con este proyecto se pretende describir cuales fueron las causas del aumento repentino de las acciones de dicha compañía, mediante el análisis de lenguaje natural que se encuentra en las publicaciones del foro.

Los datos utilizados para el proyecto se obtienen de dos fuentes. Por una parte, los datos de las publicaciones en el foro de Reddit provienen de una base de datos ya extraída de Kaggle. Por otro lado, los datos de los precios de las acciones se han obtenido del sitio web oficial de NASDAQ, National Association of Securities Dealers Automated Quotation.

El objetivo principal del proyecto es analizar si los comentarios del sub-foro WallStreetBets son una fuente fiable de la cual se puede extraer valor para invertir en acciones del mercado de valores, utilizando un sistema de procesado de lenguaje totalmente automatizado que aporte una ventaja competitiva frente a otros inversores.

# **INTRODUCCIÓN**

En enero de 2021, se produjo un [estrangulamiento de posiciones cortas](https://es.wikipedia.org/wiki/Estrangulamiento_de_posiciones_cortas) de las acciones del minorista estadounidense de videojuegos [GameStop](https://es.wikipedia.org/wiki/GameStop" \o "GameStop) y otros valores en varias bolsas de valores, lo que provocó importantes pérdidas financieras para los vendedores en corto, incluidos ciertos [fondos de cobertura](https://es.wikipedia.org/wiki/Fondo_de_cobertura). La contracción corta provocó que su precio alcanzase más de 500 dólares por acción, casi 30 veces el precio de principios de mes. Aproximadamente el 140 por ciento de las acciones de GameStop se habían vendido al descubierto, y la prisa por comprar acciones para cubrir esas posiciones a medida que el precio subía hizo que el precio incrementase aún más. El breve apretón fue inicial y principalmente provocado por usuarios del foro de Internet WallStreetBets en [Reddit](https://es.wikipedia.org/wiki/Reddit).

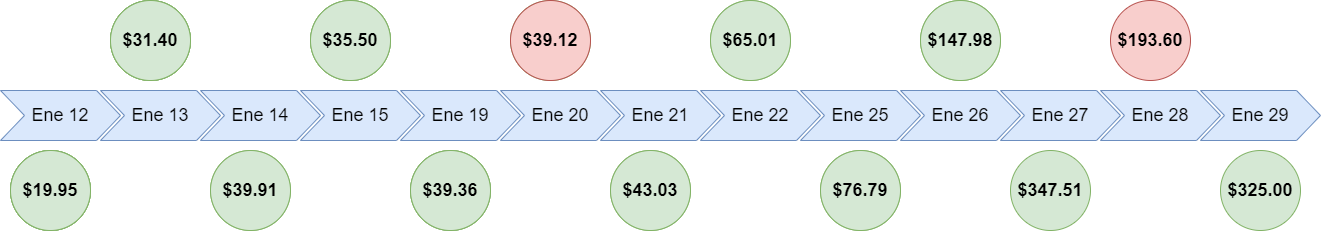


Ilustración 1. Evolución de precios GME

El 28 de enero, varias casas de bolsa, incluida Robinhood, detuvieron la compra de GameStop y otros valores. Esta decisión atrajo críticas y acusaciones de manipulación del mercado por parte de destacados políticos y empresarios de todo el espectro político. Se presentaron demandas colectivas contra Robinhood en los Tribunales de Distrito de EE. UU. para el Distrito Sur de Nueva York y el Distrito Norte de Illinois. Muchos otros valores muy cortos también experimentaron aumentos de precios. (Wikipedia, 2021).

La Ilustración 1. Evolución de precios GME resume la evolución que sufrió el precio de las acciones GME en el tiempo hasta el bloqueo de transacciones en la plataforma de inversión Robinhood.

Históricamente los inversores han tomado sus decisiones basadas en grandes cantidades de datos de precios, noticias e indicadores técnicos para definir con mayor precisión la fuerza de tendencias en el precio de las acciones. Todos estos datos e información en tiempos pasados se han analizado de forma manual. Uno de los problemas que supone esta manualidad en el procesado de datos reside en que las personas tienen días mejores y peores por diferentes motivos, y un mal día puede suponer perder una gran cantidad de dinero para un inversor. Otro de los problemas puede surgir al confiar en exceso de los indicadores técnicos, son una buena fuente de información para identificar tendencias. Pero tienden a reflejar el comportamiento del mercado con retraso, lo que puede hacer que perdamos la “ventaja” ante el mercado o resto de inversores.

