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](#_Toc83834360)

[**TABLA DE ILUSTRACIONES** 3](#_Toc83834361)

[**RESUMEN** 5](#_Toc83834362)

[**INTRODUCCIÓN** 5](#_Toc83834363)

[**ESTADO DEL ARTE** 7](#_Toc83834364)

[**OBJETIVOS** 16](#_Toc83834365)

[Objetivo general 16](#_Toc83834366)

[Objetivos específicos 16](#_Toc83834367)

[**SOLUCIÓN PLANTEADA** 16](#_Toc83834368)

[Metodología 16](#_Toc83834369)

[Desarrollo de cada etapa 18](#_Toc83834370)

[Comprensión del negocio 18](#_Toc83834371)

[Comprensión de los datos 18](#_Toc83834372)

[EDA Reddit 18](#_Toc83834373)

[EDA GME 22](#_Toc83834374)

[Preparación de los datos 24](#_Toc83834375)

[Análisis de sentimientos: Clasificación de sentimientos en los títulos 24](#_Toc83834376)

[Librería BERT 24](#_Toc83834377)

[Librería NLTK 25](#_Toc83834378)

[Librería NLTK 29](#_Toc83834379)

[Modelado 30](#_Toc83834380)

[Evaluación 43](#_Toc83834381)

[Despliegue 43](#_Toc83834382)

[**EVALUACIÓN** 44](#_Toc83834383)

[**RESULTADOS** 44](#_Toc83834384)

[**CONCLUSIONES Y TRABAJOS FUTUROS** 44](#_Toc83834385)

[**Referencias** 45](#_Toc83834386)

# **TABLA DE ILUSTRACIONES**

[Ilustración 1. Evolución de precios GME. 5](#_Toc83834289)

[Ilustración 2. Etapas CRISP-DM, por Daniel Álvarez Gil (https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/) 17](#_Toc83834290)

[Ilustración 3. Datos de partida Reddit. 19](#_Toc83834291)

[Ilustración 4. Datos faltantes Reddit. 19](#_Toc83834292)

[Ilustración 5. . Ejemplo de registro con columna body nula. 20](#_Toc83834293)

[Ilustración 6. Información descriptiva del conjunto de datos de Reddit. 20](#_Toc83834294)

[Ilustración 7. Evolución del like medio de las publicaciones por fecha. 20](https://d.docs.live.net/29c7d1ce3830647b/Escritorio/MDBI%20-%20Plantilla.%20docx.docx#_Toc83834295)

[Ilustración 8. Evolución de la cantidad de publicaciones por fecha. 20](https://d.docs.live.net/29c7d1ce3830647b/Escritorio/MDBI%20-%20Plantilla.%20docx.docx#_Toc83834296)

[Ilustración 9. Relación entre la longitud de las publicaciones y likes. 21](https://d.docs.live.net/29c7d1ce3830647b/Escritorio/MDBI%20-%20Plantilla.%20docx.docx#_Toc83834297)

[Ilustración 10. Relación entre la longitud de los títulos y likes de las publicaciones. 21](https://d.docs.live.net/29c7d1ce3830647b/Escritorio/MDBI%20-%20Plantilla.%20docx.docx#_Toc83834298)

[Ilustración 11. Distribución del número de comentarios en las publicaciones. 21](https://d.docs.live.net/29c7d1ce3830647b/Escritorio/MDBI%20-%20Plantilla.%20docx.docx#_Toc83834299)

[Ilustración 12. Relación entre el número de comentarios y los likes de las publicaciones. 21](https://d.docs.live.net/29c7d1ce3830647b/Escritorio/MDBI%20-%20Plantilla.%20docx.docx#_Toc83834300)

[Ilustración 13. Distribución de likes en las publicaciones. 21](https://d.docs.live.net/29c7d1ce3830647b/Escritorio/MDBI%20-%20Plantilla.%20docx.docx#_Toc83834301)

[Ilustración 14. Datos de partida GME. 22](#_Toc83834302)

[Ilustración 15. Datos faltantes GME. 22](#_Toc83834303)

[Ilustración 16. Información descriptiva del conjunto de datos GME. 23](#_Toc83834304)

[Ilustración 17. Evolución del precio de apertura por fecha. 23](#_Toc83834305)

[Ilustración 18. Gráfico de velas japonesas GME. 24](#_Toc83834306)

[Ilustración 19. Resultado de tokenizar el título ejemplo. 24](#_Toc83834307)

[Ilustración 20. Resultados análisis sentimiento negativo-negativo/neutro-neutro-positivo/neutro-positivo. 25](#_Toc83834308)

[Ilustración 21.Resultados modelo BERT en el título. 25](#_Toc83834309)

[Ilustración 22. Modelado de los datos titulos. 25](#_Toc83834310)

[Ilustración 23. Parametrizamos para evitar valor 0. 26](#_Toc83834311)

[Ilustración 24. Demostración valores sentimientos títulos. 26](#_Toc83834312)

[Ilustración 25. Ecuación sentimiento predominante. 26](#_Toc83834313)

[Ilustración 26. Unión dataset con scoring obtenido con NLTK. 27](#_Toc83834314)

[Ilustración 27.Tabla visualización resultados BERT y NLTK. 27](#_Toc83834315)

[Ilustración 28. Cálculo de sentimientos con modelo BERT en los posts. 27](#_Toc83834316)

[Ilustración 29. Visualización del resultado previo. 28](#_Toc83834317)

[Ilustración 30. Unión de dataset con scoring títulos. 28](#_Toc83834318)

[Ilustración 31. Resultados del merge. 28](#_Toc83834319)

[Ilustración 32. Visualización. 29](#_Toc83834320)

[Ilustración 33. Modelado de los datos posts. 29](#_Toc83834321)

[Ilustración 34. Demostración valores sentimientos posts. 29](#_Toc83834322)

[Ilustración 35. Unión de dataset con scoring títulos. 30](#_Toc83834323)

[Ilustración 36. Resultados del merge. 30](#_Toc83834324)

[Ilustración 37. Datos BERT y NLTK. 30](#_Toc83834325)

[Ilustración 38. Reddit Dataset. 30](#_Toc83834326)

[Ilustración 39. 31](#_Toc83834327)

[Ilustración 40. 31](#_Toc83834328)

[Ilustración 41. 31](#_Toc83834329)

[Ilustración 42. Generación de columnas por sentimiento. 31](#_Toc83834330)

[Ilustración 43 . Visualización de resultados. 31](#_Toc83834331)

[Ilustración 44. Filtros por scoring. 32](#_Toc83834332)

[Ilustración 45. Funciones de agregación. 32](#_Toc83834333)

[Ilustración 46. Visualización de resultados. 32](#_Toc83834334)

[Ilustración 47. Creación de variables. 32](#_Toc83834335)

[Ilustración 48. Variables por score de representación de likes. 33](#_Toc83834336)

[Ilustración 49. Variables por score de representación de posts. 33](#_Toc83834337)

[Ilustración 50. Aplicación de ponderación del análisis de los títulos. 33](#_Toc83834338)

[Ilustración 51. Análisis de sentimientos de los posts. 34](#_Toc83834339)

[Ilustración 52. Cuadro de visualización. 34](#_Toc83834340)

[Ilustración 53. Visualización dataset GME. 34](#_Toc83834341)

[Ilustración 54. Merge. 35](#_Toc83834342)

[Ilustración 55. Visualización tabla datos. 35](#_Toc83834343)

[Ilustración 56. Relación clasificación sentimientos BERT. 35](#_Toc83834344)

[Ilustración 57. Detalle relación clasificación sentimientos BERT títulos y volumen. 36](#_Toc83834345)

[Ilustración 58. Relación clasificación sentimientos BERT post y volumen. 36](#_Toc83834346)

[Ilustración 59. Detalle relación clasificación sentimientos BERT post y volumen. 37](#_Toc83834347)

[Ilustración 60. Relación clasificación sentimientos NLTK títulos y volumen. 37](#_Toc83834348)

[Ilustración 61. Detalle relación clasificación sentimientos NLTK títulos y volumen. 38](#_Toc83834349)

[Ilustración 62. Relación clasificación sentimientos NLTK cuerpo posts y volumen. 38](#_Toc83834350)

[Ilustración 63. Detalle relación clasificación sentimientos NLTK cuerpo posts y volumen. 39](#_Toc83834351)

[Ilustración 64. Matriz de correlación. 40](#_Toc83834352)

[Ilustración 65. Concatenación de los títulos. 40](#_Toc83834353)

[Ilustración 66. Nube de palabras títulos positivos BERT. 41](#_Toc83834354)

[Ilustración 67. Nube de palabras post positivos BERT. 41](#_Toc83834355)

[Ilustración 68. Nube de palabras títulos positivos y negativos NLTK. 42](#_Toc83834356)

[Ilustración 69. Nube de palabras títulos positivos y negativos NLTK. 43](#_Toc83834357)

[Ilustración 70. Captura de pantalla de la imagen 200 del dataset r/WallStreetBets. 44](#_Toc83834358)

[Ilustración 71. Texto raw obtenido de utilizar Tesseract en la Ilustración 69. 45](#_Toc83834359)

# **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) 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).

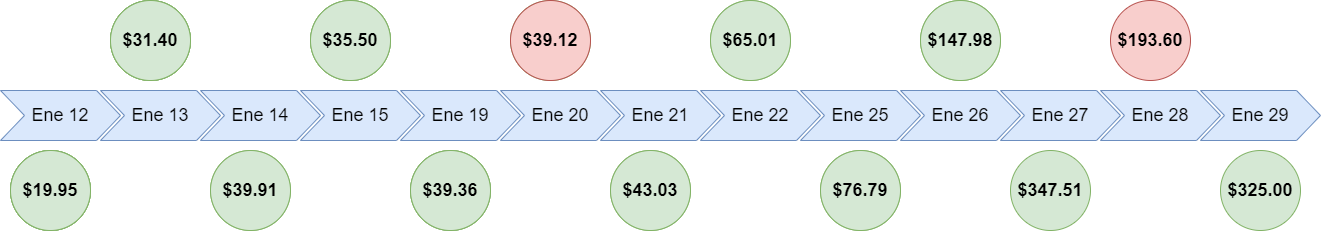


Ilustración 1. Evolución de precios GME.

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 acción 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 está 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, etc. 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 desafío 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*, reduce 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 vía 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 lingüístico 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>

**Aprendizaje no Supervisado y Detección de Anomalías: ¿Qué es una Anomalía?**

El autor, Alejandro Alcalde, *Data Scientist and Computer Scientist*, realiza un trabajo de divulgación sobre el aprendizaje no supervisado, que utilizamos en este trabajo, y la detección de anomalías. Entre otras, nos presenta dos formas de representación de los datos:

* Regresión lineal
* Clustering

Además, introduce el concepto de los Sistemas de detección de intrusiones en red (*Network Intrusion Detection Systems - NIDS*). Con respecto a estos, desarrolla cómo se aplican en los modelos supervisados, donde se desarrolla una clasificación entre elementos anómalos y legítimos. En el caso de que se presente un mayor número de datos anómalos se considera que tratamos con una clasificación desbalanceada y plantea dos métodos para lidiar con esta casuística: Los métodos basados en Instancias y los basados en algoritmos. En el primero de los casos, son los datos los que se modifican antes de incluirlos en el algoritmo, en el seguro, es el algoritmo el que debe ser modificado.

