PROYECTO FIN DE GRADO

TÍTULO: Diseño y desarrollo de un sistema de vigilancia según la metodología CRISP-DM. Validación mediante un caso práctico sobre la influencia de las redes sociales en el mercado financiero.

AUTOR: Alberto Antonio López Pérez-Villamil

TITULACIÓN: Grado en Ingeniería Telemática

TUTOR: Aurelio Berges García

DEPARTAMENTO: Ingeniería de Telemática y Electrónica

VºBº TUTOR

Miembros del Tribunal Calificador:

PRESIDENTE: Carlos Felipe Rueda Frías

TUTOR: Aurelio Berges García

SECRETARIO: Óscar Ortiz Ortiz

Fecha de lectura: 19/07/2024

Calificación:

El Secretario,

Resumen

En la actualidad, todas las empresas tienen acceso a una gran cantidad de datos, tanto internos como externos. Durante mucho tiempo, se han estado implementando soluciones y enfoques para aprovechar esos datos y convertirlos en información útil para la toma de decisiones.

Una manera de obtener información valiosa es mediante el uso de la vigilancia e inteligencia, un proceso que engloba todas las etapas, desde la recopilación de los datos hasta la distribución de la información a los responsables de la toma de decisiones dentro de la organización.

En este proyecto de fin de grado se lleva a cabo la creación de un modelo genérico de un sistema de vigilancia e inteligencia, siguiendo las pautas establecidas por metodología CRISP-DM. Además, con el objetivo de validar este modelo, se realiza un estudio de caso práctico en el que se implementan las diversas etapas características de un sistema de vigilancia e inteligencia. El objetivo es obtener información de calidad sobre la influencia o relación de los comentarios de los usuarios de la red social Reddit sobre el mercado financiero.

Abstract

Nowadays, all companies have access to a large amount of data, both internal and external. For a long time, solutions and approaches have been implemented to harness this data and turn it into useful information for decision making.

One way to obtain valuable information is through the use of surveillance and intelligence, a process that encompasses all stages from data collection to the distribution of information to decision makers within the organization.

In this final degree project, the creation of a generic model of a surveillance and intelligence system is carried out, following the guidelines established by the UNE 166.006:2018 standard. In addition, with the aim of validating this model, a practical case study is carried out in which the various characteristic stages of a surveillance and intelligence system are implemented. The objective is to obtain valuable information on the influence or relationship of the comments made by users of the Reddit social network on the financial market.

Índice de contenidos

Índice de figuras

Índice de tablas

Lista de acrónimos

AENOR	Asociación Española de Normalización y Certificación
API	Application Programming Interface
CRISP-DM	Cross-industry Standard Process for Data Mining
EFPA	European Financial Planning Association
ETF	Transferencia electrónica de fondos
NoSQL	No Structured Query Language
KDD	Knowledge Discovery in Databases
KPI	Key Performance Indicator
ODS	Objetivos de Desarrollo Sostenible
OLAP	Online Analytical Processing
PFG	Proyecto de Fin de Grado
P3QT	Product Place Price Time Quantity
SEMMA	Sample Explore Modify Model Assess
SQL	Structured Query Language

Capítulo 1. Introducción y objetivos del proyecto

1.1 Contexto y justificación del proyecto

En los últimos años la información a la que pueden acceder las organizaciones, tanto externa como internamente, ha crecido exponencialmente [1].

Hay dos tipos de información, externa e interna. En cuanto a la primera, se pueden diferenciar los datos no fiables, información generalmente gratuita que se puede usar libremente y que no verificada, o los datos fiables, los compartidos u obtenidos a través de APIs u otros medios, pero provenientes de fuentes verificadas y con experiencia en el tema que está tratando [2]. Con respecto a la información interna, se observan diferentes tipos, como la proveniente del departamento financiero, la de recursos humanos, etc. [3]

Además de su origen, interno o externo, la información se puede presentar de manera estructurada, es decir, en datos cuantitativos que pueden ordenar fácilmente en bases de datos relaciones. También se puede dar que sea información semiestructurada, datos que no siguen un modelado fijo y, por lo tanto, no se pueden introducir en bases de datos relacionales de una manera tan sencilla como los anteriores. Por último, se tiene la información no estructurada, se trata de datos cualitativos que no encajan en modelos predefinidos [4].

La información a la que se expone una organización crece de manera exponencial y puede ser determinante para llevar a cabo diferentes actividades o tareas, es decir, para la toma de decisiones. También, sus formatos pueden variar, lo que dificulta que se pueda organizar.

Para ayudar a las organizaciones en el procesado de toda la información que les rodea, se han desarrollado diversas metodologías, entre ellas CRISP-DM. Esta especifica cómo desarrollar un proyecto de explotación de datos. Así, en este Proyecto de Fin de Grado (PFG) se aplica esta metodología para apoyar el desarrollo de un sistema de vigilancia e inteligencia, para procesar una gran cantidad de datos y transformarlos en información de calidad. La metodología facilita la estructuración y normalización de los procesos para la obtención de la información, su análisis y, finalmente, su implementación y comunicación al entorno de la empresa u organización [5]. Un ejemplo del funcionamiento de estos sistemas, son las empresas de la comunidad valenciana, que están implementando soluciones para analizar grandes volúmenes de datos para identificar patrones e información que antes no eran detectables. Esto les ha provocado una mejorara de su competitividad, gracias a la optimización de procesos y provocando una adaptación a sus clientes [6].

En este PFG se propone realizar un sistema de vigilancia e inteligencia, que siga la metodología CRISP-DM, con la finalidad ayudar a las organizaciones en la toma de decisiones.

1.2 Objetivos técnicos y académicos

El propósito fundamental de este proyecto consiste en crear un **modelo genérico de un sistema de vigilancia e inteligencia, de acuerdo con las pautas establecidas por la metodología CRISP-DM**, para gestionar información de manera aplicable a cualquier tipo de organización. Además, se llevará a cabo una validación del modelo de sistema de vigilancia e inteligencia mediante el estudio de un caso práctico.

Con la implementación de este caso práctico, se busca simplificar las etapas de adquisición, procesamiento, almacenamiento, análisis y difusión de información. Entre los objetivos secundarios establecidos para este PFG se incluyen los siguientes:

- Diseñar e implementar los pasos necesarios para procesar los datos obtenidos de las fuentes de información, a partir de diferentes archivos que contengan la información, y así obtener datos normalizados.
- Analizar y seleccionar las herramientas y plataformas disponibles en el mercado actual para el desarrollo e implementación de las distintas etapas que conforman el sistema de vigilancia e inteligencia propuesto en este proyecto.
- Implementar un *data warehouse* para almacenar los datos obtenidos en las fases del sistema de vigilancia e inteligencia.
- Obtener información de calidad sobre la red social Reddit, como parte del análisis de información para la validación del caso práctico.

1.3 Metodología del trabajo

Para este proyecto se va a seguir una metodología que tenga fases claras y estructuradas, con la flexibilidad necesaria para poder adaptarse a cambios y necesidades del proyecto. Una metodología que cumple estos requisitos es *Waterfall* con Iteraciones Ágiles. Consta de las siguientes fases, tal como se puede apreciar en la Figura 1.1 [7]:

- Definición y análisis: Establecer los objetivos, identificar los requisitos y analizar las necesidades del sistema.
- Diseño: Crear la arquitectura y el diseño detallado del sistema de vigilancia e inteligencia.
- Implementación: Desarrollar el sistema de acuerdo con el diseño aprobado. Es posible volver a modificar el diseño si fuera necesario.
- Pruebas: Validar que el sistema cumple con los requisitos.
- Despliegue y Entrega: Implementar el sistema en un entorno productivo.

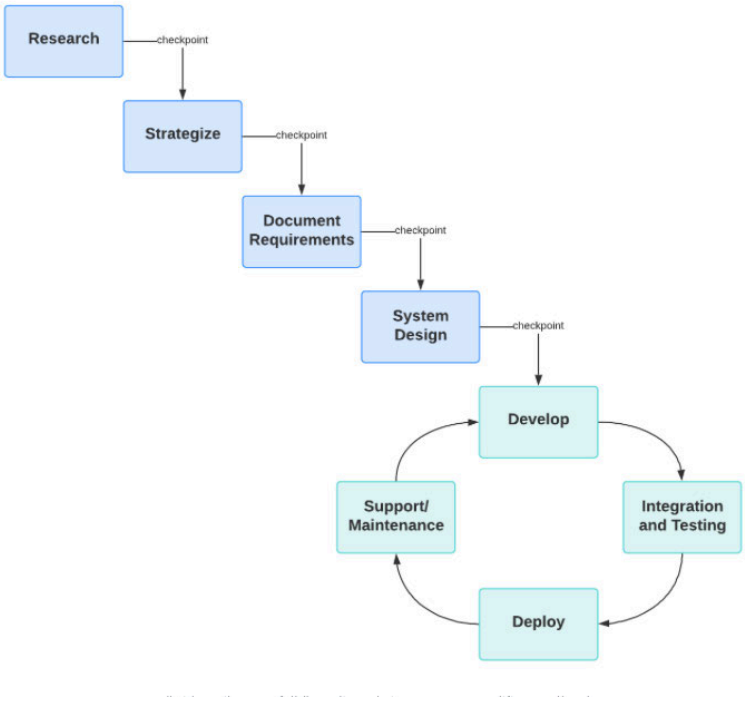

Figura 1.1 - Metodología híbrida Waterfall-Agile. Fuente [7].

1.4 Estructura del resto de la memoria

A continuación, se proporciona una breve descripción del contenido de los distintos capítulos que conforman la memoria de este PFG:

- Capítulo 2: Estado del arte. En este capítulo se exponen detalladamente las características de la metodología CRISP-DM, junto con otras normas o metodologías que también se pueden utilizar. Además, se exploran las numerosas herramientas y plataformas necesarias para implementar cada una de las fases del sistema de vigilancia en inteligencia según CRISP-DM.

- Capítulo 3: Necesidad de información para la toma de decisión en las organizaciones. Planteamiento del problema. En este capítulo se plantea el problema y se presenta el contexto en el cual los sistemas de vigilancia e inteligencia operan, junto con los desafíos que enfrentan las organizaciones al gestionar la información y adquirir conocimiento.

- Capítulo 4: Implementación de un sistema de vigilancia e inteligencia según la metodología CRISP-DM. En este capítulo se propone un modelo genérico de un sistema de vigilancia e inteligencia.

- Capítulo 5: Validación del modelo propuesto mediante un caso práctico. En ese capítulo se valida dicho modelo a través de la realización de un caso práctico. Se describe paso a paso la implementación del caso práctico, incluyendo los procesos llevados a cabo y las herramientas utilizadas para establecer un sistema de vigilancia e inteligencia.

- Capítulo 6: Presupuesto y planificación del proyecto. En este capítulo se presenta el presupuesto seguido para la realización de este Proyecto de Fin de Grado, así como su planificación.
- Capítulo 7: Impacto del proyecto: En este capítulo se exponen las implicaciones ambientales, sociales, tecnológicas y económicas del proyecto, así como su contribución a los Objetivos de Desarrollo Sostenible (ODS).
- Capítulo 8: Conclusiones y desarrollo de trabajos futuros. En este capítulo se presentan las conclusiones obtenidas a partir de la realización de este proyecto, así como las posibles áreas de desarrollo para futuros trabajos, con el objetivo de ampliar y mejorar el presente Proyecto de Fin de Grado (PFG).
- Capítulo 9: Referencias. En este capítulo se presentan toda la bibliografía utilizada para el desarrollo del PFG.

Capítulo 2. Estado del arte

2.1 Introducción

En los últimos años la sociedad está viviendo una revolución tecnológica que está provocando un crecimiento exponencial de la creación de datos provenientes de Internet. En el año 2019, la información generada por todos los sistemas que se utilizan alcanzó la cifra de 45 zettabytes, (1 zettabyte equivale a 1.000 millones de terabytes), y según diferentes estudios [8] [9], para el año 2025 es probable que esta cantidad de datos generada llegue a los 175 zettabytes [10]. Este fenómeno se debe al aumento del número de dispositivos conectados, el aumento del crecimiento electrónico y el uso, por cada vez más personas, de las redes sociales.

Lo realmente importante de toda esta gran cantidad de datos obtenida de Internet, es que son solamente datos que, sin un proceso previo de refinado y procesamiento para transformarlos en información útil, no van a servir de nada a las personas u organizaciones que quieran acceder a ellos. Esto último, **la transformación de datos en información es el reto que se está viviendo en la actualidad**. La capacidad de analizar y extraer puntos clave de los datos es fundamental para la toma de decisiones informadas, que pueden mejorar la competitividad y eficiencia de las organizaciones [11].

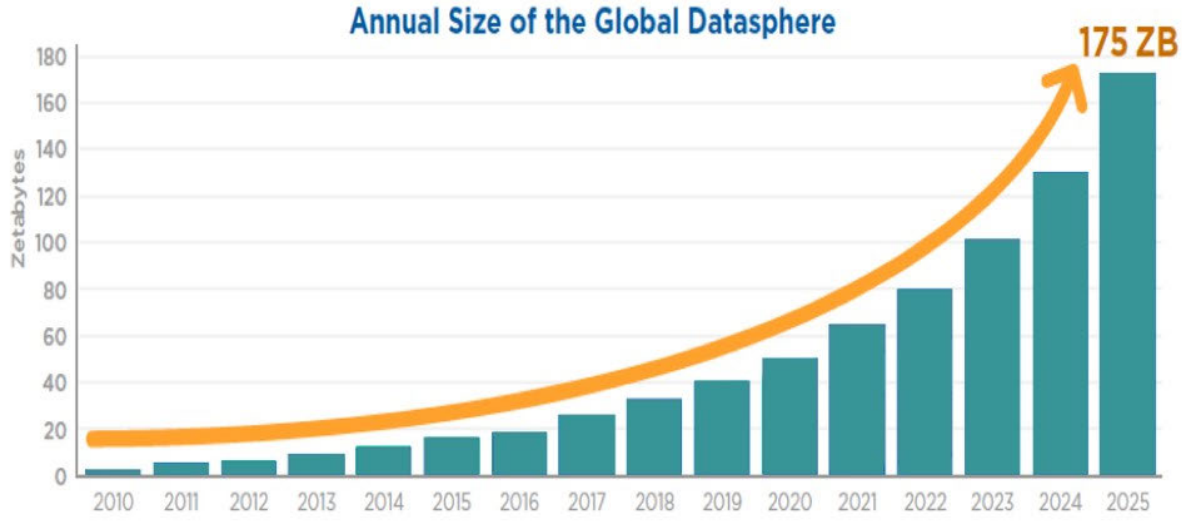

Figura 2.1 - Cantidad anual de datos compartidos globalmente. Fuente: [1]

En lo relativo al mundo empresarial, el volumen de datos que las organizaciones manejan o gestionan incrementó un 569% entre los años 2016 y 2018, de 1’45 Petabytes a 9’20. La gran mayoría de las empresas, entorno al 92%, sabe del valor que tiene esta cantidad de datos, y un gran número de ellas, ha empezado a llevar a cabo prácticas que les obtengan un rendimiento económico de estos [12].