Para el desarrollo de la solución planteada se extraen los sentimientos predominantes, mediante librerías de Python públicas, de los títulos y los cuerpos de las entradas en el sub-foro WallStreetBets de Reddit. Una vez contamos con los sentimientos se realiza un proceso de feature engineering. Finalmente se ejecuta un análisis descriptivo del precio de la accion de la compañía GameStop, con el fin de conocer si mediante las variables generadas es posible rescribir el comportamiento del precio de la acción.

Los seres humanos somos animales sociales, por lo que tenemos una tendencia innata en interactuar y compartir conocimiento. El problema es que muchas veces este conocimiento no es apto o no esta preparado para ser consumido por ordenadores. No hay que perder de vista que, contantemente los humanos, hemos ido cambiando la forma de comunicarnos desde el lenguaje antiguo y más formal, pasando por la jerga utilizada en SMS, a los cortos y resumidos tweets que combinan texto con iconos. Por ello tiene mucho sentido y puede aportar mucho valor programar los ordenadores para que ellos mismos sean capaces de extraer la información esencial de estas cadenas de caracteres que codifican tanto conocimiento. Y disponerlo para que otros modelos puedan procesar esta información.

“En los próximos diez años, la investigación en PNL se disparará. Podemos esperar que se produzcan grandes avances en la investigación fundamental de la PNL, las tecnologías básicas y las aplicaciones importantes.” (Ming Zhou, 2018)

Para lograr la solución planteada, después de cargar los datos en crudo en dataframes, se ha realizado un tratamiento inicial de estos. Posteriormente se ha realizado análisis exploratorio de los datos para identificar distribuciones, columnas con datos faltantes, extraer estadísticos descriptivos de los datos, … El objetivo del este análisis es tener una idea general de los datos con los que se trabaja. A continuación, se ha realizado un proceso de enriquecimiento de los datos inicialmente cargados. Una parte del proceso ha consistido en realizar un análisis de sentimientos de los títulos y los cuerpos de las entradas en el foro mediante dos modelos. El primer modelo es conocido como BERT y el segundo Vader, mediante la librería nltk. Adicionalmente para obtener unos datos más valiosos se han combinado los datos de los posts con los precios diarios de las accione. Una vez con los datos procesados se procede a analizar los resultados e identificar señales que permitan predecir el comportamiento del mercado únicamente con las entradas que se han ido escribiendo por los usuarios del foro.

* Resumen brevísimo de los resultados obtenidos.

El presente documento se divide en ocho apartados. En el primero de ellos se resume el problema abordado, se presenta la solución planteada y se justifica porque esta solución es buena. En el segundo apartado se introduce el problema detectado, soluciones anteriores al mismo problema, la solución planteada en el trabajo, el procedimiento para lograr aplicar la solución planteada y finalmente un breve resumen de los resultados obtenidos. En el tercer apartado se aborda el estado del arte, se referencian artículos o proyectos relevantes que describan como se ha afrontado el problema tratado. En el quinto apartado se define cual es el objetivo final del proyecto y cuatro objetivos específicos que describen como llevó a cabo el objetivo final. El sexto apartado se presenta la solución planteada y la metodología de desarrollo aplicada. En el sexto apartado se plasman los resultados obtenidos en el proyecto y la evaluación de los mismos. Por último, en el séptimo apartado se plantean las conclusiones obtenidas tras el trabajo y como se podría seguir avanzando para mejorar la solución planteada.

# **ESTADO DEL ARTE**

**Análisis de sentimientos de twitter usando NLP y redes neuronales**

En este proyecto se aborda el problema de superar el desafio de determinar los sentimientos de los tweets, clasificándolos entre positivo o negativo en función del sentimiento primario del mismo.

Para la resolución del problema se utiliza Python, como lenguaje de programación. Tensorflow como framework para el desarrollo del modelo de clasificación. Las métricas para evaluar el resultado del modelo han sido Accuracy, una matriz de confusión y la curva ROC.