Con respecto a los métodos semi-supervisados destaca el proceso de clasificación, dando opciones para mejorarlo como son los clasificadores basados en reglas o la generación de máquinas de estados finitos. Los primeros buscan patrones frecuentes y reglas de asociación, por ejemplo, LERAD, que aprende reglas para encontrar eventos extraños en una serie temporal. Por otro lado, los métodos basado en Kernel como Máquinas de Soporte Vectoriales construye una región en el espacio que considera normal, y tacha como anómalo todo lo que salga de dicho espacio.

Ya tratando los métodos no supervisados, plantea las dificultades a la hora de reconocer los datos anómalos y plantea varias alternativas.

* Aproximaciones gráficas
* Aproximaciones paramétricas
* Aproximaciones basadas en vecinos cercanos
* Aproximaciones basadas en *clustering*

Por un lado, los métodos de visualización que, como contrapunto, tienen como desventaja la necesidad de un empleo mayor de tiempo y, puntualiza, un sesgo subjetivo. Como sistema de visualización recomienda el biplot, donde plos datos se proyectan en dos dimensiones. Con respecto al uso de un modelo paramétrico, se centraen la descripción de la distribución de los datos y aplica tests estadísticos que determinan los *outliners*. Son ejemplos de esta aproximación el test Grubb, los tests de Tietjen y Moore. También introduce la distancia de Mahalanobis en casos con varias dimensiones (p dimensiones).

Para casuísticas en la que la distribución de los datos no sea normal, el autor aborda el asunto desde una perspectiva basada en vecinos cercanos, mediante una función de distancia que mida la cercanía entre dos puntos o asignando una puntuación de anomalía a un punto en función de su distancia frente al resto.

Por último, si trabajamos con clusters previamente construidos, podremos medir la distancia de un punto a su centroide para determinar si se trata de una anomalía. En este caso, el autor nos plantea dos posibilidades, la distancia Euclídea y la distancia relativa.

Finalmente nos plantea la ecuación TP/TP+FP que indica qué porcentaje en la predicción de anomalías es correcto, donde TP es un acierto verdadero y FP es un falso positivo.

<https://elbauldelprogramador.com/aprendizaje-nosupervisado-anomalias/#basados-en-instanciasa-idsec-2-2-1-namesec-2-2-1a>

**TechRules diseña un algoritmo basado en Machine Learning para descubrir nuevas posibilidades de inversión**

En este artículo se expone la generación de un algoritmo que funciona como un sistema de recomendación aplicado a inversión para la comunidad financiera Rankia. Su desarrollador, TechRules trabaja con datos obtenidos a tiempo real de los casi 40.000 usuarios de esta comunidad online para luego agruparlos en lo que en banca se denominan grupos operativos que gestionan carteras de valores y condiciones similares. Una vez realizado el *clustering*, el sistema se basa en los comportamientos de los usuarios comunes para generar recomendaciones adecuadas al tipo de cliente y cartera, de manera que pueda adaptarse a las necesidades del cliente.

Se trata, por tanto, de una nueva aportación del *machine learning* a la banca tradicional y a la inversión en la bolsa de valores. Falta, sin embargo, el tratamiento de los comentarios de los usuarios en los foros de la comunidad para que, a través de procesamientos de lenguaje natural, se incluya la aportación de los análisis de sentimiento en el algoritmo aplicado.

<https://www.rankia.com/blog/rankia-com-blog-t-advisor/4360785-techrules-disena-algoritmo-basado-machine-learning-para-descubrir-nuevas-posibilidades-inversion>

**Estudio de G.Magesh y P. Swarnalatha sobre el uso de algoritmos de Machine Learning para la predicción del precio de stocks**

Muchos de los inversores bursátiles han optado por técnicas de inversión basadas en el Value Investing, buscando empresas cuyas proyecciones a años vista sean favorables. Sin embargo, a día de hoy muchos inversores siguen realizando técnicas de inversión cortoplacistas: buscando identificar qué empresas subirán o bajarán de precios en períodos de tiempo mínimos, incluso realizando operaciones intradías. Uno de los motivos por los cuales este tipo de inversiones son cada vez más minoritarias es la amplia presencia de bots en el mercado que utilizan algoritmos de Machine Learning para predecir los cambios de precio, haciendo así que los márgenes de beneficio de este tipo de operaciones sean cada vez más ajustados.

Esto es algo que lograron repetir estos profesores de la Universidad de Singapur en 2018: utilizando un algoritmo de clasificación (Weka toolkit, que incluía modelos de árboles de decisión, decisiones de soporte vectoriales, árboles aleatorios, bosques aleatorios y Naive Bayes) y utilizando datos de los 4 años anteriores (el 70% de los datos para entrenar el modelo y el 30% para problarlo). Clasificaron los stocks estudiados en “SAFE” y “UNSAFE” dependiendo de si el precio a 6 meses vista era mayor o menor y tras eliminar y testear con diferentes variables obtuvieron un 75% de acierto.

De este estudio tenemos que tener en consideración que es ciertamente criticable por un factor determinante: al operar en un período de 6 meses sobre los precios de los stocks se podría criticar que el cambio de precio en los stocks está vinculado directamente en el estado general del mercado: puede que esos seis meses finales fueran mayoritariamente positivos para el mercado en un alto porcentaje y por ende el 75% de acierto no sea realista en un contexto más negativista.

Magesh, G., y P. Swarnalatha. «Predictive Analysis of Stocks Using Data Mining». En *Smart Intelligent Computing and Applications*, editado por Suresh Chandra Satapathy, Vikrant Bhateja, y Swagatam Das, 105:283-89. Smart Innovation, Systems and Technologies. Singapore: Springer Singapore, 2019. <https://doi.org/10.1007/978-981-13-1927-3_30>.

**Artículo de la Social Science Computer Review (SAGE) sobre las consideraciones al utilizar datos de Reddit**

Este estudio se centra en la utilización de datos de Reddit para análisis de datos de tipo puramente social. Al realizar un estudio que busca vincular los comentarios de Reddit con el precio de un stock este artículo nos puede aportar información muy valiosa.

Una de las particularidades de la incorporación de datos de Reddit en el estudio es la diversidad de formatos: los comentarios en Reddit pueden incluir texto plano, imágenes, vídeos, GIFs, enlaces a páginas webs completamente distintas, enlaces a artículos académicos, enlaces de descarga de archivos de todo tipo y varios otros. Por eso se menciona que incorporar datos de esta red social puede ser especialmente tedioso en algunos casos.

Además, se hace mención a la no perfección de los datos: gran parte de los datos obtenidos pueden ser datos muy difíciles de procesar para nuestros algoritmos. Incluso los algoritmos de PLN más avanzados fallarán estrepitosamente cuando la fila de texto a analizar es un comentario con faltas de ortografía en respuesta a otro comentario que puede estar no vinculado con el mensaje original del tema.

Se hace especial mención a la importancia de seleccionar los tipos de datos a aceptar y los tipos de datos a ignorar: quizá para la simplicidad de nuestro estudio nos interese eliminar todos aquellos enlaces externos o los vídeos de más de X fotogramas. Además, se deberían responder a toda una serie de preguntas, como, por ejemplo:

* + ¿Cuál es el target de tu estudio?
  + ¿Cuál es la pregunta principal de tu estudio?
  + ¿Quién consideras que es el usuario de tu subrredit analizado?
  + ¿Qué variables utilizaras en tu análisis y cómo la codificarás?

*Amaya, Ashley, Ruben Bach, Florian Keusch, y Frauke Kreuter. «New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data». Social Science Computer Review 39, n.o 5 (octubre de 2021): 943-60.* [*https://doi.org/10.1177/0894439319893305*](https://doi.org/10.1177/0894439319893305)*.*

**Análisis de sentimientos en redes sociales para la predicción de precios de stocks**

El objetivo de este estudio es la creación de un modelo para predecir las variaciones en el precio de un stock acorde a los sentimientos en redes sociales. Este estudio se diferencia en sus antecesores en que evita tratar el sentimiento de los consumidores como un sentimiento general (positivo, negativo) si no que trata de analizar las opiniones en ámbitos específicos de una compañía.

Dieciocho fueron las acciones estudiadas y el método logro mejorar un 2.07% los resultados obtenidos respecto a un método que consideraba únicamente los precios históricos. Además, seleccionando solo las acciones con peor resultado (las más difíciles de predecir) este método obtuvo un 9.83% de mejores resultados que el método histórico y un 3.03% que el método de análisis de sentimientos generalista.

Dos métodos son los propuestos finalmente por este estudio:

* + Un método basado en algoritmos JST que utiliza la bibliografía interior.
  + Otro método, el central en el estudio, que utiliza análisis de sentimientos por “aspectos” (específicos) donde los temas y los sentimientos son analizados particularmente.

La principal limitación de este estudio es la preasignación de parámetros a ser analizados: se menciona una gran posible mejora si el algoritmo se aplicase de una forma no paramétrica (donde el propio algoritmo es capaz de discernir cuantos “aspectos” y tipos de sentimiento ha de analizar en cada caso). Otra de las limitaciones es que este método solo es capaz de discernir variaciones positivas o negativas en el precio del stock pero no trata especialmente a las variaciones drásticas.

Nguyen, Thien Hai, Kiyoaki Shirai, y Julien Velcin. «Sentiment Analysis on Social Media for Stock Movement Prediction». *Expert Systems with Applications* 42, n.o 24 (diciembre de 2015): 9603-11. <https://doi.org/10.1016/j.eswa.2015.07.052>.

# **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

A continuación, se presenta la metodología utilizada, haciendo una introducción de las etapas que componen el proceso. Seguidamente, se hace un repaso del trabajo desarrollado en cada una de las etapas.

La metodología que se ha seguido a lo largo del proyecto ha sido CRISP-DM (*Cross Industry Standard Process for Data Mining*), que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos. Esta metodología proporciona un marco en el que se especifican, en cada fase, las tareas concretas a realizar y las salidas deseables a obtener tras cada fase. (www.sngular.com, s.f.)