A raíz de este incremento en la cantidad de información disponible, las organizaciones han comenzado a experimentar la necesidad de convertir dicha información en conocimiento. Para lograr ello, se requiere el uso de metodologías, que indiquen los pasos y procesos que hay que seguir para analizar toda esta información con el fin de obtener datos de utilidad que respalden la toma de decisiones, identificación de amenazas potenciales, adelantarse a las

oportunidades de mercado, prever tendencias, conocer necesidades de los clientes o anticiparse a cambios, entre otras.

La **minería o explotación de datos**, que se centra en el procesado de grandes cantidades de datos para transformarlos en información útil, investiga técnicas para la obtención de patrones y la explotación de la información. Este proceso da lugar a diferentes metodologías que consiguen que la minería de datos sea sistemática [13]. Estas metodologías ayudan a la organización a entender los diferentes procesos que forman la explotación de datos, desde el descubrimiento del conocimiento hasta planificar proyectos y ejecutarlos. Definen las tareas que se deberían llevar a cabo en cada proceso, además de precisar las etapas de ejecución.

A mediados de los años 90, surge la primera metodología conocida como **KDD** (*Knowledge Discovery in Databases*), en ella se describe un proceso interactivo e iterativo organizado en diversas etapas, donde destacan la integración y recopilación de los datos, preparación de estos, minería, evaluación, difusión y uso de modelos [14].

En los años 2000, surgen más metodologías que definen sistemas concretos para llevar a cabo proyectos de explotación de datos, entre los cuales se encuentran **Catalyst, CRISP-DM** (*Cross-industry Standard Process for Data Mining*) y **SEMMA** (*Sample, Explore, Modify, Model, Assess*) [15]. Catalyst se centra en el análisis de datos en tiempo real, mientras que CRISP-DM presenta un enfoque estructurado y fiable para la minearía de datos. SEMMA, desarrollada por SAS, ofrece una guía de la minería de datos mediante diferentes fases [16].

Además de las metodologías, también surgen normas como la **UNE 166.006:2018**. Esta norma tiene como propósito demostrar, de una manera sencilla, la implementación un sistema de vigilancia e inteligencia en cualquier organización, sin importar su estructura o tamaño. Este sistema tiene la finalidad de proporcionar conocimiento a los ejecutivos de las organizaciones en relación con un área temática específica en la que requieran información valiosa, a partir de datos obtenidos de diferentes fuentes [17].

En el siguiente apartado se van a describir dichas metodologías y normas de manera más específica, ofreciendo un marco detallado sobre cómo éstas pueden ser aplicadas para mejorar la toma de decisiones y la competitividad en las organizaciones. Además, en siguientes apartados, se realizará una comparativa donde se discuten las ventajas e inconvenientes asociados a la implementación de cada una de estas metodologías.

2.2 Metodologías para la explotación de datos

En este apartado se va a realizar una descripción detallada de las metodologías existentes para el desarrollo de proyectos de explotación de datos, también conocida como minería de datos. Estas metodologías ofrecen marcos estructurados que guían a las organizaciones en el proceso de transformación de grandes volúmenes de datos en información de utilidad.

2.2.1 Metodología KDD (Knowledge Discovery in Databases)

KDD fue el primer modelo que define la obtención de información procedente de bases de datos como un proceso o etapa, compuesto por diversas fases que van desde el preparado de

los datos, hasta su análisis, interpretación y difusión [18]. Además, es un proceso repetitivo, porque el resultado de cada una de sus fases permite retroceder a fases o etapas anteriores y porque son necesarias varias repeticiones para obtener datos de calidad. También es un proceso interactivo, porque permite que el usuario ayude en la preparación de datos y valide la información extraída [18] [19]. A continuación, se presentan las nueve etapas que forman la metodología [20]:

1. **Comprensión del dominio de la aplicación**: En esta etapa, siempre desde el punto de vista del usuario, se debe recolectar la mayor cantidad de conocimiento posible y llevar a cabo la identificación de objetivos del proceso.
2. **Creación del conjunto de datos**: En esta segunda etapa, se eligen las fuentes de datos que se van a utilizar en el proceso y se selección los atributos que se van a tener en cuenta para realizar la minería de datos.
3. **Limpieza y preprocesamiento de los datos**: Eliminación tanto de datos nulos como anómalos.
4. **Reducción y proyección de los datos**: Detectar características de los datos útiles para su representación. Se usan técnicas para la reducción de la dimensionalidad de datos y de la cantidad de variables.
5. **Determinar la tarea de minería de datos:** Determinar qué tipo de minería se va a utilizar para llevar a cabo el análisis, como regresión, agrupamiento, asociación o agrupación.
6. **Determinar el algoritmo de minería**: Dependiendo de lo escogido en el punto anterior, se tiene que elegir un algoritmo propio para la búsqueda de patrones.
7. **Minería de datos:** Etapa en la que se aplica el algoritmo seleccionado anteriormente.
8. **Interpretación**: Traducir y visualizar los patrones obtenidos.
9. **Utilización de la nueva información obtenida**: Implementar el conocimiento adquirido en la toma de decisiones, incluyendo su verificación e identificación de conflictos con información de calidad previamente obtenida.

Si bien estas son las nueve etapas que forman KDD, la metodología se puede resumir en cinco etapas, tal como se puede apreciar en la Figura 2.2:

1. Seleccionar los datos.
2. Procesamiento previo de los datos para seleccionar los que son correctos.
3. Transformar los datos y su reducción.
4. Obtener patrones de interés según el tipo de minería de datos que se esté utilizando (predictiva o descriptiva).
5. Interpretar el nuevo conocimiento obtenido y su evaluación.

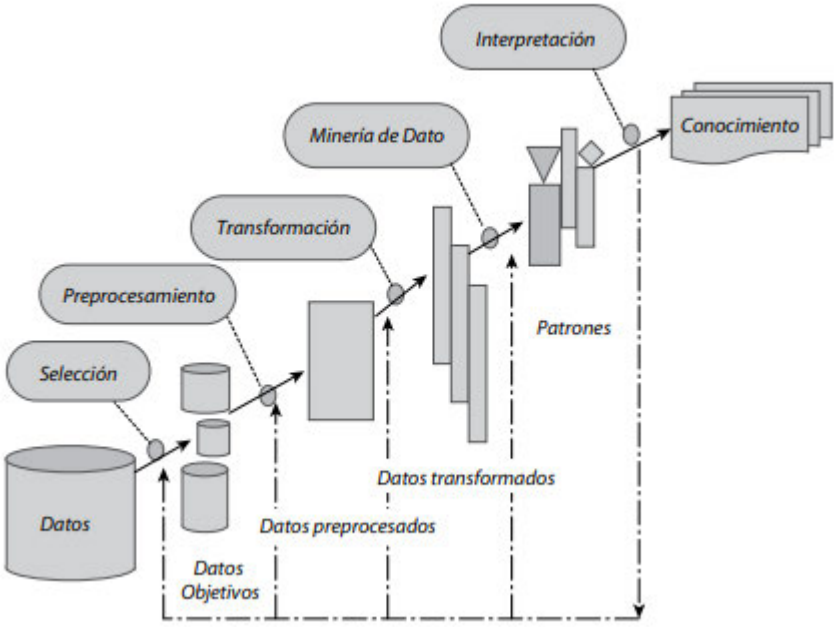

Figura 2.2 - Modelo KDD. Fuente [19]

Cabe destacar que la metodología KKD especifica las fases del proyecto de explotación de datos, pero no las tareas a realizar en cada una de ellas [21].

2.2.2 Metodología SEMMA (Sample, Explore, Modify, Model, Assess)

En la metodología SEMMA se establecen cinco fases para realizar el proyecto de explotación de datos, entre los que se encuentran el muestreo, la exploración, la modificación, el modelado y la evaluación. Esta metodología se orienta más al desarrollo del proyecto y deja fuera el estudio del problema o la planificación de este.

Como se puede apreciar en la Figura 2.3, las etapas que forman la metodología SEMMA son las siguientes [22]:

1. **Muestreo** (Sample): Consiste en obtener una muestra del conjunto de datos que tiene la organización a su disposición, siempre y cuando sea lo suficientemente grande para contener información de calidad y lo suficientemente pequeña para que se realice rápidamente. Esta etapa es opcional, se aconseja su realización cuando la cantidad de datos es demasiado extensa.
2. **Exploración** (Explore): En esta etapa se realizan datos en temas no relacionados con el principal, para poder encontrar nuevas hipótesis y realizar un mejor análisis.
3. **Modificación** (Modify): Se preparan los datos, eliminando valores nulos o extraños. También se modifican las variables con las que se va a trabajar.
4. **Modelado** (Model): Trata en crear un modelo para poder predecir la respuesta que se va a obtener, utilizando técnicas predictivas como redes neurales o árboles de decisión.

5. **Evaluación** (Assess): Se evalúa la exactitud y la utilización de los modelos obtenidos con el proceso de explotación de datos.

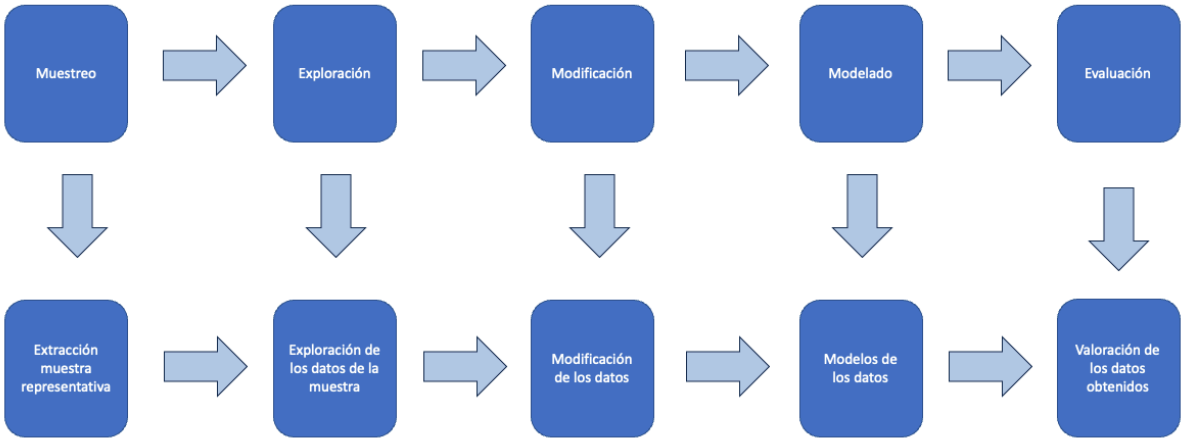

Figura 2.3 - Modelo SEMMA. Elaboración propia.

En esta metodología, al acabar las cinco fases, se vuelven a generar hipótesis para volver repetir el proceso. Al igual que KKD, SEMMA no proporciona una guía de tareas a seguir en cada una de sus fases [16].

2.2.3 Metodología Catalyst

La metodología **Catalyst**, también conocida como **P3QT** (*Product, Place, Price, Time, Quantity*), es un proceso que recomienda que la minería de datos siempre se desarrolle para ayudar a la organización en su situación actual. También se aconseja que no se trabaje directamente con los datos, sino que previamente se establezca la problemática a abordar, las expectativas que se tienen, así como las necesidades de los usuarios y el personal involucrado en el proyecto. Este primer punto es fundamental para que la organización justifique el desarrollo del proyecto [23].

En cuanto a la **estructura** de esta metodología, está formada por dos categorías o sub-metodologías, una metodología para el **Modelado de Explotación de Información** y otra para el **Modelado de Negocio** [23]. En la siguiente Figura 2.4 se pueden apreciar los dos modelados [24].

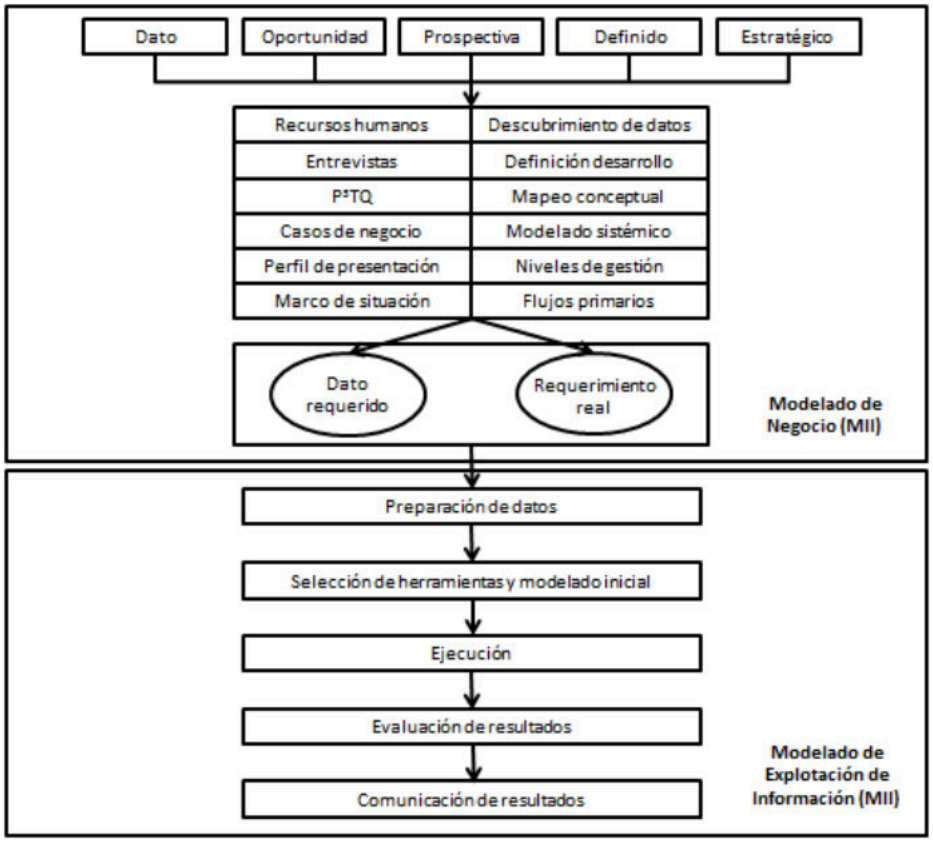

Figura 2.4 - Modelo Catalyst. Fuente [24]

Como se puede apreciar en la Figura 2.4 en la **metodología para el Modelado de Negocio**, se proponen cinco escenarios para el proyecto. Son los siguientes [23] [25]:

- Primer escenario: **Datos**.