El modelo de clasificación, que es el core del proyecto, utilizado ha sido un a red neuronal de 8 capas contando las capas de entrada y de salida. El resto de capas utilizadas, por orden, son las siguientes:

* Embeding, proporcionar la presentación de palabras y sus significados relativos. Así, estamos alimentando el límite de palabras máximas, la longitud de las palabras de entrada y las entradas de la capa anterior.
* LSTM, (long short term memory) guarda las palabras y predice las siguientes palabras basándose en las palabras anteriores. LSTM es un predictor de secuencia de las próximas palabras que vienen.
* Dense, reduzce las salidas obteniendo entradas de la capa Flatten. La capa densa utiliza todas las entradas de las neuronas de la capa anterior y realiza cálculos y envía 256 salidas
* Activation, neurona que ayuda a decidir qué neurona debe pasar y qué neurona debe dispararse. Entonces, la función de activación del nodo define la salida de ese nodo dada una entrada o un conjunto de entradas.
* Dropout, borra algunas neuronas de capas anteriores para evitar problemas de sobreajuste. En el sobreajuste, el modelo proporciona una buena precisión en el tiempo de entrenamiento, pero no es bueno para el tiempo de prueba.
* Dense.

<https://medium.com/analytics-vidhya/introduction-bd62190f6acd>

**Predicción de precios de acciones basado en redes sociales**

En esta tesis Heng Gui, el autor, analiza la relación entre los precios de las acciones del mercado de valores y las publicaciones en redes sociales. El primer paso para predecir los precios de las acciones es aplicar análisis de sentimientos a las publicaciones para extraer las emociones u opiniones individuales de los usuarios sobre las acciones. A continuación, el autor analiza las características via estadística descriptiva y predictiva usando el coeficiente de correlación de Pearson y modelos de *machine* *learning*.

El análisis predictivo esta diseñado para examinar la dependencia entre los datos de las redes sociales y los datos del cambio en el precio de las acciones evaluando el rendimiento de las predicciones. Los indicadores de rendimiento utilizados son:

* Precisión en la predicción de la dirección del cambio del precio.
* Beneficio total con estrategia de negociación para acciones individuales.
* Rentabilidad diaria de la estrategia comercial.
* Beneficio total con estrategia de negociación de cartera.

Los modelos aplicados para la resolución del problema propuesto en el proyecto son:

* Clasificador K-nearest neighbors.
* SVM.
* Regresión logística.
* Árbol de decisión.
* Clasificador random forest.
* Gradient boosting.

<https://trepo.tuni.fi/bitstream/handle/10024/118242/GuiHeng.pdf?sequence=2&isAllowed=y>

**Operación de criptomonedas basado en análisis de sentimientos de publicaciones en Twitter**

Stuart Colianni, Stephanie Rosales y Michael Signorotti en este trabajo pretenden comprobar si los datos de Twitter, referentes a las criptomonedas, pueden ser utilizados para desarrollar estrategias de operación que aporten ventajas competitivas.

Los autores, a través del aprendizaje supervisado, proponen una serie de modelos de machine learning con el fin de identificar movimientos en el mercado de las criptomonedas. En particular en el trabajo se centra el foco de atención en la que puede ser considerada hoy en día la criptomoneda más conocida, Bitcoin. Los modelos analizados en el estudio han sido:

* Regresión logística.
* Naive Bayes.
* SVM.

<http://cs229.stanford.edu/proj2015/029_report.pdf>

**Análisis de sentimientos de noticias financieras en Reddit y publicaciones en Financial Times**

Michael Lubitz, en su tesis, trata de evaluar si los sentimientos en las publicaciones del sub-foro *ecconomics* de Reddit pueden predecir el rendimiento a futuro del mercado de valores. Para ello realiza una extracción de las entradas en el sub-foro mencionado entre 2008 y 2017, fecha en la que se publicó el artículo. En el trabajo se evalúa el poder de predicción de Reddit comparando los sentimientos de todas las publicaciones de un día con el correspondiente precio de cierre del índice S&P 500.

El autor realiza el cálculo de los sentimientos de los artículos aplicando un enfoque basado en diccionarios y modelos de aprendizaje automático supervisado que se apoyan en diferentes métodos de extracción de características. En el trabajo adicionalmente se realiza un cálculo ponderado del sentimiento incorporando el número de los votos y comentarios en los modelos.