Las etapas que componen la metodología CRISP-DM son las que se muestran en la Ilustración 2. Etapas CRISP-DM, por Daniel Álvarez Gil :  
(https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/)

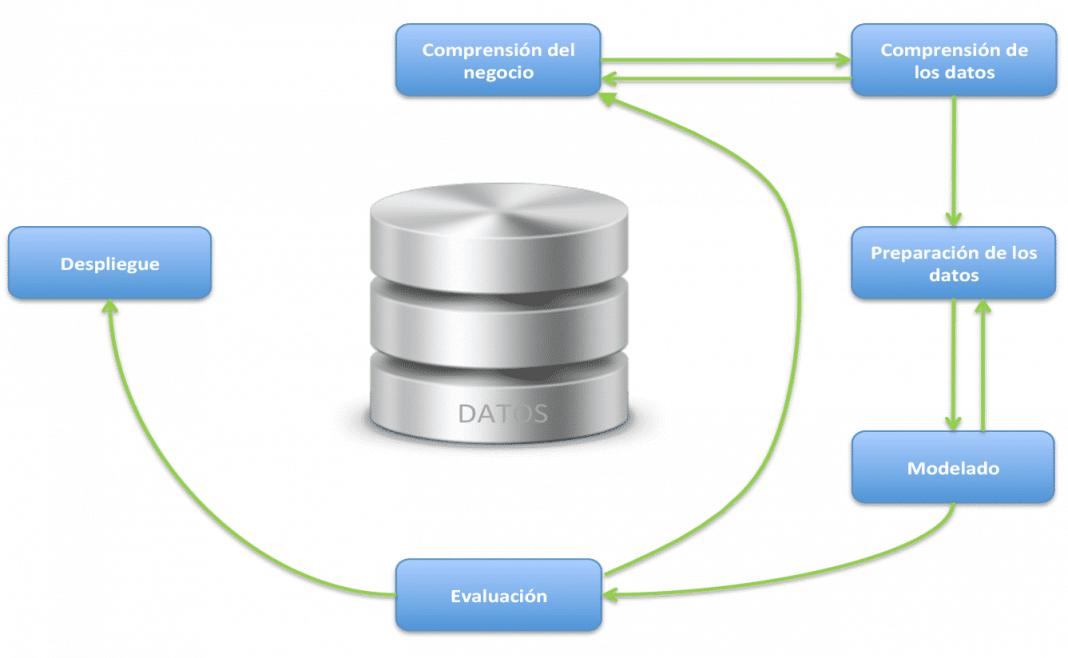


Ilustración 2. Etapas CRISP-DM, por Daniel Álvarez Gil (https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/)

**Etapa 1. Comprensión del negocio**

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto.

**Etapa 2. Comprensión de los datos**

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

**Etapa 3. Preparación de los datos**

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos, es decir, los datos que se utilizarán en las herramientas de modelado. Las tareas que se incluyen en esta fase son:

* Selección de tablas, registros y atributos.
* Transformación y limpieza de datos.

**Etapa 4. Modelado**

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Como algunas técnicas de modelado tienen requerimientos específicos en los datos es común, en esta fase, tener que volver a la fase tres de preparación de los datos. Esta posible situación queda patente en la Ilustración 2.

**Etapa 5. Evaluación**

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.

**Etapa 6. Despliegue**

Esta etapa depende directamente de los requisitos del proyecto y puede ser tan simple como la generación de un informe o tan compleja como la realización periódica de un proceso de análisis de datos. De cualquier modo, el objetivo es organizar el conocimiento obtenido y presentarlo a los usuarios.

## Desarrollo de cada etapa

A continuación, se detallan las actividades que se han desarrollado en cada una de las fases del proyecto, anteriormente definidas. El desarrollo del código se ha realizado en Python, utilizando librerías para acceder a funcionalidades que no se encuentran en Python de serie y que agilizan mucho el desarrollo de código. El entorno de desarrollo seleccionado ha sido *Jupyter Notebook*. Este entorno nos permite crear documentos que contienen celdas de código, visualizaciones y texto para explicar funcionalidades del código.

Como el proyecto se ha realizado de una forma descentralizada y colaborativa se ha utilizado Git para el mantenimiento de versiones. Todo el código y archivos, a los que se hace referencia a continuación, se encuentran en el repositorio <https://github.com/david-dlta/recursosTFM>.

Para facilitar el trabajo colaborativo y la reproducibilidad del proyecto en cualquier máquina, en el repositorio remoto, se puede encontrar el archivo *enviromentTFM.yml*. Este archivo contiene todas las librerías necesarias y sus versiones para ejecutar el código.

### Comprensión del negocio

Objetivos detallados en la sección **OBJETIVOS**.

### Comprensión de los datos

En esta etapa se realiza un Análisis Exploratorio de Datos (EDA, por sus siglas en inglés) EDA es el proceso de utilizar resúmenes numéricos y visualizaciones para explorar sus datos e identificar posibles relaciones entre variables. Es un proceso de investigación para llegar a conocer los datos y comprender lo que se puede averiguar de ellos. Se pueden hallar anomalías en los datos, como valores atípicos u observaciones inusuales, revelar patrones y generar preguntas o hipótesis interesantes. Es un paso fundamental en cualquier proyecto de analítica de datos. (www.jmp.com, s.f.)

En este proyecto se han utilizado dos grupos de datos, que provienen de fuentes diferentes. Por una parte, tenemos los datos de las publicaciones del sub-foro de Reddit y por otra parte los datos de los precios de la acción de la compañía GameStop (GME). Los datos de Reddit se han extraído de la plataforma [Kaggle](https://www.kaggle.com/gpreda/reddit-wallstreetsbets-posts), una comunidad de científicos de datos. Los datos de los precios de la acción GME se han obtenido del sitio web oficial de [NASDAQ](https://www.nasdaq.com/es/market-activity/stocks/gme), *National Association of Securities Dealers Automated Quotation*. Ambos conjuntos de datos se han extraído en formato CSV y se encuentran en el repositorio anteriormente mencionado.

#### EDA Reddit

El proceso comienza con una consulta para conocer el número de registros, si existen valores nulos, el tipo de datos con el que contamos y la memoria que requieren los datos. El resultado se puede ver en la Ilustración 3. Para facilitar la identificación de registros nulos se genera una visualización, el resultado se puede ver en la Ilustración 4.

Texto, Tabla

Descripción generada automáticamente

Ilustración 3. Datos de partida Reddit.

Imagen que contiene Gráfico de barras

Descripción generada automáticamente

Ilustración 4. Datos faltantes Reddit.

Si aislamos los datos faltantes del conjunto, apoyándonos en la variable URL, podemos ver que las filas con valor nulo en la columna *body* se tratan de archivos multimedia (imágenes, videos, gif, …), como podemos observar en la Ilustración 5.

Texto

Descripción generada automáticamente

Ilustración 5. . Ejemplo de registro con columna body nula.

A continuación, extraemos información descriptiva del conjunto de datos como: el número de registros, cuantos registros únicos existen para cada columna, la media, valor mínimo, valor máximo, etc. Los resultados se pueden ver en la ilustración 6.

Interfaz de usuario gráfica

Descripción generada automáticamente

Ilustración 6. Información descriptiva del conjunto de datos de Reddit.

El siguiente paso en la exploración de los datos ha sido la generación de algunas gráficas, entre pares de variables interesantes del conjunto de datos, para analizar si existe alguna relación entre ellas.

Gráfico

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente

Ilustración 7. Evolución del like medio de las publicaciones por fecha.

Ilustración 8. Evolución de la cantidad de publicaciones por fecha.

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

Ilustración 9. Relación entre la longitud de las publicaciones y likes.

Ilustración 10. Relación entre la longitud de los títulos y likes de las publicaciones.

Gráfico, Histograma

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

Ilustración 11. Distribución del número de comentarios en las publicaciones.

Ilustración 12. Relación entre el número de comentarios y los likes de las publicaciones.

Gráfico

Descripción generada automáticamente

Ilustración 13. Distribución de likes en las publicaciones.

*Comentarios sobre las gráficas.*

#### EDA GME

El proceso de exploración del conjunto de datos de los precios de la acción GME comienza de la misma manera, una consulta para conocer el número de registros, si existen valores nulos, el tipo de datos con el que contamos y la memoria que requieren los datos. El resultado se puede ver en la Ilustración 3. En la Ilustración 4 podemos ver que los datos no contienen registros nulos.

Texto

Descripción generada automáticamente

Ilustración 14. Datos de partida GME.

Imagen que contiene electrónica, computadora

Descripción generada automáticamente

Ilustración 15. Datos faltantes GME.

Seguidamente, extraemos información descriptiva del conjunto de datos del precio de la acción GME.

Tabla

Descripción generada automáticamente

Ilustración 16. Información descriptiva del conjunto de datos GME.

El proceso de exploración de los datos de este conjunto continua con la representación gráfica de algunos pares de variables que consideramos relevantes.

Interfaz de usuario gráfica, Gráfico, Gráfico de líneas, Histograma

Descripción generada automáticamente

Ilustración 17. Evolución del precio de apertura por fecha.

Pero el precio de apertura únicamente no nos aporta mucho valor. En economía se suele utilizar muy comúnmente un tipo de gráficos llamado velas japonesas, que permiten representar en un formato muy compacto el precio de apertura de la acción, el precio de cierre, el precio máximo que ha alcanzado la acción y el precio mínimo. De esta manera maximizamos el valor que el gráfico nos puede ofrecer, incluso representando si la acción ha tenido un precio de cierre inferior o superior al precio de apertura mediante colores.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Ilustración 18. Gráfico de velas japonesas GME.

### Preparación de los datos

#### Análisis de sentimientos: Clasificación de sentimientos en los títulos

#### Librería BERT

Con el modelo BERT (*Bidirectional Encoder Representations from Transformers*) o Representación de Codificador Bidireccional de Transformadores, vamos a tokenizar el texto que queremos analizar. En nuestro ejemplo, el título seleccionado sería el siguiente: *“Math Professor Scott Steiner says the numbers spell DISASTER for Gamestop shorts”*. El primer paso es tokenizar el texto que queremos analizar y el codificador nos devuelve una lista de la codificación del texto que le pasamos por parámetro.

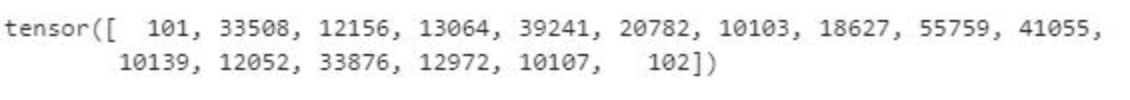


Ilustración 19. Resultado de tokenizar el título ejemplo.

Con el texto codificado podemos alimentar el modelo. El modelo nos devuelve otra lista, *logits*, con 5 valores que representan la probabilidad de que el texto que estamos analizando tenga un sentimiento negativo - negativo/neutro - neutro - positivo/neutro - positivo, respectivamente. El valor máximo de la lista nos indicará cual es el sentimiento predominante del texto analizado.

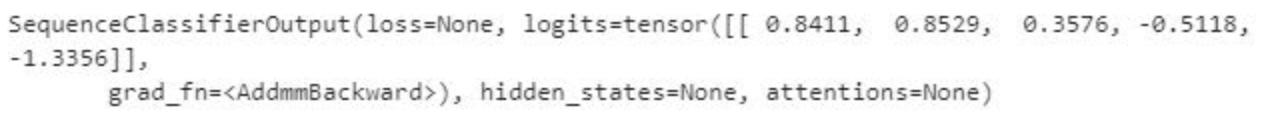


Ilustración 20. Resultados análisis sentimiento negativo-negativo/neutro-neutro-positivo/neutro-positivo.

En el ejemplo podemos ver que para el texto analizado el modelo nos devuelve un sentimiento predominante negativo/neutro.

Para poder extraer el sentimiento predominante de todos los títulos vamos a crear una función que ejecutaremos para cada registro del *dataset*. La función realizará cada unos de los pasos anteriormente explicados.

Aplicamos la función para calcular el sentimiento de los títulos con el modelo BERT y lo almacenamos en una nueva variable en el *dataset*. Para facilitar el desarrollo, y no calcular constantemente el sentimiento de los títulos, una vez generado el sentimiento de los títulos se ha exportado a csv el *dataset*. Si el csv no se encuentra disponible, sería el caso de una ejecución inicial, se calcularía el sentimiento de los títulos del *dataset*.