    1. Buscar datos con relaciones útiles.
    2. Determinar las fuentes de información.
    3. Identificar el personal relevante en el proyecto.
    4. Discutir el proyecto con el personal.
    5. Parametrizar los datos en función a las relaciones P3QT.
    6. Descubrir la motivación que tiene negocio para la recoleta y almacenamiento de los datos.
    7. Descubrir quién o qué áreas inicio el proyecto.
    8. Descubrir el problema, identificando las relaciones P3QT, relacionando lo que representan los datos con los objetos que posee la organización, enmarcando el problema, etc.

- Segundo escenario: **Oportunidad/Problema**.

    Ver como la extracción de datos y su procesado pueden resolver un problema o crear una oportunidad.

    1. Identificar al personal interesado.

2. Estudiar el caso con el personal interesado.
3. Enmarcar la situación de negocio.
4. Obtener los objetivos de negocio.
5. Identificar qué datos se tienen que explotar.
6. Desarrollar el caso de negocio.
7. Presentar la situación de negocio a las personas de interés.
8. Describir el caso de negocio que se ha obtenido con la extracción de datos.
9. Definir los requisitos y parámetros para la implementación.

- Tercer escenario: **Prospección**.

  Consiste en averiguar dónde el desarrollo de un proyecto de minería de datos puede aportar valor para la empresa. Está formado por los siguientes pasos:

  1. Obtener las relaciones P3QT relevantes para la organización.
  2. Obtener los principales procesos de la organización,
  3. Identificar a las personas de interés.
  4. Entrevistar a las personas de interés.
  5. Descubrir qué cambios, para el usuario, pueden resultar de mayor interés.
  6. Buscar las fuentes de información.
  7. Preparar un borrador con cada caso de negocio, con cada oportunidad que se puede conseguir.
  8. Presentar los casos de negocio a las personas de interés.
  9. Definir qué caso se va a llevar a cabo.
  10. Definir los requisitos y parámetros para la implementación.

- Cuarto escenario: **Modelo definido**.

  Consiste en desarrollar un proyecto de explotación de datos para un caso en concreto. Se deben tener en cuenta los siguientes pasos:

  1. Identificar a las personas de interés.
  2. Definir los requisitos para la implementación junto con las personas de interés.
  3. Definir el caso de negocio.
  4. Buscar las fuentes de información.
  5. Definir los requisitos y parámetros para la implementación.

- Quinto escenario: **Estrategia**.

Utilizar la minería de datos para llevar a cabo una planificación estratégica y que ayude en la toma de decisiones. Teniendo en cuenta los siguientes puntos:

1. Obtener a las personas de interés.
2. Entrevistar a dicho personal.
3. Definir el caso de negocio.
4. Crear un mapa estratégico con el personal de interés.
5. Crear un modelo a partir del mapa.
6. Identificar las relaciones P3QT más relevantes de la organización.
7. Relacionar el mapa con las relaciones P3QT.
8. Simular la situación estratégica para identificar errores y posibles mejoras.
9. Identificar que relaciones P3QT son de mayor relevancia.
10. Identificar las más viables.
11. Buscar las fuentes de información.
12. Definir el caso de negocio.

En cuanto a la metodología para el **Modelado de Explotación de Información**, se define el orden de tareas que hay que realizar para descubrir patrones en los datos, en relación con la problemática de negocio identificada. Las tareas son las son las siguientes [23] [25]:

- **Preparar los datos:**
    1. Evaluación de las variables de estudio.
    2. Tener en cuanta posibles errores en las variables.
    3. Tener en cuenta posibles errores en las bases de datos.
    4. Evaluar posibles variables anacrónicas.
    5. Obtener una cantidad de datos suficiente.
    6. Cubrir todos los valores posibles de las variables.
    7. Posible necesidad de recodificación de variables.
- **Seleccionar modelado inicial y herramientas:**
    1. Categorizar los datos para el proceso de minería.
    2. Definir variables de entrada y salida.
    3. Seleccionar algoritmo para el procesado de datos.
    4. Evaluación los problemas que puedan originar los datos faltantes.
    5. Crear un modelado inicial. Este puede ser de tipo descriptivo, de predicción o de clasificación.
- **Refinar el modelo:**

1. Si es de tipo descriptivo, se tienen que describir los resultados obtenidos.
2. Si es predictivo o de clasificación, verificar su capacidad clasificatoria o predictiva.
3. Verificación del modelado con las personas de interés.

- **Implementar el modelo:**
    1. Si es descriptivo, de tienen que revisar los requerimientos del problema, describir los resultados en un informe, tener en cuenta los valores extremos, añadir evidencias experimental y negativa y conseguir opiniones de los usuarios.
    2. Si es de predicción o de clasificación, se tienen que revisar los requerimientos del problema, realizar una explicación del modelo y revisar los requerimientos de implementación.
    3. Comunicación de resultados finales.

Como se puede apreciar, Catalyst ofrece una manera de trabajar detallada que especifica todas las actividades que hay que desarrollar en cada una de sus tareas o escenarios.

2.2.4 Metodología CRISP-DM (Cross-industry Standard Process for Data Mining)

CRISP-DM fue presentada por las empresas Daimer Chrysler, NCR y SPSS en el año 1999, es una metodología que no está ligada a ningún producto [22]. Se trata de una metodología estructurada en niveles jerárquicos, formados por cuatro niveles distintos, que van desde un ámbito más general a lo específico [18].

En el nivel más alto, CRISP-DM propone seis fases para el proceso de explotación de datos, entre los que se encuentran el entendimiento del negocio y los datos, la preparación de dichos datos, el modelado, la evaluación y su posterior implementación. Según propone esta metodología, la sucesión de estas fases no tiene por qué ser en orden, como se puede observar en la Figura 2.5.

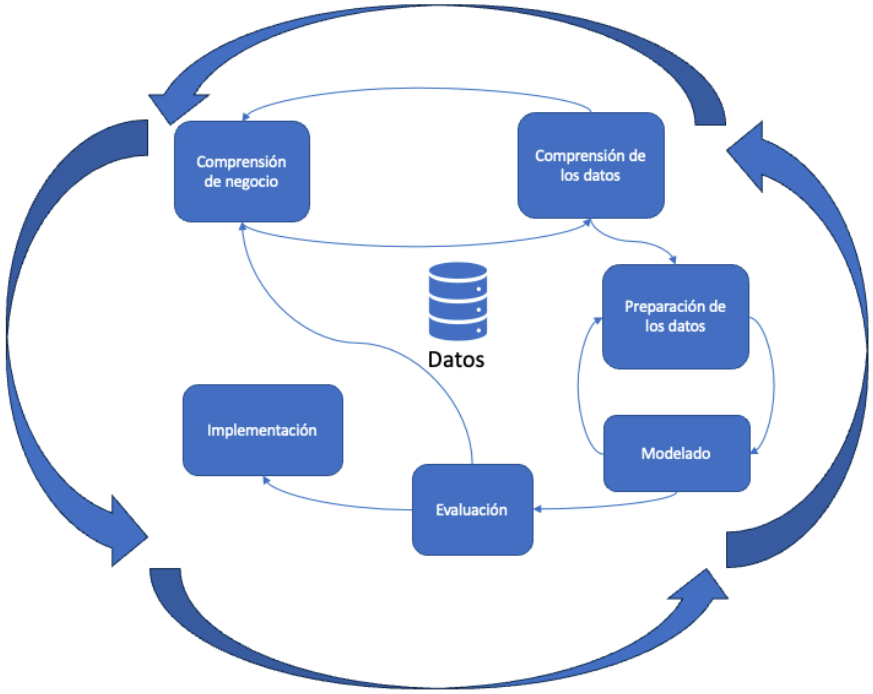

Figura 2.5 - Modelo CRISP-DM. Elaboración propia.

Cada una de las fases recientemente mencionadas, se descompone en una serie de tareas generales de segundo nivel. Se tratan de tareas de carácter general, ya que intentan ser de utilidad para la mayoría de las situaciones posibles que se pueden dar en la minoría de datos. Ya en el tercer nivel, se realiza el paso de estas tareas genéricas a casos más específicos, dando lugar a que en el cuarto nivel se describan la serie de acciones, decisiones y resultados obtenidos del proyecto de explotación de datos en particular. En Figura 2.6 se pueden apreciar las tareas generales de segundo nivel, asociadas a las del primer nivel.

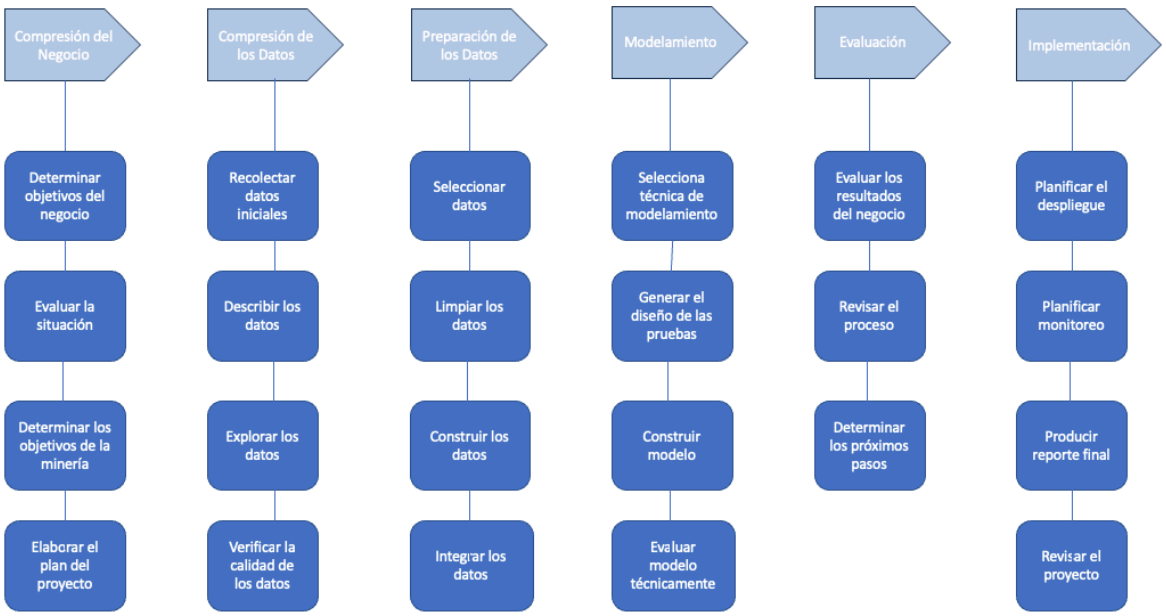

Figura 2.6 - Pasos del modelo CRISP-DM. Elaboración propia.

A continuación, se presentan con más detalle las tareas de segundo nivel [18]:

- **Comprensión del negocio**. Es la fase donde se establecen los objeticos y requerimientos del proyecto de minería de datos desde la vista de negocio, definiendo el plan de trabajo a seguir. Entre sus tareas se encuentras:
    1. Definir los objetivos del negocio: Aquí se obtienen la situación en la que se encuentra la organización y la descripción del problema, los objetivos del cliente y él éxito que se quiere alcanzar con el proyecto.
    2. Evaluar la situación: Se evalúa la situación actual de negocio y se obtienen el inventario de recursos disponibles, la lista de requerimientos del proyecto, se detectan los posibles riesgos junto con sus planes de contingencia y un análisis de costo-beneficio que supone llevar a cabo el proyecto.
    3. Determinar los objetivos de la explotación de datos: Se trata de los objetivos técnicos del proyecto. Con esta tarea se obtienen los objetivos describiendo los resultados previstos a obtener, así como definir las condiciones que hagan que se acepten los resultados conseguidos.
    4. Desarrollar plan para el proyecto: Se define una planificación para el proyecto, de aquí de obtienen el plan de proyecto, formado por las diversas tareas que hay que realizar, sus duraciones, recursos necesarios, etc.
- **Compresión de los datos**. Es la fase donde se obtienen los datos que se van a utilizar en el proyecto, en este punto pueden aparecer planes sobre qué hacer con información que se encuentre oculta. Sus tareas son las siguientes:
    1. Recolecta de datos: Se obtiene una recolección de datos junto con un informe que describe la manera en la que han sido obtenidos y los problemas que han aparecido mientras se hacía.
    2. Descripción de datos: Se describen los datos, junto con su tamaño y formato.
    3. Exploración de los datos: Observar las variables con mayor relevancia y su distribución. Al realizar esta tarea, se obtiene un reporte inicial con los resultados del análisis y el posible impacto en el proyecto.
    4. Verificación de datos: Se examinan los datos determinando su calidad. Con esto se consigue un reporte con el análisis de calidad realizado junto con soluciones a errores o problemas encontrados.
- **Preparación de los datos.** Es la fase en la que se obtiene el conjunto de datos final sobre el cual se aplicarán las técnicas de explotación de datos. Está formada por las siguientes tareas:
    1. Selección de datos: Se seleccionan los datos que se van a utilizar para el análisis, para ello hay que justificar porque la inclusión de unos y la exclusión de otros.
    2. Limpieza de datos: En esta etapa se mejora la calidad de los datos, como cambio de nombres de variables o borrado de datos ausentes. Al realizar esta

actividad, se tiene que elaborar un informe que detalle qué se ha hecho con los datos y porqué.