Para evaluar los resultados se utilizan noticias publicadas en *Financial Times* como línea base. Como resultado del trabajo se consigue demostrar la hipótesis planteada, Reddit tiene una capacidad de predicción ligeramente superior a la obtenida con noticias de medios estándar.

<https://ad-publications.cs.uni-freiburg.de/theses/Bachelor_Michael_Lubitz_2018.pdf>

**Arquitectura Web para análisis de sentimientos en Facebook con enfoque semántico**

En este trabajo, se propuso una arquitectura de software que considera el aprendizaje supervisado para clasificar publicaciones de Facebook en tres emociones: alegría, enojo y tristeza.

Los autores dividieron el trabajo en 3 partes:

* Extraer texto: Recuperar el texto de interés considerando la complejidad de la arquitectura de datos de las fuentes de información.
* Normalizar texto: Por medio de un tratamiento lingúístico se recuperan únicamente palabras clave que se puedan asociar con emociones, tomando en cuenta polisemia, sinonimia, negaciones de verbos entre otros.
* Clasificar texto: Diseñar mecanismos de reconocimiento de patrones, a fin de alcanzar una mejor precisión al momento de clasificar las publicaciones en las categorías de sentimientos definidas

Se presenta como una opción útil para análisis de sentimientos en Facebook, combinando un clasificador semántico y un clasificador Naive Bayes entrenado con publicaciones seleccionadas que alcanza un desempeño del 63 %.

<https://www.rcs.cic.ipn.mx/2014_75/Arquitectura%20Web%20para%20analisis%20de%20sentimientos%20en%20Facebook%20con%20enfoque%20semantico.pdf>

**La eficacia del análisis de sentimientos para la empresa: el caso de estudio Dell** **Technologies Inc.**

La empresa lleva a cabo un análisis semestral con el fin de detectar eventuales cambios en la satisfacción analizando los puntos de vista de más de 30.000 clientes como promedio. La medición se lleva a cabo utilizando una escala métrica que divide y etiqueta aquellos que facilitan una opinión dependiendo de la puntuación que hayan dado.

Debido a que los datos obtenidos son de tipo textual utilizan una serie de herramientas NLP capaces de traducir los datos textuales (el texto no estructurado) a información significativa y valiosa para la toma de decisiones. Para poder cuantificar adecuadamente esta información.

Entre las herramientas que utiliza la empresa para llevar a cabo el análisis destacan:

* Dashboards
* Haddop
* R

En este trabajo realizado por un alumno de la Universidad Complutense de Madrid se pone de manifiesto las bondades que ofrece el Análisis de sentimiento mostrando el uso de esta técncia por parte de la empresa Dell Technologies Inc.

**Del Data-driven Al Data-feeling: Análisis De Sentimiento En Tiempo Real De Mensajes En Español Sobre Divulgación Científica Usando Técnicas De Aprendizaje Automático**

En esta obra los autores utilizan técnicas de aprendizaje automático para desarrollar un clasificador de sentimiento relacionados con mensajes publicados a tiempo real en Twitter. El nombre del prototipo es OpScience. Para desarrollarlo, descargaron 200.000 tweets para construir el grupo de entrenamiento y etiquetaron 10.000 textos etiquetados, 5.000 positivos y 5.000 negativos, para analizar la polaridad de las opiniones en Twitter y centrando el estudio en la comunicación de la ciencia en español. De esta manera, el equipo refleja el desarrollo de un clasificador en tiempo real que hace uso de técnicas de aprendizaje automático.

Este trabajo se basa en un estudio previo que observó que un detector de sentimiento basado en Twitter puede llegar a replicar la confianza del consumidor. El gran avance que presenta este proyecto es la utilización de un sistema de inteligencia artificial para evitar el continuo etiquetado manual de los textos del estudio.