Ilustración 21.Resultados modelo BERT en el título.

#### Librería NLTK



Ilustración 22. Modelado de los datos titulos.

Separamos en columnas diferentes cada uno de los sentimientos estimados para cada título.

Añadimos un término a cada sentimiento para evitar la situación en la que uno de ellos es igual a cero, tener un valor mínimo nos previene de tener que aplicar una transformación logarítmica en valores cero.

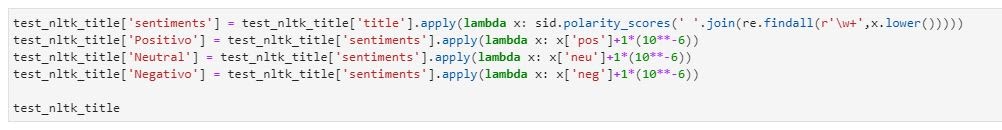


Ilustración 23. Parametrizamos para evitar valor 0.

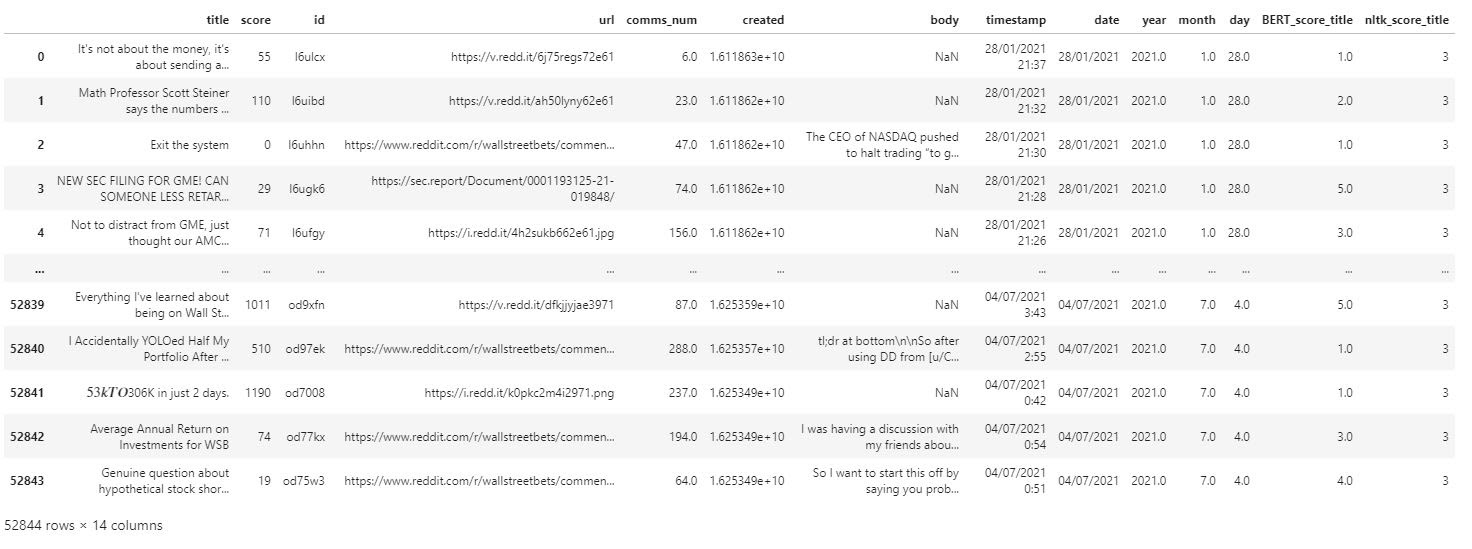


Ilustración 24. Demostración valores sentimientos títulos.

Creamos una función para identificar el sentimiento predominante de los títulos y lo almacenamos en una nueva columna. El resultado se codifica con un valor numérico que responde a: sentimiento negativo = 1, sentimiento neutral = 3, sentimiento positivo = 5.

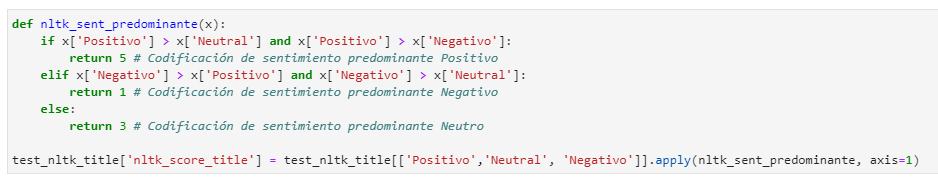


Ilustración 25. Ecuación sentimiento predominante.

Unimos el *dataset* con el *scoring* de los títulos obtenidos con la librería \*NLTK\* al *dataset* \*maestro\*.



Ilustración 26. Unión dataset con scoring obtenido con NLTK.

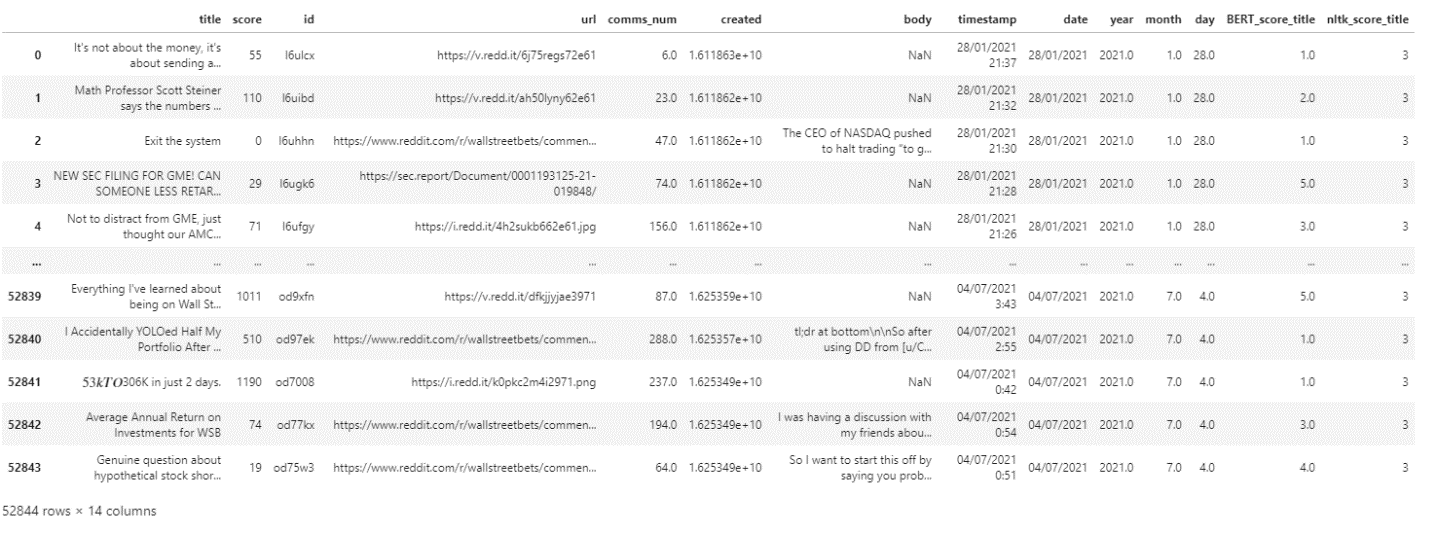


Ilustración 27.Tabla visualización resultados BERT y NLTK.

De una manera similar aplicamos la función para calcular el sentimiento del cuerpo de los *posts* con el modelo BERT y lo almacenamos en una nueva variable en el *dataset*. Para facilitar el desarrollo, y no calcular constantemente el sentimiento de los *posts*, una vez generado el sentimiento de los posts se ha exportado a csv el *dataset*. Si el csv no se encuentra disponible, sería el caso de una ejecución inicial, se calcularía el sentimiento de los *posts* del *dataset*.



Ilustración 28. Cálculo de sentimientos con modelo BERT en los posts.

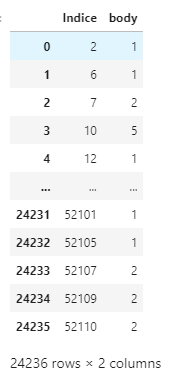


Ilustración 29. Visualización del resultado previo.

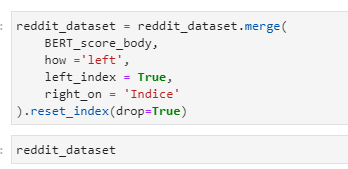


Ilustración 30. Unión de dataset con scoring títulos.

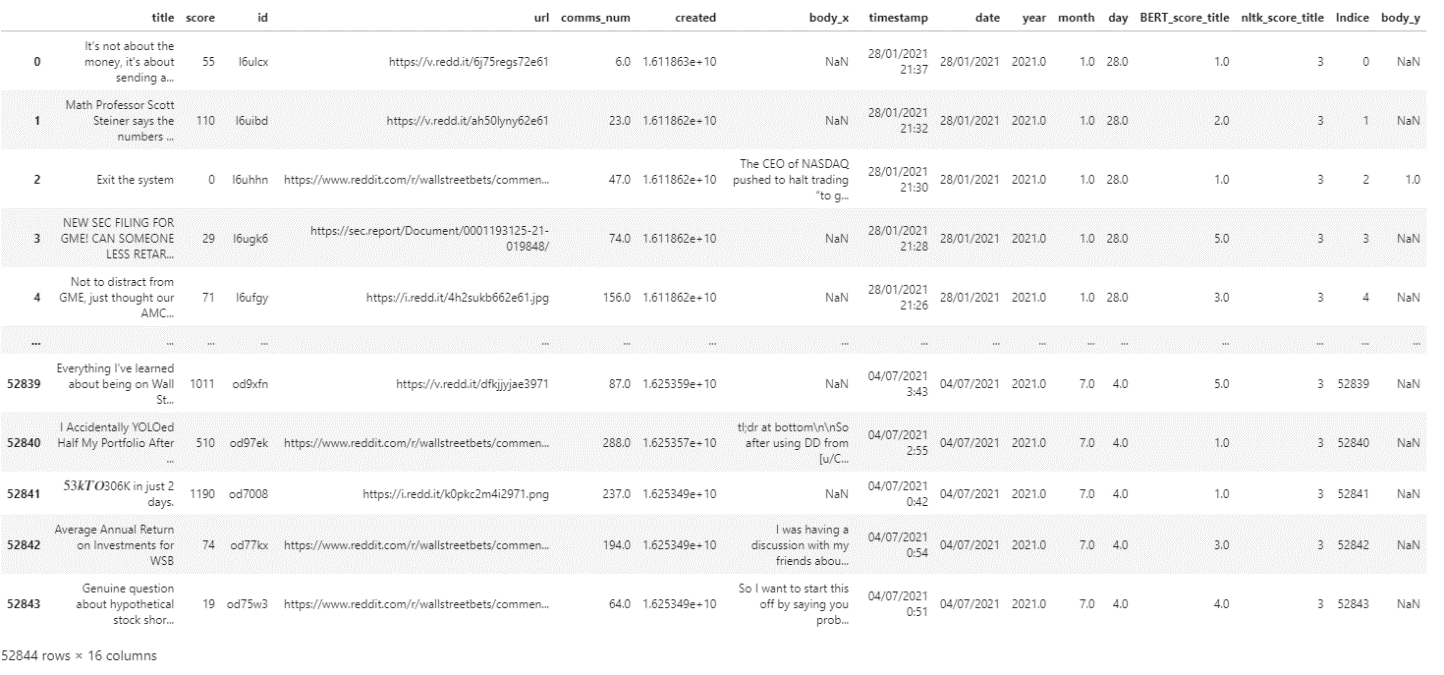


Ilustración 31. Resultados del merge.