3. Construcción de los datos: Se realiza la construcción de nuevos datos que pueden servir de interés para el proyecto gracias a los primeros datos obtenidos, como pueden ser atributos derivados, nuevos registros creados.
4. Integración de los datos: Consiste en la creación de nuevas tablas o registros que contengan los datos obtenidos.
5. Cambio de formato de datos: Trata del cambio de formato en los datos por necesidades técnicas.

- **Modelado**. Es la fase en la que se aplican modelos a los datos para conseguir información oculta y patrones en ellos. Está formada por las siguientes fases:
    1. Selección de las técnicas de modelado: Se selecciona la técnica de explotación de datos a utilizar. Se debe documentar qué técnica se va a emplear.
    2. Diseño de pruebas de modelado: Consiste en diversas pruebas para verificar la validez y calidad del modelado.
    3. Construcción del modelo: Trata de aplicar la técnica selecciona anteriormente en el conjunto de datos preparados en la fase anterior.
    4. Evaluación del modelado: Las personas de interés del proyecto evalúan la calidad del modelo a partir de los criterios de éxito establecidos en fases anteriores.
- **Evaluación**. Se analizan los patrones obtenidos con relación a los objetivos de negocio. Está formado por las siguientes tareas:
    1. Evaluación de resultados: Se evalúan los resultados en función de los objetivos de negocio, teniendo en cuenta su validez.
    2. Revisión del proceso: Se buscan errores o problemas que hayan surgido en todo el proceso.
    3. Definir próximas etapas de actuación: Teniendo en cuenta los dos pasos anteriores, se definen los siguientes pasos a seguir, si se pasa a la implementación o hay que volver a realizar otra fase.
- **Implementación**. Se trata de la comunicación de la nueva información obtenida su implementación. Consta de las siguientes tareas:
    1. Planificación de la implementación: Se genera un plan para llevar a cabo la implementación del proyecto.
    2. Planificación de monitoreo y de mantenimiento: De suma importancia por si los datos cambian, provocando que los resultados puedan también cambiar.
    3. Elaboración de un reporte final: Consiste en un resumen del proyecto elaborado, en el que se incluyen los resultados obtenidos.

4. Revisión del proyecto: Se identifican los puntos que fueron bien realizados y los que mal, para poder implementar mejoras para proyectos futuros.

Como se ha podido observar, se trata de un grupo de fases de carácter general, en ellas se especifican los reportes o entregables que se tienen que generar [22].

2.2.5 Norma UNE 166006:2018

Los sistemas de vigilancia e inteligencia siempre han sido relevantes para las organizaciones y han tenido una constante evolución. Prueba de ello, es que en el año 2006 se publicó la primera edición de la norma UNE 166006, que facilita el uso de la vigilancia tecnológica como medio para mejorar el desempeño en investigación y desarrollo de las empresas y otras organizaciones. En 2011 se revisó por primera vez, versión en la que se incluyó el concepto de inteligencia competitiva, la búsqueda y procesamiento de información y su uso para el análisis y la toma de decisiones. Por último, en 2018, se ha publicado la siguiente edición de la norma, que actualiza la práctica propuesta y cubre los nuevos requisitos de los usuarios en un área que siempre requiere un rápido desarrollo [26].

Esta última versión de la norma UNE 166006:2018, da a conocer con carácter general los procesos de vigilancia e inteligencia. Dichos procesos están formados por diferentes etapas necesarias para llevar a cabo en cualquier organización, así como el flujo que sigue la información entre dichas etapas y los resultados obtenidos de las mismas [17]. En la Figura 2.7 se pueden observar los diferentes procesos.

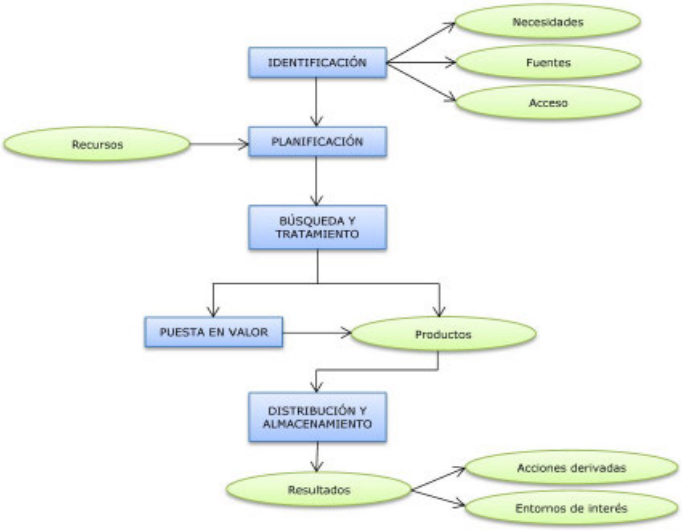

Figura 2.7 - Fases de un sistema de vigilancia e inteligencia. Fuente: [11]

A continuación, se van a explicar los diferentes procesos, y sus respectivas etapas, que son los que forman un sistema de vigilancia e inteligencia según la norma 166.006:2018 [17]:

- **Identificación**: En esta primera etapa se identifican las necesidades de la información teniendo en cuenta:

- Áreas identificadas de vigilancia e inteligencia.
- Un primer progreso sobre las posibles fuentes de información disponibles en dichas áreas.
- Un primer progreso sobre palabras clave y criterios de selección para elaborar el informe final.
- Información sobre los contenidos del producto que se elaborará y su tipo.
- También en esta etapa, usando como base las necesidades obtenidas del proceso anterior, se realiza una primera identificación de las fuentes de información y los recursos que tiene a su alcance la organización o están disponibles, como, por ejemplo:
  - Documentación perteneciente o relacionada a la organización.
  - Personas que tienen experiencia o conocimiento con dichas necesidades.
  - Personas externas de interés.
  - Otras organizaciones externas.
  - Fuentes documentales disponibles a la organización, tanto en soporte físico (catálogos y revistas, etc.) como en soporte electrónico (bases de datos, etc.), portales de información provenientes de Internet (redes sociales, noticias, etc.) o archivos multimedia (imágenes, audio, etc.).
  - Documentos técnicos como propiedad industrial e intelectual, normas, reglamentaciones o especificaciones.
  - Exposiciones, seminarios, ferias o congresos.
  - Resultados ya existentes de otros análisis como tendencias de futuro o ejercicios de previsión propios.

- **Planificación**: Para llevar a cabo la planificación, se deben tener en cuenta dos criterios, el seguimiento constante de las áreas identificadas previamente para estar al tanto de novedades y la búsqueda e identificación de nuevas áreas que puedan ser de utilidad para la organización. En base a las necesidades de las nuevas áreas identificadas, el acceso y fuentes de éstas, hay que planificar los recursos disponibles en función de experiencias pasadas. Debido a la continua evolución en un sistema de vigilancia e inteligencia, la organización debe cerciorarse de que la estructura, periodicidad y actualización en el seguimiento de las actualizaciones en las áreas ya identificadas estén establecidas.
- **Búsqueda y tratamiento de la información**: En esta etapa se distinguen dos procesos. En primer lugar, se tiene la búsqueda, la cual se debe realizar a través determinadas acciones y una estrategia establecida en las fuentes de información previamente seleccionadas. Para las siguientes fases de tratamiento y puesta en valor, las estrategias establecidas en la búsqueda pueden resultar de gran utilidad. Debido a

esto, puede ser de ayuda incluir la estrategia seguida, como las palabras clave, operadores usados, descriptores o terminología, entre otros. Después de recopilar los datos, se debe distinguir cuáles son útiles para satisfacer las necesidades de la información.

En segundo lugar, se encuentra el tratamiento de la información, el cual está estrechamente ligado a la calidad de las fuentes a través de las cuales se ha obtenido la información. Este proceso consta de una limpieza y normalización de los datos que se denomina preparación. Dependiendo del volumen de datos, esta preparación puede ser manual o mediante un tratamiento inicial. Este segundo tipo se puede diferenciar entre:

- Tratamiento inicial de la información estructurada: Su objetivo es encontrar correlaciones o estructuras que no se pueden obtener fácilmente, a través de diferentes técnicas de agrupación y reducción dimensional, y encontrar posibles maneras de análisis. Entre las posibles técnicas aplicadas, se incluyen las siguientes:
  - Diagramas de dispersión.
  - Análisis de probabilidad condicional.
  - Análisis geoposicional.
  - Distribución de variables.
  - Análisis de correlación.
  - Análisis multivariante.
- Tratamiento inicial de la información no estructurada: Este análisis se basa en técnicas como NLP (procesado de lenguaje natural) para extraer características que sirvan para procesar su contenido y en técnicas de análisis de imagen.
- Para finalizar, se aplica un análisis matemático o estadístico tras el tratamiento inicial, usando técnicas relacionadas con el problema a tratar. Durante este proceso es normal encontrarse con problemas de clasificación, optimización, asociación o predicción. En las técnicas que se pueden utilizar para resolver estos problemas, se encuentran:
  - Modelos de regresión.
  - Interferencia estadística.
  - Machine learning.

Los resultados que se obtienen de esta etapa se consideran suficientes para la toma de determinadas decisiones. Si no es así, se debe poner en valor toda esta información obtenida y realizar un análisis más detallado.

- **Puesta en valor**: La puesta en valor, es un proceso llevado a cabo por expertos que tengan los suficientes conocimientos técnicos, capacidad técnica e ingenio para

relacionar toda la información obtenida en el proceso anterior con oportunidades, adecuación de estrategias, reducción u otros aspectos. Éstos pueden ser:

  - Combinación de datos procedentes de numerosas fuentes con el objetivo de obtener una mayor relevancia y alcance que cada uno de los datos por separado.
  - Interpretar la información con el fin de determinar qué es lo verídico y qué es lo importante para la toma de decisiones.
  - Facilitar la compresión gracias a infografías o representaciones gráficas.
  - Dar significado a los datos analizados junto con sus implicaciones y consecuencias.
  - Recomendaciones en la manera de actuar de la organización.

- **Distribución y almacenamiento**: Esta etapa se basa en la distribución y el almacenamiento de los productos obtenidos en la vigilancia e inteligencia a las partes interesadas según las necesidades de la organización y asegurando que se sigue el procedimiento que la organización desea. Toda la información debe ser recuperable y se debe poder tener acceso a ella para posibles actualizaciones posteriores.

2.2.6 Comparativa entre las distintas metodologías

De los cinco modelos que se han presentado, SEMMA es la única que comienza el proceso minería de datos a partir del conjunto de datos, es decir, con la etapa inicial de muestreo de datos. En el lado opuesto, CRISP-DM, KDD (en su versión completa de nueve pasos) y la norma UNE:166006 comienzan con un análisis del negocio y del problema organizativo. Catalyst es la metodología más completa en este sentido, ya que incluye cinco posibles escenarios como puntos iniciales.

En lo que respecta a la estructura de las fases de la metodología, KDD, CRISP-DM, Catalyst y la norma UNE:166006 consideran el análisis y la comprensión del problema antes de iniciar el proceso de minería, mientras que SEMMA excluye este proceso. En todos los modelos se incluye la selección y preparación de los datos, así como en la fase de modelado, donde se aplican técnicas de minería para descubrir nuevos patrones.

La fase de evaluación de los patrones obtenidos está presente en todas las metodologías. En SEMMA, la evaluación e interpretación de estos patrones se realiza evaluando el rendimiento del modelo, mientras que en las otras metodologías la evaluación se basa en la utilidad que aportan para la solución del problema de la organización.

La implementación de los resultados obtenidos no está contemplada en el modelo SEMMA. En CRISP-DM, además, se propone una planificación para el control futuro y un análisis de cierre del proyecto. Este análisis tiene como objetivo recopilar información objetiva sobre el curso del proyecto para realizar una evaluación abierta del equipo de trabajo, las decisiones tomadas, las tecnologías utilizadas y sus consecuencias, con el propósito de aplicar lo aprendido en proyectos futuros.

En cuanto al nivel de detalle en las tareas de cada etapa varía entre las distintas metodologías. KDD y SEMMA presentan pasos generales del proyecto de explotación de datos, sin ofrecer una especificación detallada de las tareas a ejecutar en cada fase. Por otro lado, CRISP-DM, Catalyst y la norma UNE:166006 detallan más las actividades del proceso, con Catalyst yendo incluso más allá al indicar cómo realizarlas.

Por estas razones, KDD y SEMMA se asemejan más a un modelo de proceso que a una metodología. Mientras que CRISP-DM, la norma UNE:166006 y Catalyst son metodologías propiamente dichas, al llevar a cabo esa explicación de cómo realizar las fases, e incluyen actividades para la gestión del proyecto, como gestión del tiempo, costos y riesgos.

Por lo tanto, si se habla de metodologías para la gestión de un proyecto de explotación de datos, los modelos que deben considerarse son la norma UNE:166006, CRISP-DM y Catalyst, ya que presentan un nivel similar de detalle en cuanto a procesos y tareas a realizar en cada uno de ellos.

En la Tabla 2.1, se muestra un resumen de las características más importantes de cada una de las metodologías explicadas en este capítulo.

Tabla 2.1 - Comparativa de metodologías. Elaboración propia.