El objetivo principal de este prototipo surge de la pregunta “¿Podemos analizar una parte de los datos públicos disponibles en la red social Twitter para conocer actitudes, opiniones y sentimientos en torno a los temas de comunicación de la ciencia que se comparten en esa red, y avanzar hacia la predicción de tendencias o comportamientos futuros?” y cuyo objetivo se alcanza en el desarrollo de tres tareas principales:

1. La construcción de un nuevo corpus de textos de tema científico en lenguaje español, clasificado por sentimiento positivo o negativo, para el entrenamiento del modelo de aprendizaje automático.
2. Aprendizaje automático supervisado por el que se lleva a cabo el entrenamiento del modelo al dividir el corpus en entrenamiento y validación en proporción 70%-30%. Los primeros manualmente etiquetados y los segundos incluidos para ajustar los parámetros del modelo.
3. La evaluación del modelo mediante la implementación del prototipo clasificador de sentimientos en tiempo real y su testeo con datos nuevos tomados a tiempo real de la red social.

Los algoritmos planteados en esta publicación son los siguientes:

* Original Naive Bayes
* Naive Bayes for multimodal models
* Naive Bayes for multivariate Bernoulli models
* Logistic Regression
* Linear Support Vector Classification (svc)
* Linear classifiers with stochastic gradient descent training
* Votación por intervalos de características

<https://revistas.urosario.edu.co/xml/5115/511562674006/html/index.html>

**Machine Learning aplicado al Trading**

En este artículo, la autora, Macarena Salvador, realiza un acercamiento teórico mediante una recopilación de estudios relacionados con el uso del *Machine Learning* para anticiparse a los movimientos de la bolsa y trabajar en la línea del *high frequency learning*, donde las órdenes  de compras y ventas tienen un gran volumen, contando con miles de operaciones en fracciones de segundo. Estas órdenes buscan pequeños márgenes de beneficios en cada operación. Esta metodología no se basa en predecir el valor futuro de los activos financieros, sino que aprovecha las pequeñas ineficiencias del mercado.

En este documento se desarrolla la teoría relacionada con las siguientes técnicas:

* Árbol de decisión o SVM (Support Vector Mach), ​​para elegir los activos financieros que puedan dar mayor beneficio deducido a partir de sus características.
* Modelo de decisión de Markov
* Percepción Simple y multiplicada
* Funciones de Redes Recurrentes
* Redes LSTM (Long short-term memory)

<https://repositorio.comillas.edu/rest/bitstreams/295750/retrieve>

# **OBJETIVOS**

## Objetivo general

Desarrollar un proceso de análisis descriptivo que permita afirmar si es posible identificar tendencias del mercado de valores apoyándose en las publicaciones en el sub-foro WallStreetBets de Reddit. Con el objetivo de conseguir una ventaja competitiva frente a otros inversores que realicen un análisis de las acciones de forma más tradicional.

## Objetivos específicos

* 1. Obtener unos datos limpios y aptos para poder resolver el objetivo general.
  2. Determinar los modelos de análisis de sentimientos idóneos para extraer los sentimientos de las publicaciones y títulos.
  3. Determinar los sentimientos de las publicaciones y títulos.
  4. Determinar las relaciones entre los precios/volúmenes de operaciones diarias en función de los sentimientos.

# **SOLUCIÓN PLANTEADA**

En esta sección el estudiante debe describir la solución planteada, comenzando por la metodología (pasos que siguió) de desarrollo. Posteriormente la descripción del desarrollo de cada etapa seguida.

La metodología que debe utilizar el estudiante de master debe estar validada por la comunidad científica y el estudiante debe justificarlo.

## Metodología

**Etapa 1**

## Desarrollo de cada etapa

# **EVALUACIÓN**

El estudiante en esta sección debe describir cómo realizó la evaluación de la solución planteada.

La evaluación de un TFM debe basarse en una metodología validada y bien establecida que el estudiante deberá justificar.

Extensión máxima: 10 páginas.

*Escribe aquí…*

# **RESULTADOS**

El estudiante en esta sección debe describir los resultados obtenidos en el proceso de evaluación de la solución planteada.

Debe usar notación estándar para presentación de resultados de carácter científico.

Extensión máxima: 10 páginas.

*Escribe aquí…*

# **CONCLUSIONES Y TRABAJOS FUTUROS**

El estudiante plantea las conclusiones de su trabajo, y cómo considera que puede seguir avanzando en el mejoramiento de la solución planteada.

Extensión máxima: 2 páginas

*Escribe aquí…*

# **REFERENCIAS**

Lista de referencias bibliográficas utilizadas, siguiendo las normas APA.

(Wikipedia, 2021)

(Zaman, 2021)

(Ming Zhou, 2018)