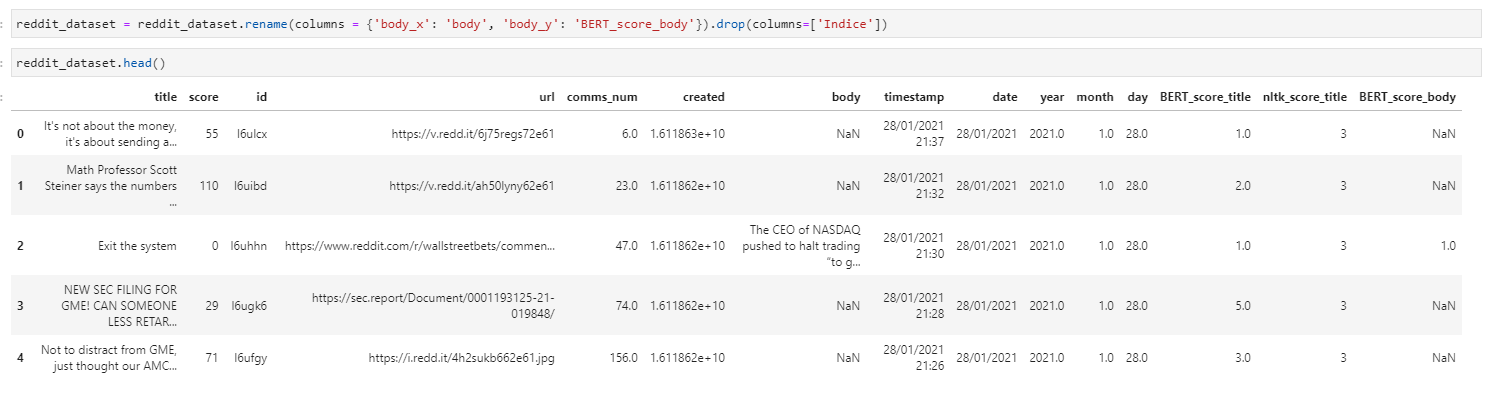


Ilustración 32. Visualización.

#### Librería NLTK



Ilustración 33. Modelado de los datos posts.



Ilustración 34. Demostración valores sentimientos posts.

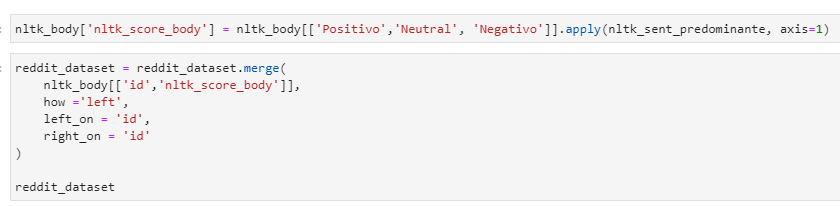


Ilustración 35. Unión de dataset con scoring títulos.

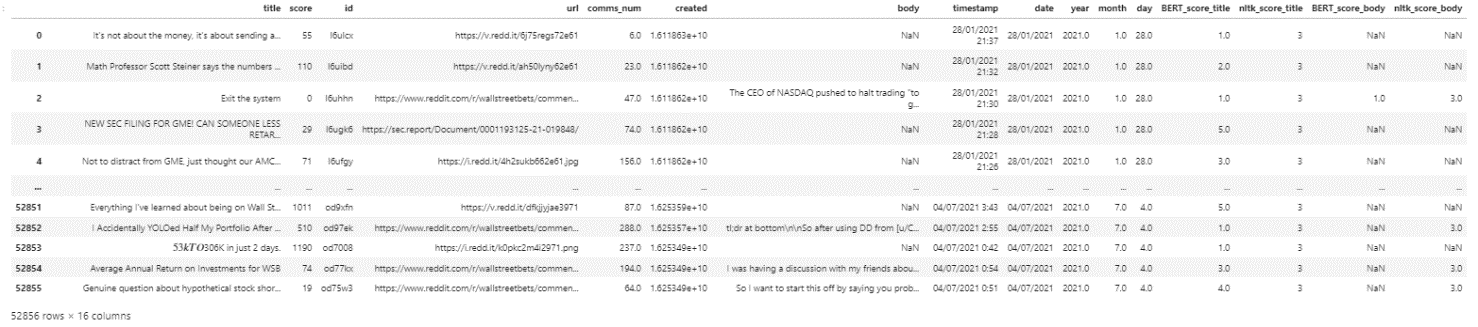


Ilustración 36. Resultados del merge.



Ilustración 37. Datos BERT y NLTK.



Ilustración 38. Reddit Dataset.

### Modelado

**Análisis relación sentimientos y precio de la acción**

Comenzamos por seleccionar los campos que vamos a utilizar para el análisis final. De esta manera las operaciones que realicemos posteriormente serán más rápidas.

https://lh6.googleusercontent.com/U7VaGr9veDc_h1imJ7YftsOobwoFZumaVyPDNGxjpsDl7-MJJ5ZaEtEzmL8zU3t4rR_a3lIOs-ikWazuy317gmV51mpE5zLhAfpBkMVnjDpYGiVK8ij-XZMfR33GHN9ZdgZvzIQe=s0

Ilustración 39.

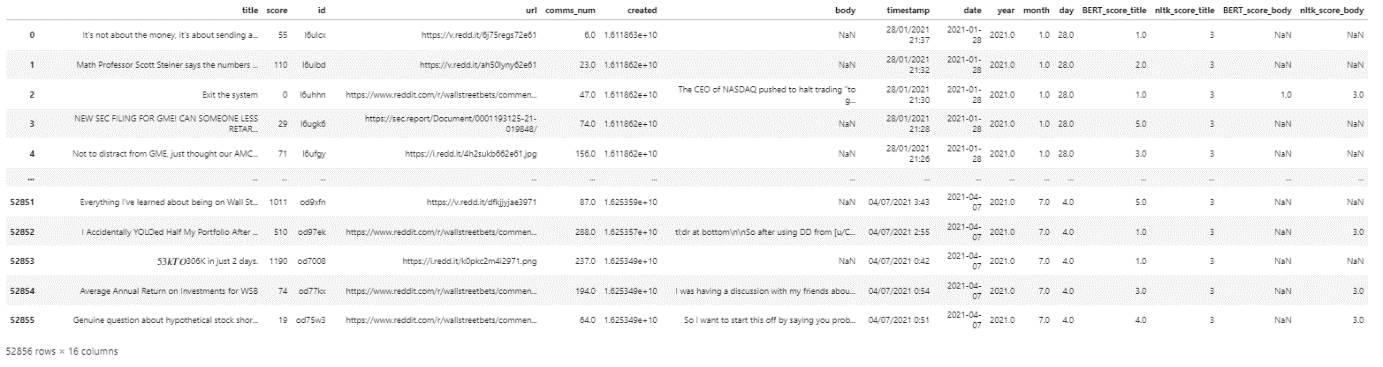


Ilustración 40.

https://lh5.googleusercontent.com/EUn5JcVBaHFvY4StCwUhokUu5Nk9mdEA0espT13T2H3IjqcNMjb99-_xGsGEfMyTotFQv35RaYjc-bOB255GUcDWu71pSWzmTVeuGsXKuRybs1vzsQNotYpBPrT33YwnkAvvdW0O=s0

Ilustración 41.

Creamos una columna para contar cada vez que identificamos un título o post con sentimiento positivo, neutro o negativo para poder realizar el análisis posterior.



Ilustración 42. Generación de columnas por sentimiento.

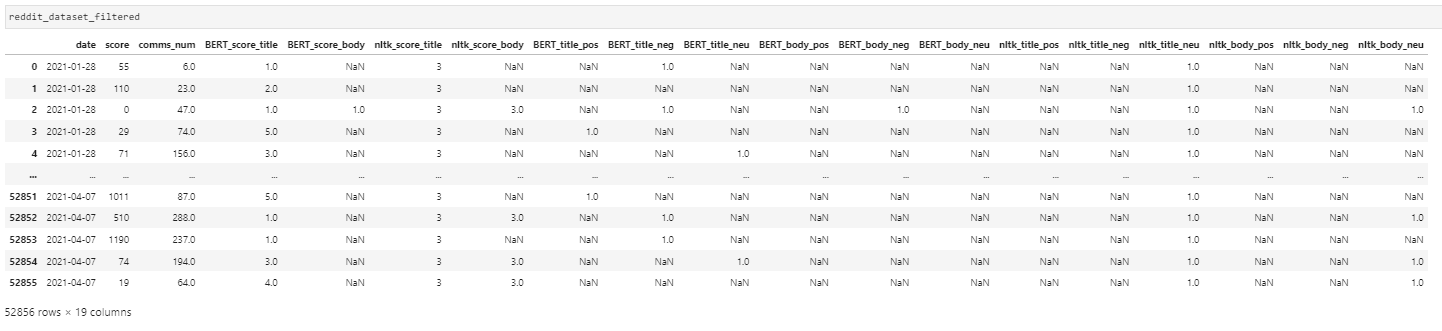


Ilustración 43 . Visualización de resultados.

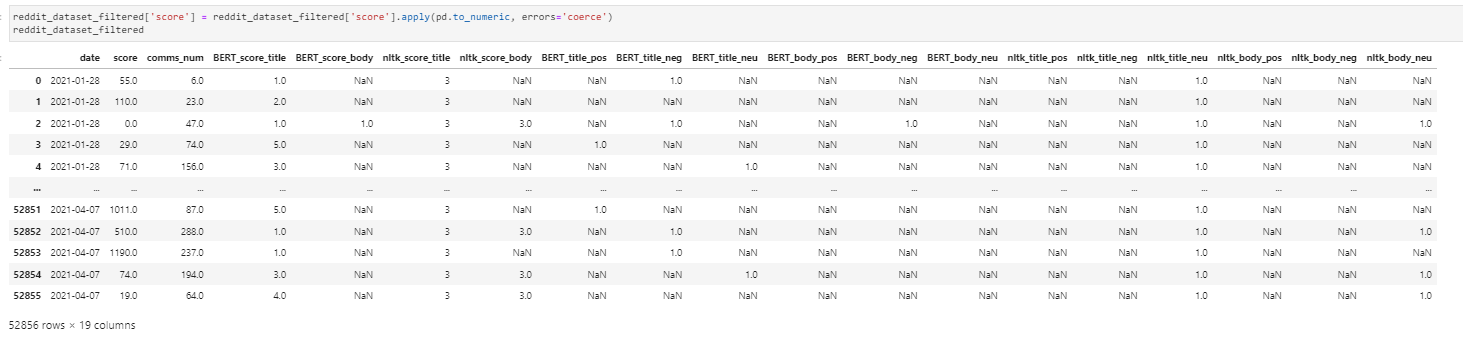


Ilustración 44. Filtros por scoring.