Aspecto	SEMMA	CRISP-DM	KDD	UNE:166006	Catalyst
Inicio del proceso	Conjunto de datos (muestreo de datos)	Análisis del negocio y del problema organizativo	Análisis del negocio y del problema organizativo	Análisis del negocio y del problema organizativo	Cinco posibles escenarios como puntos iniciales
Análisis del problema	No	Sí	Sí	Sí	Sí
Selección y preparación de datos	Sí	Sí	Sí	Sí	Sí
Fase de modelado	Sí	Sí	Sí	Sí	Sí
Evaluación de patrones	Evaluación del rendimiento del modelo	Utilidad para la solución del problema organizativo	Utilidad para la solución del problema organizativo	Utilidad para la solución del problema organizativo	Utilidad para la solución del problema organizativo
Implementación de resultados	No	Incluye planificación para control futuro y análisis de cierre del proyecto	Incluida (evaluación abierta del proyecto y lecciones aprendidas)	Incluida (evaluación abierta del proyecto y lecciones aprendidas)	Incluida (evaluación abierta del proyecto y lecciones aprendidas)

Nivel de detalle en tareas	Pasos generales sin especificación detallada	Detalle en las actividades del proceso	Pasos generales sin especificación detallada	Detalle en las actividades del proceso	Detalle en las actividades del proceso, indicando cómo realizarlas
Modelo vs Metodología	Más cercano a un modelo de proceso	Metodología completa	Más cercano a un modelo de proceso	Metodología completa	Metodología completa
Gestión del proyecto	No	Incluye gestión del tiempo, costos y riesgos	No	Incluye gestión del tiempo, costos y riesgos	Incluye gestión del tiempo, costos y riesgos

2.3 Tecnologías para la implementación de las metodologías

Una vez se han descrito las metodologías que se pueden utilizar para el desarrollo de un proyecto, se van a mencionar las distintas tecnologías que permiten desarrollar cada uno de los procesos o partes de dichas metodologías. Para la elección de una tecnología, hay que tener en cuenta que la que se escoja debe cumplir con las "**7V**" para ser lo más completa posible [27].

La primera característica es la velocidad. Se refiere al ritmo con el que se obtienen los datos y la rapidez que se procesan. Hay varios tipos de herramientas que proporcionan una visión a tiempo real.

La segunda característica es la variedad. Las tecnologías seleccionadas tienen que ser capaces de aceptar diferentes tipos de formato de los datos, es decir, incluir datos estructurados, semi-estructurados y no estructurados.

La veracidad es la tercera característica. Se trata de garantizar la calidad de los datos con tecnologías que sean capaces de descubrir los datos que son erróneos, que están duplicados o que no aportan valor.

La cuarta característica es el valor. La información obtenida siempre tiene que aportar utilidad a la organización. Lo que se traduce en que las tecnologías seleccionadas tienen que ser capaces de transformar los datos en información de valor.

En cuanto al volumen, se refiere a la capacidad de procesar grandes cantidades de datos sin comprometer a otras características.

La vulnerabilidad también debe tenerse en cuenta. Los datos pueden sufrir ataques cibernéticos y brechas de seguridad, por lo tanto, las tecnologías tienen que gestionarlos y ofrecer siempre una cantidad de datos constante y relevante.

Por último, la visualización es una característica que es crucial para poder interpretar los datos de manera correcta. Las tecnologías elegidas deben presentar los datos de manera clara y comprensible, favoreciendo la identificación de patrones o tendencias.

A continuación, se van a describir una parte de las tecnologías o herramientas existentes, teniendo en cuenta las características presentadas. Se van a agrupar en dependiendo de su funcionalidad.

Las **herramientas o tecnologías** más adecuadas **para la obtención de datos** son las siguientes [28] [29]:

- APIs: Es un conjunto de protocolos y definiciones que permiten que diferentes aplicaciones o sistemas se comuniquen entre sí. Generalmente, las organizaciones ponen a disposición de desarrolladores y empresas externas estos servicios de pago, para que puedan consultar la información que estas ofrecen.
- Web Scraping: Es una técnica por la cual un programa simula la navegación del usuario por una página web y va obteniendo la información html o JavaScript que ésta contiene. Hay numerosas herramientas que permiten emplear esta técnica, como ParseHub, OctoParse, Selenium o Playwright. Algunas de estas técnicas se tienen que implementar directamente en código, pero otras tantas presentan aplicaciones de escritorio, que permiten la obtención de datos mediante una interfaz de fácil uso para el usuario, como es el caso de WebHarvy.

También existen determinadas **herramientas** que se utilizan para la **limpieza de datos**, entre ellas destacan [30] [31]:

- Pandas: Es una herramienta muy popular para la manipulación de datos, que se puede implementar como librería dentro de Python. Se basa en estructuras de datos, denominados DataFrames, que facilitan la transformación y la limpieza de los datos.
- Numpy: Al igual que Pandas, es una librería disponible en Python. Funciona a través de funciones matemáticas que dan soporte a arrays multidimensionales y matrices de gran tamaño.
- Pentaho Data Integration: Es una herramienta de integración de datos y análisis muy conocida para procesos ETL.

Continuando con **herramientas o tecnologías** adecuadas **para el manejo de datos**, se encuentran las siguientes [32]:

- Apache Hadoop: Es una herramienta gratuita de código abierto que es capaz de procesar gran cantidad de datos usando modelos de programación simples. También es escalable, lo que hace que sea capaz de funcionar en varios servidores o en uno solo.

- Elasticsearch: Da la oportunidad de procesar grandes volúmenes de datos y ver su evolución en el mismo instante. También es capaz de proporcionar gráficos que sirven de ayuda para entender la información conseguida.

  La principal característica de esta herramienta es combinar diferentes tipos de contenidos, desde búsquedas en páginas web y aplicaciones hasta métricas de

infraestructuras, visualización de datos de tipo geoespacial, etc. Esta herramienta puede ampliarse con Elastic Stack, una combinación de productos que complementan Elastcsearch para aumentar sus capacidades.

- Apache Storm: Se trata de una herramienta de código abierto que ofrece la capacidad de procesar una gran cantidad de datos y analizarlos con la creación de topologías de metadatos.
- MongoDB: Es una base de datos de tipo no relacional y como todas ellas, brinda la capacidad de trabajar con un número grande datos.
- Apache Spark: Al igual que Apache Storm, es de código abierto y en lo que destaca, es en la utilización de varios ordenadores al mismo tiempo para procesar datos en paralelo Herramienta de código abierto que utiliza varias computadoras al mismo tiempo para procesar datos en paralelo. Esto hace que se puedan analizar datos en tiempo real. Además, permite el uso de diferentes lenguajes de programación como Python, R, Java y Scala.
- Python: De todas las herramientas mencionadas, es de las más populares debido a su sencillez y usabilidad. Por ello, cuenta con una biblioteca de librerías muy grande, lo que permite encontrar funciones que se pueden añadir a los desarrollos propios con mucha facilidad. Por otro lado, es un lenguaje bastante más lento que sus competidores.
- Apache Cassandra: Al igual que MongoDB, se trata de una base de datos no relacional.
- R: Se trata de un lenguaje de programación cuyo principal uso es el de análisis de estadísticas. Al igual que Python, tiene una gran cantidad de librerías.
- Apache Drill: Es capaz de realizar análisis interactivos, a gran escala, de lotes de datos. Diseñado para trabajar en servidores y para procesar millones de registros y petabytes de información en segundos. Una de sus principales ventajas es que permite la utilización de archivos NoSQL y SQL, generando la unión de diversas bases de datos para realizar la consulta.
- RapidMiner: Ayuda en la aplicación de modelos, en el empleo de datos y en el desarrollo de modelos que aprenden de manera automática.

Una vez procesados los datos con unas de las tecnologías o herramientas presentadas anteriormente, es fundamental realizar una **visualización** de esos datos. Esta tarea consiste en la creación de gráficas y representación visual de los patrones y tendencias encontradas en los datos, hay diferentes tipos de herramientas ayudan a realizar esta tarea. La representación de los datos es muy importante para las organizaciones debido a las siguientes razones [33]:

- Difusión de los resultados: Ayudan a la comunicación de los resultados obtenidos mediante graficas e informes que favorecen la comprensión.
- Comparación rápida de grandes volúmenes de datos.
- Obtención de patrones ocultos en los datos: Debido al gran volumen de datos, muchas ocasiones es muy difícil ver las tendencias que se encuentran en ellos.

Entre todas las **herramientas** que se encuentran en el mercado para la **visualización de datos**, todas comparten características comunes, como pueden ser [33]:

- Plantillas de gráficos.
- API para la importación de datos.
- Gráficos interactivos.
- Optimización para uso móvil.
- Historial de versiones.

A continuación, se presentan las herramientas de visualización más utilizadas [33]:

- Microsoft Excel: Permite la creación de gráficos y tablas entre otros, sus características principales son la utilización de fórmulas, un autofiltro para ordenar los datos y Power Pivot, un complemento que sirve para la creación de gráficos, tablas y más visualizaciones.
- Microsoft Power BI: Es capaz de crear visualizaciones en informes de manera individual. Cuenta con una inteligencia artificial (Microsoft AI) que prepara y analiza los datos de forma automática, también con cuadros de mando personalizables y visualizaciones en tiempo real.
- Google Charts: En la herramienta para la creación de visualizaciones, a partir de conjuntos de datos, de Google.
- Tableau: Es de las herramientas más utilizadas, capaz de realizar análisis sobre datos procedentes de distintas fuentes. Además, puede responder preguntas básicas sobre los datos que está analizando.
- Zoho Analytics: Se trata de una herramienta enfocada para el uso de varios usuarios a la vez. También es capaz de generar informes de manera automática.
- Datawrapper: Otra herramienta de visualización de datos, cuyo principal reclamo es tener una característica para ayudar a las personas con daltonismo a realizar sus informes.
- Qlik Sense: Es una herramienta de visualización que cuenta con una inteligencia artificial integrada para la generación de informes automáticos.
- FusionCharts: Se trata de una herramienta que funciona con diferentes frameworks de JavaScript y tiene la posibilidad de mostrar paneles de gráficos en tiempo real.
- Domo: Es una herramienta de generación de gráficos que destaca en el análisis de datos en vivo y un chat para comunicación entre usuarios.
- Google Analytics: Otra herramienta de Google que integra diferentes herramientas de la organización o de otras, como de WordPress, para realizar informes interactivos y en tiempo real.
- Visme: Es una herramienta de generación de gráficos que tiene una biblioteca de widgets, animaciones para mejorar la presentación de los informes. Además, tiene la posibilidad de integración con Microsoft Office.

Por último, cualquier sistema de vigilancia e inteligencia tiene que contar con un sistema para el **almacenamiento de los datos**, en el mercado se encuentran diversas **herramientas o tecnologías** [34] [35] [36]:

- Bases de datos relaciones: Son bases de datos que almacenan datos estructurados en columnas y filas. Cada fila representa un registro único y cada columna, un valor de ese registro. Las tablas se relacionan entre sí con los identificadores únicos de cada registro, denominados como claves. Las operaciones de consulta mediante SQL y entre las más utilizadas se encuentran MySQL, PostgreSQL u Oracle Database.
- Bases de datos NoSQL: Son sistemas de almacenamiento que ofrecen mecanismos para recuperar y almacenar datos modelados de distintas tablas de bases de datos relacionales. Perfectas para datos no estructurados, semi estructurados o distribuidos. De este tipo de bases de datos destacan MongoDB y Redis.
- Bases de datos OLAP: Diseñadas para análisis multidimensional y analísticas complejas. Son bases de datos utilizadas para el análisis de datos históricos e inteligencia empresarial. En este tipo de bases de datos destacan Oracle OLAP e IBM Cognos.

Como se ha podido observar por el gran número de herramientas o tecnologías disponibles, en base a su funcionalidad, la selección de éstas puede ser una tarea complicada, por lo tanto, la organización se puede apoyar en la opinión de profesionales para elegir las correctas.

El cuadrante Mágico de Gartner es un informe creado por Gartner Inc., que proporciona un análisis del posicionamiento de los distintos proveedores de tecnologías en un mercado en concreto, es muy utilizado por otras empresas y profesionales para comparar y evaluar diferentes soluciones tecnológicas. El Cuadrante Mágico se divide en los siguientes cuatro cuadrantes [37]:

- Líderes (Leaders): Empresas que tienen una visión completa del mercado, innovadores y reconocidos por su éxito en el sector.
- Aspirantes (Challengers): Proveedores de tecnologías que dominan el mercado en términos de recursos, pero no tan innovadoras como los líderes.
- Jugadores de nicho (Niche players): Organizaciones que se centran en un segmento específico, por lo tanto, no tienen una visión amplia del mercado.
- Visionarios (Visionaries): Compañías que saben hacia dónde va a ir el mercado, pero que no han demostrado su estabilidad o éxito, tal como han hecho los líderes.

En la Figura 2.8, se puede observar el cuadrante mágico de Gartner donde se muestran las plataformas de analítica e inteligencia empresarial más utilizadas, entre ellas se encuentran Microsoft y Salesforce como líderes del sector, mencionadas anteriormente [38].

Figura 2.8 - Cuadrante Mágico de Gartner 2023 para plataformas analíticas e Inteligencia Empresarias. Fuente [35].

2.4 Fuentes de información

En un proyecto de explotación de datos es fundamental determinar la fiabilidad de las fuentes de datos que se utilizan. La calidad de los datos va a influir directamente en la calidad y validez de los resultados y, a consecuencia, en la toma de decisiones. La información no verificada puede llevar a conclusiones equivocadas, mientras que las fuentes fiables todo lo contrario.

En la era actual, la información se obtiene y se comparte de manera instantánea a través de las redes sociales. Por ello, los mercados financieros están influenciados por estas. Estos medios de comunicación han modificado la forma en la que se consume y se comparte la información, ofreciendo nuevas oportunidades de comercio e inversión [39]. Como se ha explicado, la procedencia de la información es de suma importancia y teniendo en cuenta esto, se puede obtener de dos grupos: **fuentes de información fiables** o **fuentes de información no fiables**.

Las **fuentes de información fiables** se definen como medios verificados. Por ejemplo, pueden ser instituciones gubernamentales y organismos reguladores, instituciones académicas y centros de investigación, medios especializados y asociaciones profesionales y analistas financieros [40] [41].

Al igual que en resto de ámbitos, en el mundo financiero, es muy importante tener información fiable para poder tomar las decisiones acertadas. La gran cantidad de datos que hay disposición de cualquier usuario puede llegar a ser abrumadora, por lo saber qué fuentes son fiables imprescindible.

A continuación, se presentan algunas fuentes fiables en el ámbito financiero:

- Instituciones gubernamentales y organismos reguladores [42]:
    - Bancos centrales: Como la Reserva Federal de Estados Unidos o el Banco Central Europeo, que publican estadísticas, informes y realizan investigaciones relacionadas con temas financieros.
    - Comisión de valores: Como la Comisión Nacional del Mercado de Valores en España, que se encarga de regular los mercados financieros y publica información sobre empresas que cotizan e investigaciones de casos de fraude y mercado.
    - Organismos internacionales: Como el Banco Mundial que publica informes sobre la economía global.
- Instituciones académicas y centros de investigación: Como el London School of Economics o el Bruegel Institute [43] [44].
- Medios financieros especializados, como:
    - The Financial Times: Periódico británico cuyo principal objetivo es informar de la actualidad económica y empresarial [45].
    - The Wall Street Journal: Periódico estadounidense que se centra en las noticias de negocios, financieras y otros asuntos económicos [46].
    - Yahoo Finance: Sitio web gratuito que pone a disposición de sus usuarios una gran cantidad de herramientas e información financiera [47].
    - Bloomberg: Compañía líder en el mundo de las finanzas que oferta a los profesionales una gran variedad de servicios [48].
    - AlphaVantage.co: Compañía que ofrece herramientas e información relativa a las finanzas, que a diferencia de Bloomberg, está orientada a todo tipo de usuarios [49].
- Analistas financieros y asociaciones de profesionales: Como el EFPA Europa y el Instituto Español de Analistas Financieros [50].

Al contario de las fiables, **las fuentes de información no fiables** son medios que no especifican la procedencia de la información. Por ejemplo, se puede distinguir en esta categoría las páginas web sin autor o responsables definidos, correos electrónicos no solicitados que dan consejos de inversión, las redes sociales, los seminarios que ofrecen expectativas de riqueza no confiables y, en general, fuentes que no estén reguladas o que no tengan transparencia [51].

En el ámbito financiero es crucial saber qué fuentes evitar para protegerse de información engañosa, falsa o que pueda provocar una toma de decisiones perjudicial.

Principalmente, las fuentes de datos no fiables son las que no citan su información ni sus referentes, fuentes que no indican su formación o que son anónimas [52]. A continuación, se enumeran algunos tipos de estas fuentes de información.

- Redes sociales: Por lo general, las redes sociales están repletas de cuentas no verificadas que divulgan información no contrastada y que no muestran la formación financiera del autor [53]. Algunos ejemplos de redes sociales donde ocurre esto son:
    - Twitter: Red social que permite compartir mensajes de hasta 280 caracteres, principalmente para compartir opiniones, intereses o gustos.
    - Reddit: Red social que permite compartir texto, imágenes o videos, generalmente en comunidades, para expresar tu opinión o interés.
- Otros medios: Aquí se pueden encontrar medios (blogs, páginas de internet, etc) que divulgan información exagerada sobre datos financieros, promesas para ganar más dinero o contenido para tomar decisiones de manera inmediata sin un análisis previo [52].

2.5 Conclusiones

El análisis y desarrollo de un sistema de vigilancia e inteligencia, siguiendo las pautas establecidas por las diferentes metodologías presentadas, es esencial para que las organizaciones puedan gestionar y transformar una gran cantidad de datos en información de utilidad. En este capítulo se ha demostrado que la implementación de metodologías como KDD, SEMMA, Catalyst, CRISP-DM y la norma UNE:166006:2018, proporciona un marco definido y estructurado para la explotación de los datos, teniendo las peculiaridades de cada una de ellas sus.

En cuanto a las tecnologías o herramientas necesarias para llevar a cabo el desarrollo del sistema de vigilancia e inteligencia, se han presentado varias que cumplen con una o más fases de las metodologías.

Por último, se ha mencionado la importancia de contar con fuentes de información fiables para garantizar los resultados del sistema.

En conclusión, el empleo de metodologías y tecnologías en sistemas de vigilancia e inteligencia permiten a las organizaciones tomar decisiones más acertadas y mejorar su competitividad. De todas las metodologías presentadas, CRISP-DM se recomienda para este proyecto debido a su enfoque estructurado y su capacidad para integrarse con otras herramientas y procesos, por lo que asegura una implementación eficiente de un sistema de vigilancia e inteligencia.

Capítulo 3. Planteamiento del problema. Necesidad de información para la toma de decisión en las organizaciones

3.1 Introducción

El entorno empresarial actual está expuesto a una cantidad de datos abrumadora, creando tanto desafíos como oportunidades para las organizaciones. La capacidad de transformar estos datos en información de calidad es fundamental para la toma de decisiones estratégicas [54] [55]. Por lo tanto, este capítulo se va a centrar en la necesidad de esa información valiosa en las empresas y cómo esta necesidad fomenta el desarrollo e implementación de sistemas de vigilancia e inteligencia. Para comprender el contexto en el que se desarrollan estos sistemas, a continuación, se analiza la naturaleza de los datos disponibles, su origen y los desafíos que conllevan su gestión.

3.2 Contextualización

La necesidad de información de calidad es fundamental en la época actual. La capacidad de obtener analizar datos, para posteriormente utilizarlos, puede determinar el éxito o el fracaso de una organización. Un ejemplo de ello es un artículo publicado en la *Review of Finance*, donde se menciona que las empresas que añaden el análisis de datos en sus operaciones pueden mejorar procesos o identificar tendencias de mercado, lo que les permite mantense competitivos [56].

Unas fuentes muy importantes y donde las organizaciones pueden obtener una gran cantidad de datos, son las redes sociales [57]. Estas se pueden definir como aplicaciones o páginas web que operan en varios niveles, permitiendo siempre el intercambio, entre empresas o personas, de información [58]. Desde su aparición, las redes sociales han servido como medio de protestas, como plataforma para la relación entre la empresa y sus clientes, para que las personas escriban su opinión respecto a diversos temas o simplemente para comunicarse con otros [59].

Su aparición se produjo en la década de los 90, con la creación de Internet. Pero no fue hasta los años 2000, cuando comenzaron a surgir redes sociales cuyo principal servicio era la interacción entre usuarios [59].

Con el paso del tiempo, las redes sociales, como Twitter, Reddit y Facebook, han demostrado tener un fuerte impacto en la sociedad. Como ejemplo, en un artículo en *Information Systems Frontiers* se habla de cómo las redes sociales ofrecen un entorno de intercambio de información, pero que también facilitan la propagación de las noticias falsas [60]. Otro de ejemplo de ello, es un artículo publicado en *BMC Psychology*, que analiza los pros y contras de las redes sociales en la salud mental [61].

Otros estudios académicos también han demostrado el impacto de las redes sociales en los mercados financieros. Como ejemplo se encuentra un artículo que analiza cómo una comunidad de Reddit alteró todo el mercado financiero provocando la subida del precio de las acciones de GameStop [62].

Por todos estos ejemplos, es importante resaltar que las redes sociales son una herramienta esencial para obtención y el análisis de datos en las organizaciones. Un estudio de *Journal of*

Business Research menciona como las redes sociales permiten a las empresas obtener información en tiempo real sobre el comportamiento de los consumidores y sus preferencias, lo que ayuda en la toma de decisiones [63].

Sin embargo, la cantidad de datos generados en las redes sociales y su variedad, son un desafío para en términos de análisis y transformación en información de valor. Aquí es donde se presenta la necesidad de la implementación de sistemas de vigilancia e inteligencia. La utilización de estos sistemas puede ayudar a las organizaciones a analizar grandes volúmenes de datos de manera eficiente y ofreciendo, como resultado, información de calidad para mantener la ventaja competitiva y mejorar la toma de decisiones en las organizaciones [64].

En conclusión, la integración de los sistemas de vigilancia e inteligencia en las organizaciones es crucial para manejar el gran volumen de datos provenientes de las redes sociales. Como se ha mencionado, no solo ayudarán en el análisis de la información, sino que permitirá a las organizaciones a reaccionar rápidamente a cambios en el mercado o cambios en las preferencias de sus consumidores.

3.3 Implementación de un sistema de vigilancia a e inteligencia

La implementación de los sistemas de vigilancia e inteligencia en las organizaciones puede resultar complejo, ya que depende de la cómo está estructurada la organización. Sin embargo, y como se ha podido apreciar, la integración de estos sistemas en las empresas proporciona muchas ventajas.

Para lograr un sistema adecuado, las organizaciones pueden adoptar por la metodología CRIPS-DM, ya que ofrece un marco estructurado que ayuda a las organizaciones a gestionar el proceso de minería de datos. También, es una metodología adaptable a diversos datos y problemas. Además, es repetible, lo que permite una mejora continua y asegura que los sistemas de vigilancia e inteligencia se mantengan efectivos y actualizados. Por último, facilita la integración de técnicas avanzadas de análisis de datos, ofreciendo información de valor como resultado [65].

3.4 Planteamiento del problema

En este este punto se presentan los problemas a lo que se enfrentan las organizaciones a la hora de gestionar la información y aprovechar las ventajas que ofrecen los sistemas de vigilancia e inteligencia.

El primer problema es que muchas organizaciones carecen de los profesionales dedicados exclusivamente a la vigilancia, lo que resulta en pérdidas de tiempo en la búsqueda de la información y su acceso [66].

El segundo problema es la diferencia entre los datos que provienen de diversas fuentes, esto conlleva la necesidad de normalizarlos para aplicarles el mismo formato o características antes de normalizarlos [67].

El tercer problema es el que la gestión de la información requiere el uso de herramientas complejas y tener un personal cualificado que sepa utilizarlas. Esto conlleva incrementos de tiempos y costes en la interpretación de la información. Sistemas más intuitivos pueden mejorar toda esta problemática [67].

El cuarto problema es la limitación a la hora de la obtención de información, por ejemplo, la obtención de la información mediante APIs suele proporcionar una variedad de datos, siendo estos más específicos y mejorado el acceso a información de valor [67].

Ante estos problemas, se ha perseguido el desarrollo de un sistema de vigilancia e inteligencia que tenga varias características para darles solución. Como, por ejemplo, llevar a cabo una automatización que permita realizar una recopilación de datos de diversas fuentes, minimizando la intervención manual y garantizando la actualización constante de la información. También, estableciendo procedimientos que permitan la normalización de los datos, que provengan de fuentes diferentes, o diseñando un sistema escalable y flexible que se pueda adaptar a diferentes tamaños y tipos de organizaciones.

Por todo ello, la investigación llevada a cabo en este Proyecto de Fin de Grado busca un sistema de vigilancia e inteligencia flexible para cualquier organización, que cubra todos los procesos desde la recuperación de la información hasta su difusión. Este sistema se valida con un caso práctico analizando datos reales.

Capítulo 4. Implementación de un sistema de vigilancia e inteligencia según la metodología CRISP-DM

4.1 Introducción

En este capítulo se propone y diseña un sistema de vigilancia e inteligencia que empleará la metodología CRISP-DM como guía, que se podrá aplicar en cualquier organización de manera general. Se elaborará en base a todo lo recopilado en los apartados anteriores, apoyándose en las herramientas que existen para desarrollar dicho sistema y las ventajas que supone su implementación para las organizaciones.

El caso práctico que se expone, para la validación del sistema, contendrá todas las fases de diseño y desarrollo de la metodología y se justificará la selección de las herramientas para llevarlo a cabo. Se tratará de un proyecto basado en la obtención de la información a partir de técnicas de web scraping, mencionadas en el apartado 2.3, para su posterior análisis.

4.2 Descripción del modelo propuesto de sistema de vigilancia e inteligencia

El sistema de vigilancia e inteligencia, basado en la metodología CRISP-DM, va a cubrir cada uno de los procesos que se han nombrado en apartados anteriores, desde la búsqueda de las fuentes de información y la extracción de ésta, hasta el tratamiento de los datos para obtener la información que la organización necesita.

El sistema propuesto es ejecutable y utilizable desde cualquier sistema operativo, pues utiliza una máquina virtual de Windows, que se puede abrir desde numerosos programas que permitan ejecutar una máquina virtual. Por ejemplo, se encuentran los más comunes, como VMWare Player y VirtualBox. También el procesado y limpieza de los datos, para normalizarlos, se hace a través de librerías de Python, mencionadas en el apartado 2.3. Dicho código se ejecuta en la propia máquina virtual, a través de cualquier IDE para Python.

Una vez obtenidos los datos necesarios y su procesado, hay que realizar su análisis. Para ello, se va a utilizar la herramienta de tratamiento y visualización de datos Power BI.

Por último, para el almacenamiento se desarrollará un data warehouse con PostgreSQL, debido a se necesita realizar consultas complejas para llevar a cabo análisis de grandes volúmenes de datos y generar informes desde Power BI.

4.2.1 Actores

En el sistema que se presenta, actuarán diferentes actores. Se pueden distinguir los siguientes:

- Clientes: Personas cuyo principal interés es el resultado del proyecto. Pueden ser departamentos propios de la organización o clientes externos. Su tarea es definir los objetivos y proporcionar retroalimentación durante el proyecto.
- Gerentes: Personas encargadas de planificar y ejecutar el proyecto además de supervisarlo. Aseguran de que se siga el plan establecido al comienzo de este, dentro de los plazos y presupuesto fijados. Son los encargados de coordinar los diferentes equipos y actores que participan.

- Analistas de negocio: Se encargan de identificar las necesidades del cliente y traducirlos a objetivos específicos para resolverlos en el proyecto. Se comunican con los clientes para asegurarse de que los objetivos del proyecto son los adecuados para cubrir sus necesidades.
- Científicos de datos: Son los expertos técnicos que desarrollan los modelos y los aplican en la extracción de datos. Se encargan de limpiar, procesar, explorar y modelar. Además, interpretan los resultados y los traducen en información útil para el cliente.
- Ingenieros de datos: Se ocupan de la gestión y preparación de datos. Se aseguran de que estén disponibles, cumplan con los estándares de calidad y tengan un formato adecuado.
- Expertos en TI: Son los que se encargan de proporcionar la arquitectura necesaria para el desarrollo del proyecto y de la gestión del almacenamiento de los datos.
- Auditores: Se encargan de asegurar que el proyecto cumpla con las leyes y normativas vigentes en el momento de desarrollo.