Agregamos los datos para que se ajusten al nivel de agregación de los precios de las acciones, en este caso a nivel de día. Las funciones de agregación aplicadas dependen de la medida.



Ilustración 45. Funciones de agregación.

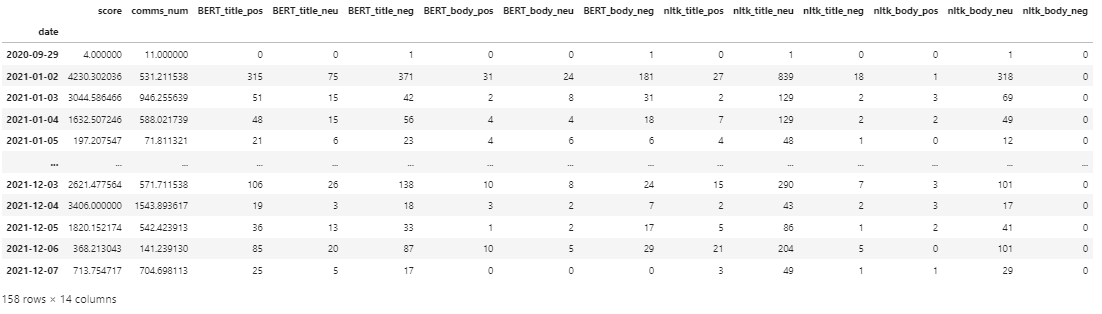


Ilustración 46. Visualización de resultados.

Creamos variables adicionales que agrupan la cantidad de comentarios/títulos, independientemente del sentimiento de cada uno de ellos, para ampliar el análisis.



Ilustración 47. Creación de variables.

Creamos otra serie de variables adicionales, en este caso se tratan de ponderaciones de los sentimientos de los títulos en función de los "likes" medios de cada uno. De esta manera damos mayor valor a los títulos de los posts que hayan tenido más repercusión.

Para calcular las variables se han seleccionado 3 valores de ajuste (0.2, 0.4 y 0.6) que se multiplican, en este caso, por la variable \*score\* que representa los "likes".

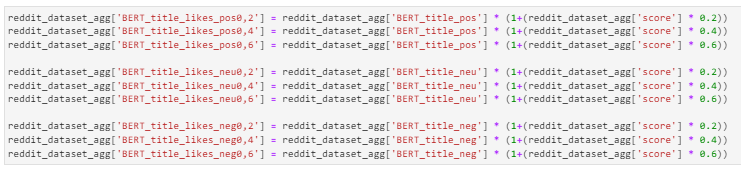


Ilustración 48. Variables por score de representación de likes.

Realizamos la misma operación para los resultados obtenidos del análisis de sentimientos del cuerpo de los posts.

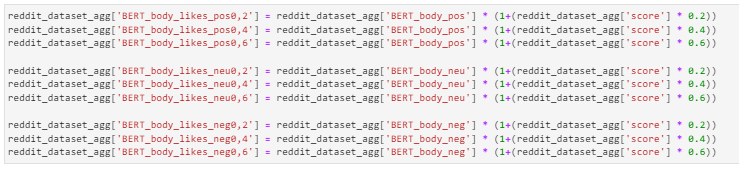


Ilustración 49. Variables por score de representación de posts.

Creamos otra serie de variables muy similares pero en este caso utilizamos otra medida de repercusión, el número medio de comentarios por post. Aplicamos la ponderación a los resultados obtenidos del análisis de sentimientos de los títulos. También aplicando los valores de ajuste comentados anteriormente.

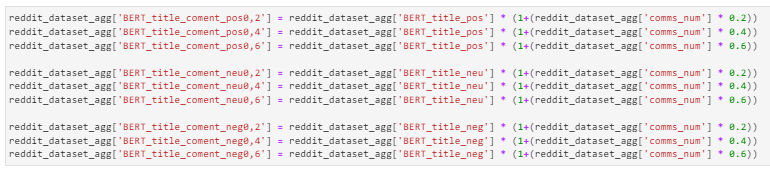


Ilustración 50. Aplicación de ponderación del análisis de los títulos.

Aplicamos la misma operación para los resultados obtenidos del análisis de sentimientos de los *posts*.



Ilustración 51. Análisis de sentimientos de los posts.

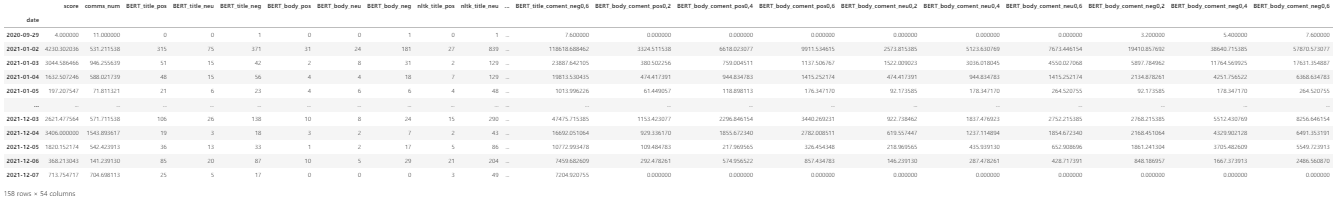


Ilustración 52. Cuadro de visualización.

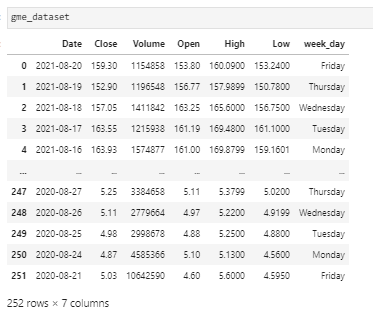


Ilustración 53. Visualización dataset GME.



Ilustración 54. Merge.

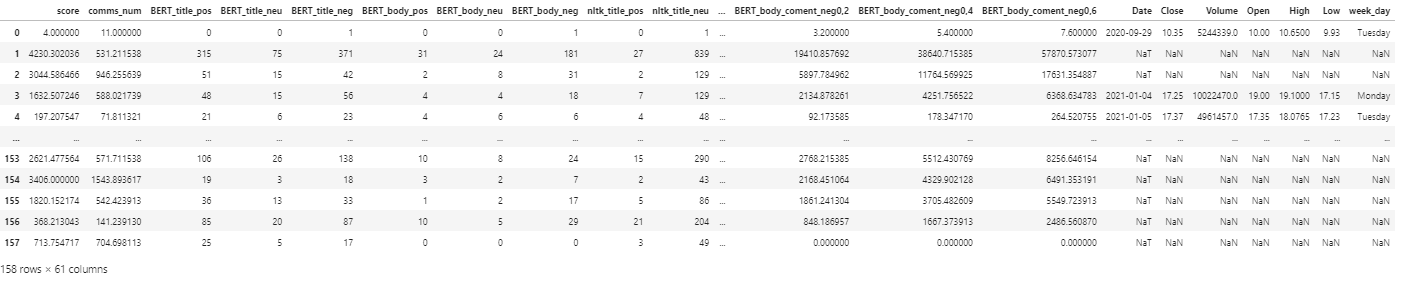


Ilustración 55. Visualización tabla datos.

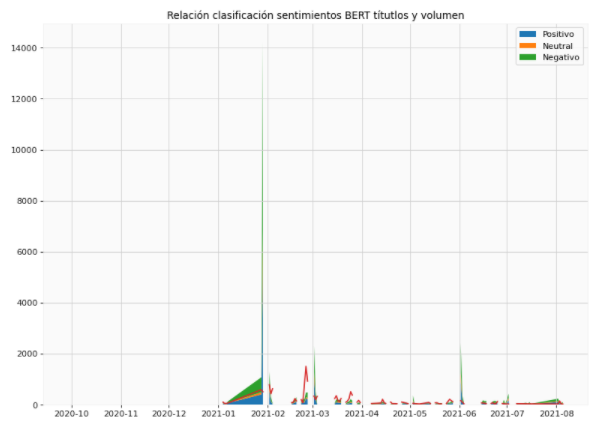


Ilustración 56. Relación clasificación sentimientos BERT.

Repetimos la gráfica, pero esta vez limitando el eje y para hacer zoom en la gráfica.

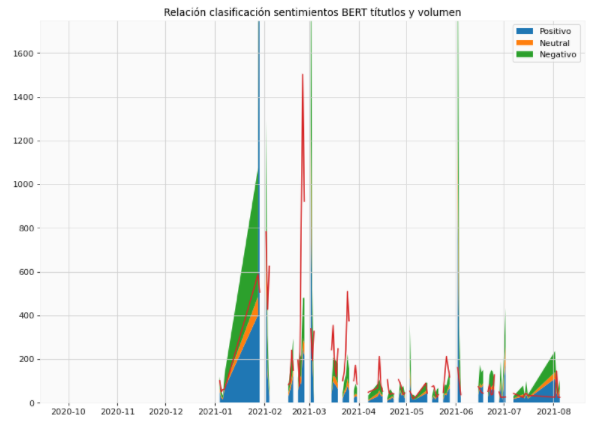


Ilustración 57. Detalle relación clasificación sentimientos BERT títulos y volumen.

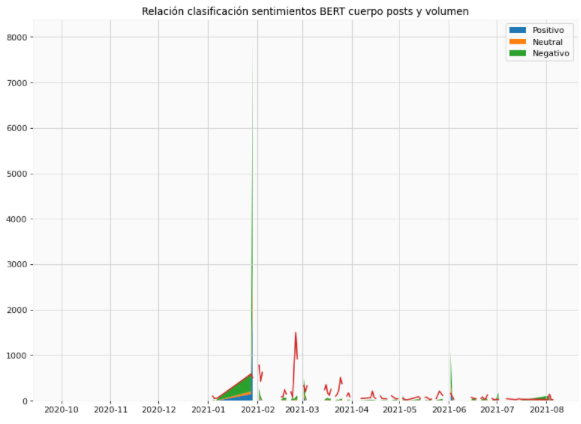


Ilustración 58. Relación clasificación sentimientos BERT post y volumen.

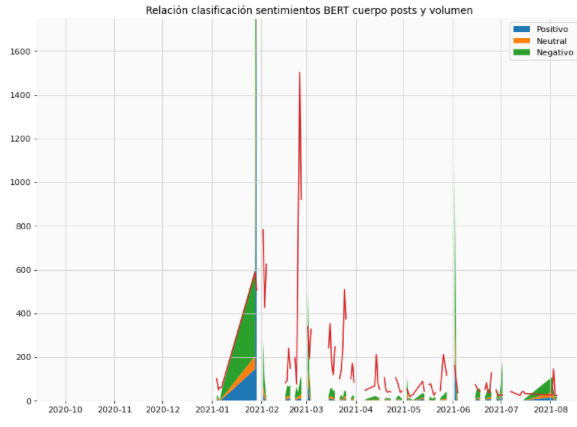


Ilustración 59. Detalle relación clasificación sentimientos BERT post y volumen.