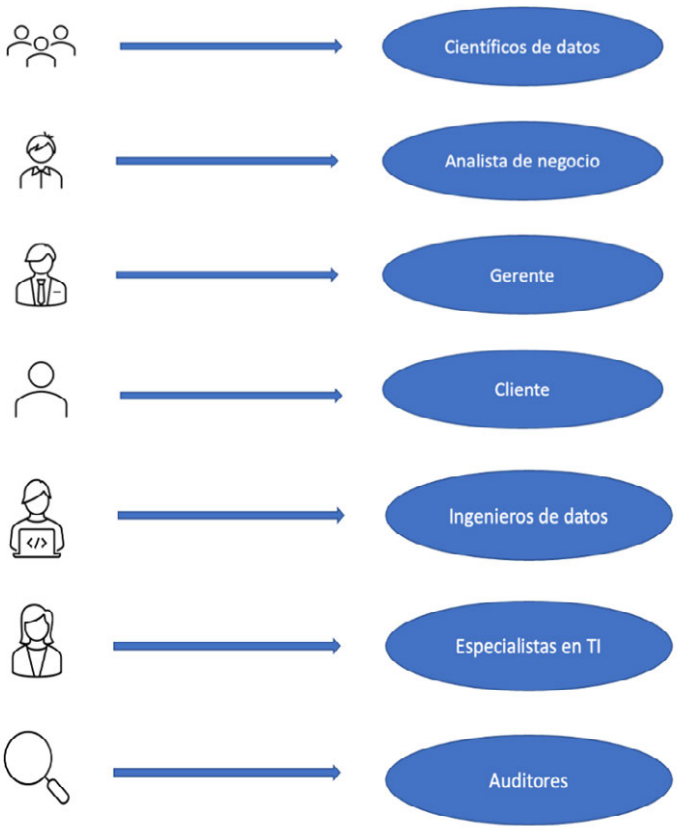

Figura 4.1 – Actores presentes en el estudio práctico. Elaboración propia.

Una vez definidos los actores que van a participar en la realización del proyecto, se presenta en la Figura 4.2 el modelo completo con las diferentes fases que forman la metodología y los actores que participan en cada una de ellas.

A continuación, se describe la figura anterior, mencionando cada una de las fases y dónde interviene cada uno de los actores mencionados:

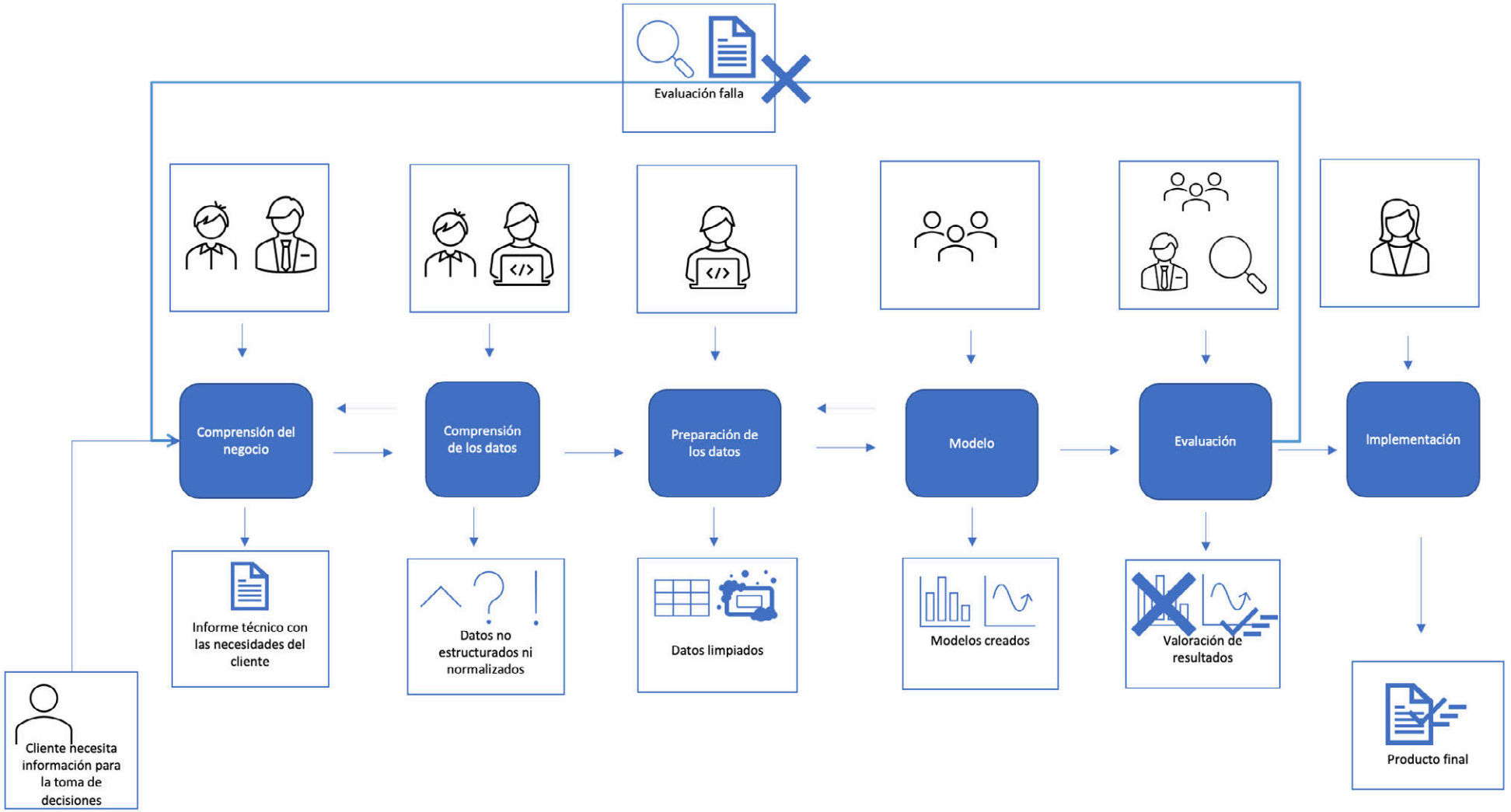

Figura 4.2 – Modelo con las diferentes fases de la metodología y actores que participan. Elaboración propia.

1. **Comprensión del negocio**: Donde los analistas de negocio y el gerente trabaja junto con el cliente para definir los objetivos del proyecto.
2. **Comprensión de los datos**: Donde ingenieros de datos y los analistas de negocio obtienen la información, la analizan y determinan su calidad y adecuación.
3. **Preparación de los datos**: Donde se limpian y transforman los datos para que sean aptos para el modelado.
4. **Modelado**: Donde los científicos de datos desarrollan y evalúan los modelos de KPIs.
5. **Evaluación**: Donde se revisan los modelos y se valida su eficacia y alineación con los objetivos del negocio. Si falla se vuelve a la comprensión del negocio, el fallo indica que no se han definido bien las pautas necesarias para obtener los objetivos pactados.
6. **Implementación**: Donde los resultados se implementan y se ponen en producción, asegurando que los resultados obtenidos sean utilizados de manera efectiva.

4.2.2 Arquitectura del modelo propuesto

La arquitectura propuesta se basa en prácticas comunes en proyectos de explotación de datos. Un ejemplo de ello es un estudio encontrado en MPDI, donde se destaca la importancia de las capas de recolección, procesamiento, almacenamiento y análisis [68]. A continuación, se representa el sistema en su conjunto, en la Figura 4.3:

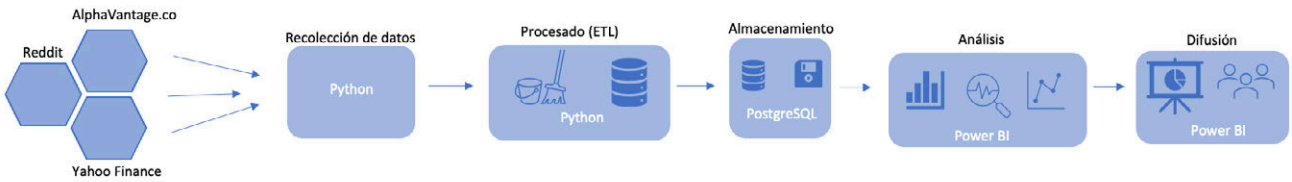

Figura 4.3 – Arquitectura del sistema. Elaboración propia

La arquitectura presentada satisface las necesidades que surgen al manejar grandes volúmenes de datos provenientes de varias fuentes, asegurando su calidad e integridad, transformándolos en información de valor para las organizaciones y haciendo que sea accesible para la toma de decisiones. Se ha elegido este diseño para que sea modular y escalable, dando la posibilidad de la integración de varias herramientas en cada fase.

En cuanto a los requisitos funcionales de la arquitectura, la primera fase de recolección de datos tiene que cumplir con diversas necesidades como obtener datos de múltiples fuentes y adaptarse a cambios en dichas fuentes. La fase de procesado tiene que ser capaz de realizar

la limpieza y normalización de los datos, eliminando los irrelevantes y estructurándolos para facilitar el análisis. En el almacenamiento se tienen que almacenar grandes cantidades de datos de manera eficiente y segura, permitir consultas complejas y ofrecer escalabilidad y flexibilidad a la hora de manejar los datos. En el modelado se tiene que asegurar una integración de los datos de distintas fuentes y se tienen que ofrecer herramientas para la visualización de los datos. Por último, en la difusión se tienen que generar informes y facilitar su distribución a las partes interesadas.

4.3 Implementación del modelo propuesto

Para el desarrollo del sistema y su implementación, se necesita emplear diversas herramientas que permitan realizar todos los procesos que forman las diferentes fases mencionadas. En este apartado, se presentan las mismas, así como la arquitectura que se va a usar en el caso práctico propuesto para el sistema de vigilancia e inteligencia según la metodología CRISP-DM.

4.3.1 Entorno tecnológico empleado

Como se ha explicado en puntos anteriores, el número de herramientas que se pueden utilizar para el desarrollo de una metodología de explotación de datos es muy elevado y no hay una herramienta que cubra todos los procesos que la forman. Por esto mismo, se van a seleccionar diferentes herramientas para cubrir uno o varios procesos. El análisis y elección de cada herramienta irá acorde a la arquitectura predefinida anteriormente. A continuación, se presenta la propuesta tecnológica en función de las capas de la arquitectura del sistema propuesto, detallando con todos los procesos y las herramientas que intervienen en cada uno de ellos:

- **Capa de recolección de datos:**

En primer lugar, la obtención de los datos se realizará a través de la técnica conocida como web scraping, existen varias librerías de Python que permiten el uso de esta técnica, como *selenium* o *playwright*, ambas tienen sus ventajas o inconvenientes, pero se selecciona *playwright* al ser óptima. Ambas simulan la navegación de un usuario por la web y van obteniendo la información necesaria, pero al ejecutarse *selenium* se abre una página de navegador y se puede ver como realiza la navegación, lo que consume más recursos. También, en este proceso, se realizará la recopilación de datos mediante APIs.

- **Capa de preparación de datos:**

Cuando ya se ha recopilado la información necesaria, es necesario llevar a cabo la preparación de datos. Esta fase consiste en limpiar y filtrar los datos, realizando una normalización y obteniendo solamente la información de valor. Este proceso también se realizará con el mismo código de elaboración propia del proceso anterior. Este punto se va a ejecutar apoyándose en la librería Pandas de Python, una librería de código abierto que cuenta con estructuras para la manipulación y el tratamiento de datos. Pandas se caracteriza por tener dos estructuras de datos principales, que son [69]:

- **Series**: se trata de un array unidimensional que es capaz de contener cualquier tipo de objeto.
- **DataFrame**: se trata de la unión de series.

Para este proyecto, se utilizará la estructura de DataFrames, donde los datos obtenidos se almacenan en distintas columnas para su posterior procesado, obteniendo diversos archivos de salida como por ejemplo un .csv.

- **Capa de almacenamiento de datos:**

Una vez se han recolectado y preparado los datos, se almacenan en una base de datos PostgreSQL, ya que es una base de datos relacionar y gratuita, que se puede instalar localmente, y que permite manejar una gran cantidad de datos. El almacenamiento de los datos se puede realizar una vez estos se han recolectado, pero como el sistema diseñado realiza la recolección y el procesado sobre el mismo programa, se ha decido hacerlo después. También, se ha decido realizar la carga de datos en dicha base de datos desde archivos .csv, pues es sencillo y pueden ser limpiados, transformados y cargados en diversas tablas [70].

- **Capa de modelado y análisis:**

Para la fase de modelado, se va a emplear Power BI. Es una herramienta que es capaz de unificar datos procedentes de diferentes fuentes, analizarlos y representarlos, creando información de valor. Proporciona modelos de datos que ayudan a contar una problemática o un asunto gracias a visualizaciones de datos y gráficos, permitiendo a sus usuarios desarrollar y compartir instantáneas actuales de cualquiera tema a tratar. También, por haber sido desarrollada con Machine Learning, puede hacer modelos predictivos en base a datos introducidos del pasado o presente [71].

Power BI contiene distintos componentes que ayudan en el proceso analítico y su representación; entre ellos se encuentran:

- Query: Más conocida como Power Query, permite modificar, eliminar y extraer datos de diversas fuentes.
- Pivot: Herramienta para modelar datos complejos y simples, estableciendo relaciones entre diferentes tablas.
- View: Es el componente fundamental para la visualización de datos. Recupera metadatos para el análisis conectándose a las fuentes de datos.
- Map: Se utiliza para la visualización de datos de características geoespaciales en 3D.
- Q&A: Motor lingüístico para resolver dudas.
- BI Destktop: Integra los componentes Query, Pivot y View, permitiendo crear modelos, informes y consultas de calidad.

Además, Power BI permite relacionar datos que provienen de distintos orígenes e integrarlos en un solo informe. Esta herramienta permite acceso a datos desde tablas de Excel o .csv (el tipo de tabla que se ha obtenido en el proceso anterior y, por tanto, la que se va a utilizar) hasta conexiones con bases de datos [72].

- **Capa de reportes y difusión:**

Para el último proceso, el de la difusión de la información y resultados obtenidos, así como al aplicar el sistema de vigilancia e inteligencia, Power BI permite exportarlos en diversos formatos, como Excel o PDF, facilitando este último punto.