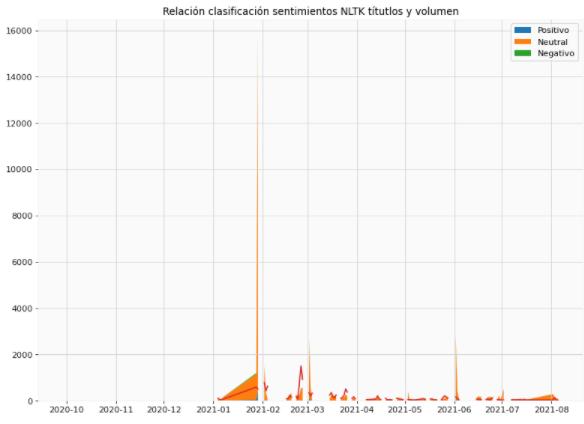


Ilustración 60. Relación clasificación sentimientos NLTK títulos y volumen.

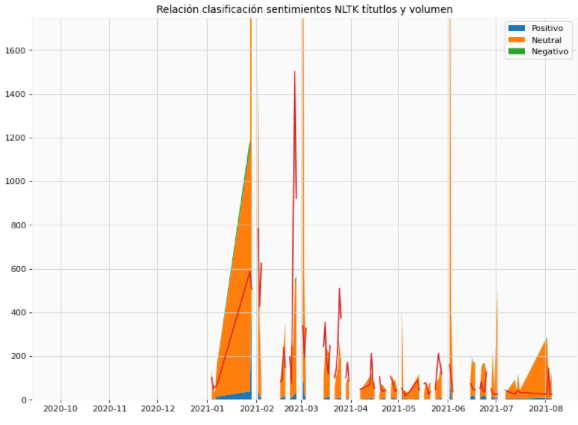


Ilustración 61. Detalle relación clasificación sentimientos NLTK títulos y volumen.



Ilustración 62. Relación clasificación sentimientos NLTK cuerpo posts y volumen.

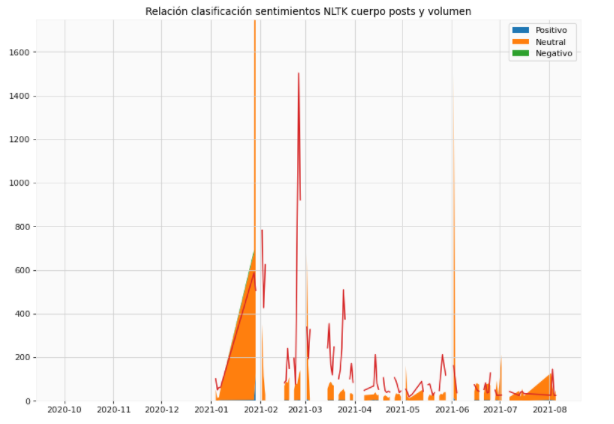


Ilustración 63. Detalle relación clasificación sentimientos NLTK cuerpo posts y volumen.

Generamos la matriz de correlación con las variables obtenidas para analizar si existe alguna relación lineal entre alguna de ellas. Y representamos de forma gráfica la matriz de correlación utilizando un mapa de calor, para facilitar la visualización.

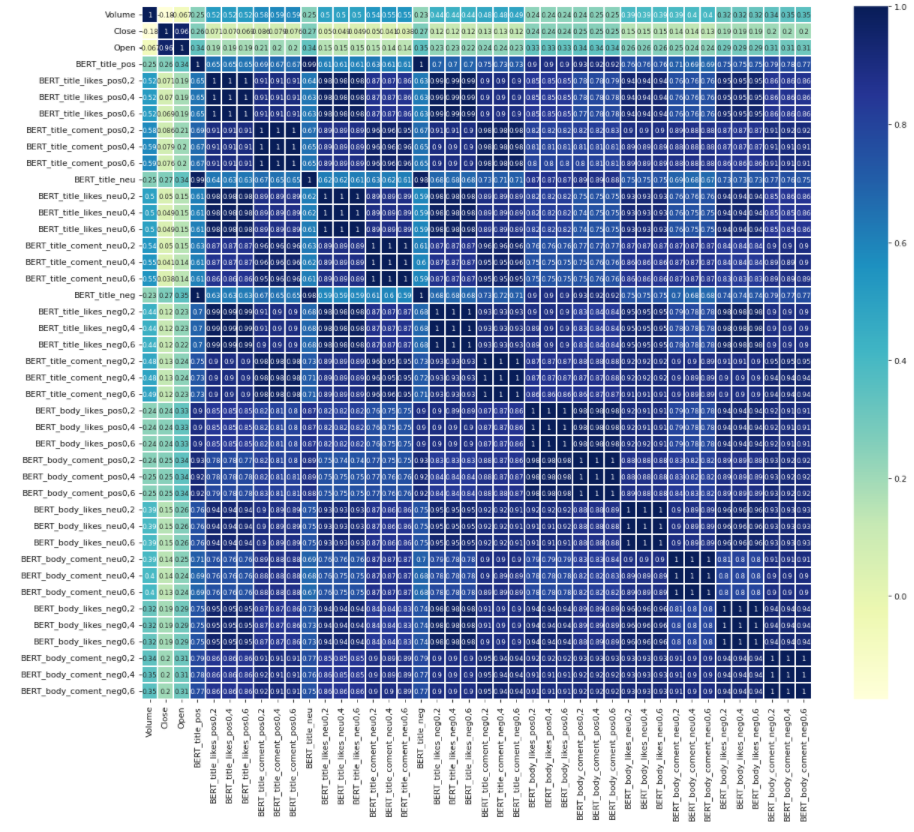


Ilustración 64. Matriz de correlación.

Adicionalmente vamos a analizar las palabras que más se repiten en función del sentimiento del título o post para cada uno de los modelos.

Empezamos por aislar los títulos que han sido identificados como positivos y negativos por el modelo BERT. A continuación, concatenamos los títulos en un solo string.

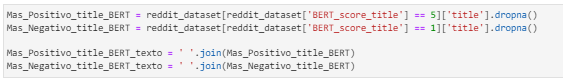


Ilustración 65. Concatenación de los títulos.

Generamos las nubes de palabras. Son una técnica de representación gráfica que permite visualizar las palabras más utilizadas en una cadena de caracteres. Teniendo las palabras con mayor número de apariciones un mayor tamaño.

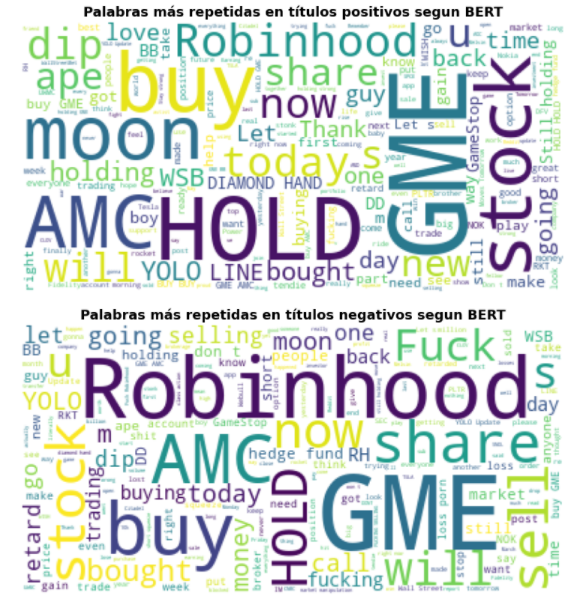


Ilustración 66. Nube de palabras títulos positivos BERT.

Aplicando las mismas operaciones generamos las nubes de palabras para el cuerpo de las publicaciones.



Ilustración 67. Nube de palabras post positivos BERT.

De manera similar generamos las mismas visualizaciones con las palabras más utilizadas de los títulos y cuerpos de los *posts* según la librería NLTK.



Ilustración 68. Nube de palabras títulos positivos y negativos NLTK.

Un periódico con texto

Descripción generada automáticamente con confianza media

Ilustración 69. Nube de palabras títulos positivos y negativos NLTK.

### Evaluación

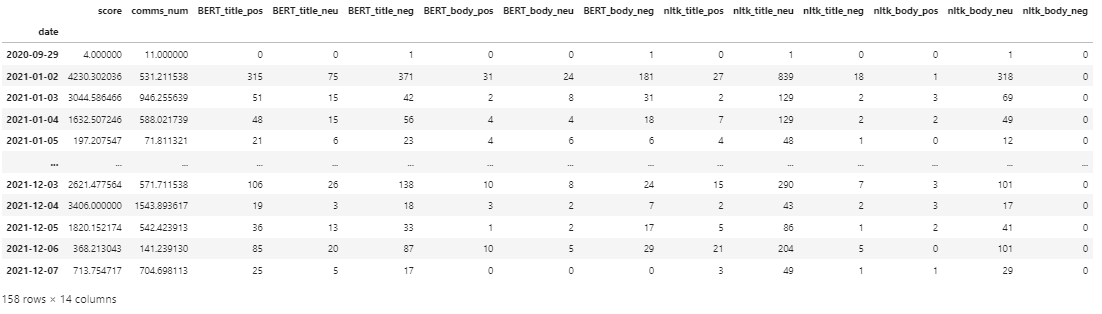
Evaluación de los resultados detallados en la sección **EVALUACIÓN**.

### Despliegue

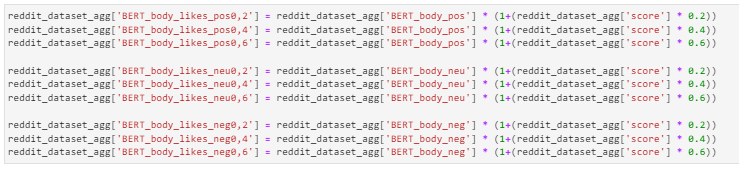
En el caso particular de este proyecto la fase de despliegue, entendido como habilitar un modelo en un entorno de producción para que los usuarios puedan interactuar con él, no ha tenido lugar como tal. Al tratarse de un proyecto de carácter de investigación la fase de despliegue podría considerarse en este propio documento. En particular el apartado **RESULTADOS**, en el que se presentan los resultados obtenidos en el proyecto.

# **EVALUACIÓN**

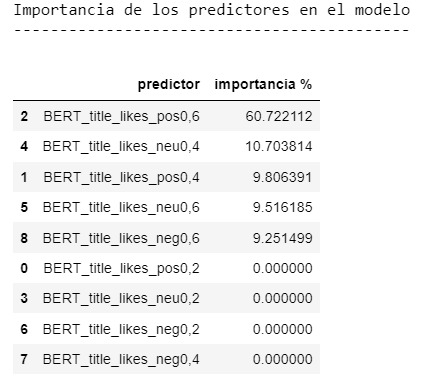
En primera instancia nuestros modelos de análisis de sentimientos tanto para BERT como NLKT trabajan bastante bien y se ajustan a las necesidades de nuestro problema.



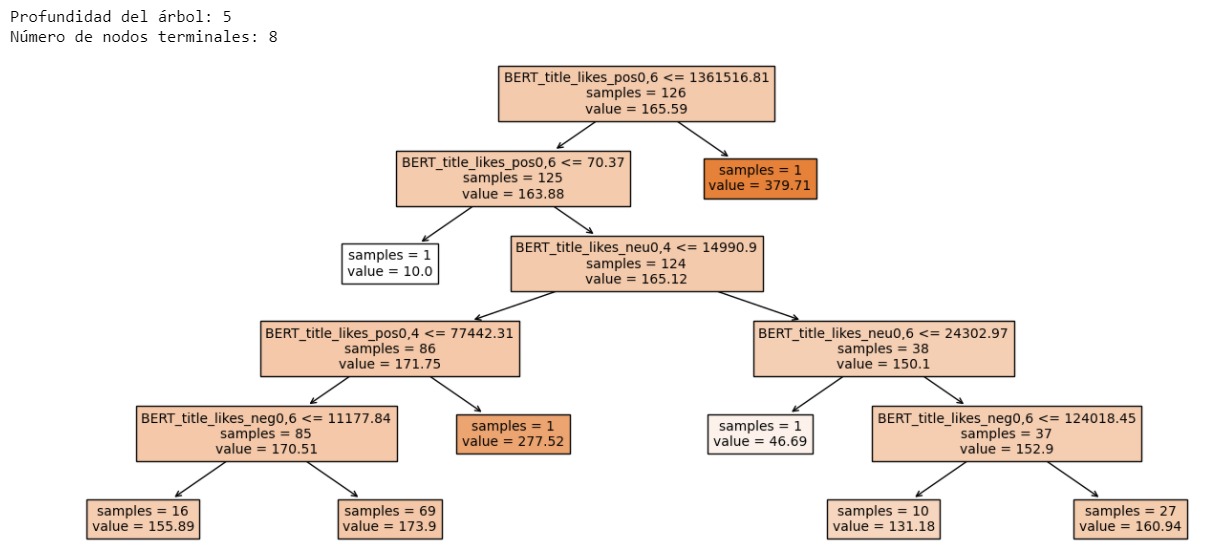
Luego se realizaron algunas transformaciones correspondientes a través de ponderaciones a los scores arrojados para poder trabajar con ellos como variables inputs para nuestro modelo de predicción.



Cada una de estas nuevas variables al utilizarlas en nuestro modelo de predicción tuvieron un impacto respectivo. Solo se utilizaron aquellos scores de likes relacionados a los títulos de los posts utilizando el modelo BERT.



Al construir el Modelo con estas variables y medir la variable respuesta “Precio de apertura” (Open) se llegó a la conclusión de que el RMSE del modelo bordeaba las 45 unidades. Si bien el error no es el mejor, ya que se esperaba uno más bajo, tampoco está del todo mal tomando en cuenta que el valor máximo del precio de apertura supera las 300 unidades.



# **RESULTADOS**

Se logró exitosamente entrenar 2 modelos de análisis para medir los sentimientos de las publicaciones del sub-foro WallStreetBets a través de scores para los likes y comentarios de los títulos y el cuerpo del post.

Los modelos utilizados BERT y NLKT nos permitió adecuar los scores para etiquetarlos principalmente en 3 categorías: negativo, neutro y positivo. Esta información fue de vital importancia para realizar el cruce de información por fecha con los datos de los precios/volúmenes de operaciones financieras de GME.

Al tener la data consolidada de Reddit y GME se generó una matriz de correlaciones para determinar que variables de score se relacionan más con las variables de precios y volúmenes. Las variables de scoring que guardaban relación con las variables de precio y volumen no fueron muchas, por lo que las seleccionadas para nuestro modelo de predicción solo fueron aquellas relacionadas a BERT\_title\_likes (scores de likes de títulos utilizando el modelo BERT).

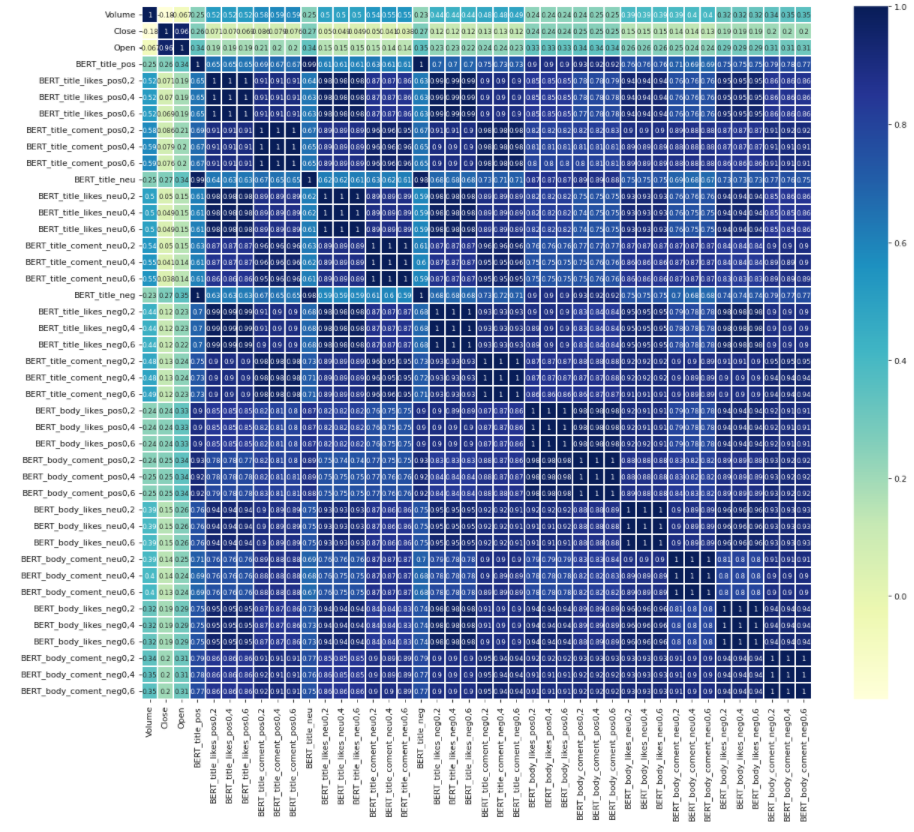
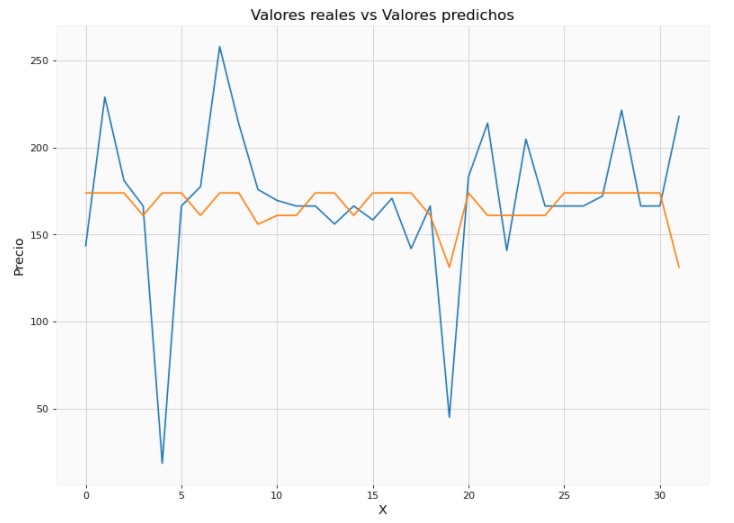


Ilustración 64. Matriz de correlación

Después de determinar dichas variables se modeló un Árbol de Decisión para predecir el Precio de Apertura en función de los scores de las publicaciones hallados previamente con el análisis de Sentimientos.

El RMSE (Error cuadrático medio) hallado en nuestro modelo no fue el mejor pero tampoco es desalentador. Por lo que se puede afirmar que el Modelo funciona bastante bien.



# **CONCLUSIONES Y TRABAJOS FUTUROS**

*Los resultados obtenidos después de realizar el análisis de sentimientos a la data de WallStreetBets* *demuestran que esta técnica es una de las más poderosas para clasificar las publicaciones. No solo porque es precisa, sino también porque debido a su dinamismo permite utilizar sus resultados en trabajos posteriores, como es el caso de nuestro análisis de predicción basado en un Árbol de Decisión (análisis de regresión) para predecir el Precio de Apertura de GME a partir de los títulos y el cuerpo de las publicaciones de WallStreetBets.*

*Este modelo de árbol de Decisión aún tiene oportunidad de mejora, ya que se puede optimizar a través del método de Pruning (Podar el árbol) que incluye la optimización de los parámetros del modelo a través de la validación cruzada.*

*Otro de los trabajos futuros a considerar es la incorporación de otras variables al modelo. Muchos de los comentarios en Reddit son imágenes, vídeos o enlaces a webs externas. Incorporar el análisis a estos datos nos podría aportar un horizonte más donde posiblemente obtener resultados. Una propuesta de formalización de este proceso se ha realizado ya en el documento Jupyter Notebook utilizando el ejecutable Tesseract para convertir imágenes en texto raw al cual se le podría PLN tras filtrar los resultados: actualmente los textos obtenidos por esta API diferencian texto útil y texto no útil.*

**

Ilustración 70. Captura de pantalla de la imagen 200 del dataset r/WallStreetBets.

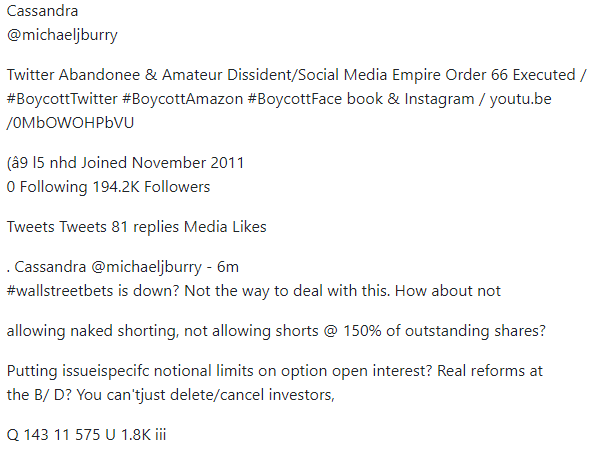


Ilustración 71. Texto raw obtenido de utilizar Tesseract en la Ilustración 69.

*Este mismo método podría utilizarse también para los vídeos y GIF’s del dataset seleccionando cada qué número de frames nos interesa aplicar tessseract y tratando los distintos frames como imágenes estáticas. Para aplicar este método se debería diseñar también un sistema de refinamiento capaz de eliminar redundancias (si el mismo texto permanece en distintos frames del video) y de filtrar el texto no útil.*

# **Referencias**

Ming Zhou, N. D. (13 de 12 de 2018). *microsoft.com*. Obtenido de microsoft.com: https://www.microsoft.com/en-us/research/lab/microsoft-research-asia/articles/next-10-years-natural-language-processing/

Wikipedia. (6 de Agosto de 2021). *Wikipedia*. Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Caso\_GameStop

www.jmp.com. (s.f.). *www.jmp.com*. Obtenido de www.jmp.com: https://www.jmp.com/es\_pe/statistics-knowledge-portal/exploratory-data-analysis.html

www.sngular.com. (s.f.). *www.sngular.com*. Obtenido de www.sngular.com: https://www.sngular.com/es/data-science-crisp-dm-metodologia/

Zaman, M. I. (16 de 1 de 2021). *Twitter Sentiment Analysis using NLP and Neural Networks Techniques medium.com*. Obtenido de medium.com: https://medium.com/analytics-vidhya/introduction-bd62190f6acd