Capítulo 5. Validación del modelo propuesto mediante un caso práctico

5.1 Introducción

En este capítulo se valida el sistema propuesto mediante su aplicación en un caso práctico de un modelo de un sistema de vigilancia e inteligencia. Este modelo va a ser un estudio práctico de la extracción y tratamiento de datos de Reddit en el que se comparará la información obtenida de dicha red social con información financiera proveniente de otras fuentes de información. A continuación, se exponen las diferentes fases del modelo presentadas en el punto 4.2.

5.2 Comprensión del negocio y los objetivos

5.2.1 Definición del problema

Las redes sociales han revolucionado la manera en la que las personas se comunican, consumen información y toman decisiones, lo que ha impactado en diferentes sectores, siendo uno de ellos el mundo financiero. Hoy en día, las plataformas generan un volumen enorme de datos que contienen información importante sobre opiniones, expectativas y actitudes hacia distintas circunstancias o eventos, que pueden afectar el comportamiento del mercado. Sin embargo, el gran volumen de estos datos y su complejidad, dificultan su análisis de manera efectiva, así como la extracción de información de valor para la toma de decisiones [73] [74].

5.2.2 Objetivos de negocio

El objetivo de este proyecto es analizar el impacto de las redes sociales en los mercados financieros. Este sistema permitirá a los usuarios:

- **Identificar tendencias:** Predecir tendencias del mercado a partir de datos de Reddit, como el análisis de volumen de comentarios con determinados tickers y tendencias relacionadas con activos financieros, empresas y eventos económicos.
- **Análisis de sentimiento**: Evaluar el sentimiento general del mercado hacia organizaciones específicas o diferentes eventos mediante el análisis de los comentarios.
- **Evaluación del impacto**: Medir el impacto de eventos ocurridos en las redes sociales sobre el precio de las acciones u otros indicadores, analizando foros dedicados a la discusión de mercados financieros como WallStreetBets en Reddit.

5.2.3 Éxito

El éxito del proyecto se medirá en función de la capacidad del sistema para alcanzar:

- **Precisión de las predicciones**: La precisión de las predicciones del sistema sobre las tendencias del mercado, el sentimiento y el impacto de los eventos de las redes sociales en el mercado financiero.
- **Oportunidades de inversión identificadas**: La capacidad del sistema para identificar oportunidades de inversión a través de los comentarios de los foros de Reddit.

5.3 Comprensión y recolección de los datos

5.3.1 Identificación de las fuentes de datos

En apartados anteriores se han presentado diversas fuentes de información útiles para desarrollar el sistema. A continuación, se indican las que se han seleccionado.

5.3.1.1 AlphaVantage.co

AlphaVantange.co es una plataforma y fuente de información valiosa, pues proporciona una gran cantidad de datos de material financiero de una manera sencilla. En un mundo dónde el panorama financiero se vuelve cada vez más complejo, ser capaz de obtener datos de mercado actualizados y fiables es de suma importancia. Esta plataforma lo hace posible y sus usuarios pueden conseguir ese conocimiento, de manera gratuita, a través de una API [75].

Uno de los puntos en los que destaca esta plataforma, es su amplia cobertura de los mercados financieros. Esta herramienta brinda datos sobre acciones, fondos cotizados (ETF), criptodivisas, divisas y noticias relacionadas con el ámbito bursátil, entre otros [75]. Proporciona, además, acceso a datos históricos y en tiempo real, ofreciendo a los usuarios la posibilidad de realizar un seguimiento de mercado, controlar tendencias y llevar a cabo un análisis en profundidad.

Como se ha mencionado, la información de esta fuente se obtiene a través de una API, la cual es intuitiva y de fácil manejo, haciendo posible que, a través de diferentes tipos de llamadas, se obtenga información estructurada y comprensible.

Además de la enorme cantidad de datos que proporciona y su fácil obtención, AlphaVantage.co también ofrece gráficos personalizados a los usuarios para que puedan realizar mejores análisis.

En conclusión, **AlphaVantage.co es una importante fuente de información bursátil por su gran cantidad de datos, herramientas analíticas, facilidad de uso y accesibilidad**. Esta plataforma brinda los recursos necesarios para navegar eficazmente por el complejo panorama financiero actual, ayudando a sus usuarios en la toma de decisiones.

5.3.1.2 Reddit

Reddit es una plataforma donde usuarios registrados pueden publicar contenido como texto, imágenes y videos y compartirlos. El contenido es votado hacia arriba o hacia abajo por otros usuarios, lo que determina su visibilidad, siendo más visto cuantos más votos positivos tenga. Este contenido se organiza en diversas comunidades o “subreddits”, diseñadas para tratar temas específicos.

Uno de estos “subreddits" es WallStreetBets, una comunidad de inversores que se caracteriza por un enfoque en estrategias arriesgadas y agresivas [76]. En concreto, esta comunidad ha demostrado la capacidad que tienen las redes sociales de influir en los mercados financieros. WallStreetBets, es conocida por el papel que tuvo en el caso de GameStop, donde sus usuarios

se coordinaron para comprar de forma masiva las acciones de ésta y hacer que suba su precio [76].

Como conclusión, **la comunidad WallStreetBets de Reddit, es óptima para examinar la influencia de las redes sociales en el mercado de valores**. La manipulación en el pasado del mercado ofrece diversas perspectivas sobre la conexión entre las comunidades en línea con los mercados financieros.

Reddit cuenta con una API que facilita la recopilación de datos, pero está muy limitada. Para este proyecto, se implementará una técnica de web scraping con el fin obtener una gran cantidad de datos, mencionada en al apartado 2.3.

5.3.1.3 Yahoo Finance

Yahoo Finance es una plataforma financiera que ofrece todo tipo de información, desde cotizaciones en tiempo real, datos de mercado o análisis de empresas [77]. También proporciona una amplia gama de análisis de noticias y comentarios sobre mercados financieros y tendencias económicas.

Al igual que AlphaVentage.co, Yahoo Finance cuenta con una API, disponible para varias plataformas, de aplicación sencilla, lo que facilita la explotación de sus datos.

Esta plataforma cuenta también con una sección de comentarios, donde los usuarios pueden discutir sobre los valores de la bolsa, intercambiar opiniones o hacer preguntas.

Durante muchos años, ha sido un portal respetado, debido a su información fiable y al gran número de usuarios que lo utilizan [78].

En conclusión, **Yahoo Finance es una fuente de datos fiable, debido a su longevidad e historial de información contrastada**, aparte de contar con una importante comunidad de usuarios que intercambian información día a día.

5.3.2 Exploración de los datos

A continuación, se presentan los datos que se va a obtener de las fuentes recientemente mencionadas, entre ellos podemos encontrar:

5.3.2.1 AlphaVantage.co

Se obtienen las noticias relacionadas con el mercado financiero. Como ejemplo del formato del archivo de salida se consigue lo siguiente:

Tabla 5.1 – Tabla de contenido AlphaVantage.co. Elaboración propia.

Campo	Tipo	Valor
Title	Título de la noticia	Will Nvidia Be Worth More Than Apple by 2030?
URL	Enlace	https://www.fool.com/investing/2023/12/17/will-nvidia-be-worth-more-than-apple-by-2030/
Time Published	Fecha y hora de publicación	2023-12-17T15:24:00
Authors	Autores	['Harsh Chauhan']
Summary	Resumen	Nvidia is growing faster than Apple right now, but will that be enough to help it overtake the world's most valuable company?
Banner Image	Imagen de la noticia	![Banner Image](https://media.ycharts.com/charts/600bc381a9b8de01b7bbe99cbb92e364.png)
Source	Fuente	Motley Fool
Category Within Source	Categoría sin fuente	n/a
Source Domain	Página web de la fuente	www.fool.com
Topics	Temas	[{'topic': 'Financial Markets', 'relevance_score': '0.538269'}, {'topic': 'Manufacturing', 'relevance_score': '0.5'}, {'topic': 'Earnings', 'relevance_score': '1.0'}, {'topic': 'Technology', 'relevance_score': '0.5'}]
Overall Sentiment Score	Puntuación global de sentimiento	0.24832
Overall Sentiment Label	Etiqueta global de sentimiento	Somewhat-Bullish
Ticker Sentiment	Indicadores bursátiles	[{'ticker': 'NVDA', 'relevance_score': '0.549362', 'ticker_sentiment_score': '0.421136', 'ticker_sentiment_label': 'Bullish'}, {'ticker': 'AAPL', 'relevance_score': '0.357507', 'ticker_sentiment_score': '0.185516', 'ticker_sentiment_label': 'Somewhat-Bullish'}]

5.3.2.2 Reddit

Se obtienen los comentarios de los distintos usuarios y se analizan sentimentalmente. Como ejemplo del formato del archivo de salida se tiene lo siguiente:

Tabla 5.2 – Tabla de contenido Reddit. Elaboración propia

Campo	Tipo	Valor
Date	Fecha en formato DD/MM/YYYY	31/05/2024
Comment	Comentario del usuario	Sold my DELL 155p for 14.4 at open at 22.7 now big sad

5.3.2.3 Yahoo Finance

Se obtiene el histórico de las capitalizaciones en bolsa de las empresas que se van a analizar, es decir, el precio de las acciones de las diez empresas más mencionadas en Reddit. Se obtienen estos datos para poder hacer una comparación de los comentarios de Reddit con el precio de las acciones diariamente. Como ejemplo del formato del archivo de salida, se consigue la Tabla 5.3:

Tabla 5.3 – Tabla de contenido Yahoo Finance. Elaboración propia.

Campo	Tipo	Valor
Date	Fecha en formato YYYY-MM-DD HH:MM:SS	2024-05-23 00:00:00-04:00
Open	Valor de apertura	532.9600219726562
High	Valor más alto	533.0700073242188
Low	Valor más bajo	524.719970703125
Close	Valor de cierre	525.9600219726562
Volume	Cantidad total de acciones	57211200
Dividends	Dividendos	0.0
Stock Splits	Acciones	0.0
ticker	Identificador bursátil	SPY
Capital Gains	Ganancias de venta de acción por venta a un precio mayor que el de compra	0.0

5.4 Preparación de los datos

En este apartado se realiza el siguiente proceso de limpieza de datos, presente en la metodología CRISP-DM.

5.4.1 Limpieza de los datos

En este apartado se van a eliminar los datos irrelevantes, erróneos o incompletos, para garantizar la calidad del análisis. Para ello se van a utilizar diferentes técnicas estándar dentro de este proceso, entre ellas se pueden encontrar [79]:

- Eliminación de los valores nulos: Se identifican y eliminan valores nulos en los distintos datasets. Un ejemplo de ello es en el dataset de AlphaVantage.co, donde se verifica que todos los valores críticos como *'title*', '*url*', '*time_published'* y '*authors*' no contengan valores nulos.
- Corrección de errores ortográficos: Se estandarizan nombres de autores y empresas para evitar inconsistencias.
- Identificación de valores duplicados.

5.4.2 Transformación de los datos

Con el objetivo de preparar los datos para el modelo, se van a realizar la siguiente transformación en los datos:

- Se hace un análisis sentimental de los comentarios de Reddit a través de la librería VADER [80]. Se trata de una librería para el análisis sentimental, en inglés, de texto procedente de redes sociales. Al realizar el análisis se añaden las columnas '*Sentiment'* y '*Sentimient_Score'* al conjunto de datos de Reddit.
- Se obtiene el ticker contenido en los comentarios de Reddit. Se añade la columna 'Ticker' al conjunto de datos.
- Se crea un nuevo dataset con los diez tickers más mencionados en dichos comentarios de Reddit, para centrar el análisis sobre ellos. Sólo se obtendrán datos de Yahoo Finance y de AlphaVantage.co sobre ellos.

5.4.3 Integración de los datos

En este apartado se utilizarán técnicas para integrar los datos de diferentes fuentes con el objetivo de crear un conjunto de datos consistente y coherente. Estas técnicas son:

- Resolución de la falta de coincidencia de identificadores: Se estandarizan los identificadores entre los datasets. Como, por ejemplo, se hace que todos los conjuntos de datos tengan una columna '*Ticker*' y con el mismo formato.
- Unificación de formatos de fecha y hora.

5.4.4 Datos obtenidos

Tras los procesos realizados sobre los datos, se obtiene lo siguiente:

- Reddit: En este conjunto de datos se crean las columnas '*Sentiment*', '*Sentiment_Score*', '*Ticker*' y se modifica el valor de la columna '*Date*', como se puede observar en la Tabla 5.4.

Tabla 5.4 - Tabla de contenido Reddit. Elaboración propia.

Campo	Tipo	Valor
Date	Fecha en formato YYYY-MM-DD	2024-05-31
Comment	Comentario del usuario	Sold my DELL 155p for 14.4 at open at 22.7 now big sad
Ticker	Identificador bursátil	DELL
Sentiment	Sentimiento del comentario	negative
Sentiment_Score	Polaridad del sentimiento	-0.4767

- Yahoo Finance: Como se puede aprecier en la Tabla 5.5, se ha modificado el formato de la fecha para que coincida con el del resto de conjuntos de datos.

Tabla 5.5 – Tabla de contenido Yahoo Finance. Elaboración propia

Campo	Tipo	Valor
Date	Fecha en formato YYYY-MM-DD	2024-05-23
Open	Valor de apertura	532.9600219726562
High	Valor más alto	533.0700073242188
Low	Valor más bajo	524.719970703125
Close	Valor de cierre	525.9600219726562
Volume	Cantidad total de acciones	57211200
Dividends	Dividendos	0.0
Stock Splits	Acciones	0.0
ticker	Identificador bursátil	SPY
Capital Gains	Ganancias de venta de acción por venta a un precio mayor que el de compra	0.0

- AlphaVantage.co: En este caso, se han eliminado varias columnas que no eran de utilidad para el caso de estudio. También se han creado nuevas columnas, como las

numerosas columnas de '*ticker*', '*ticker_revelance_score*' y '*ticker_sentiment_score*', que antes estaban contenidas en las columnas '*Ticker Sentiment*' y las diversas columnas de '*topics*' que antes estaban contenidas en la misma columna '*Topics*'. Esto se puede observar en la Tabla 5.6.

Tabla 5.6 – Tabla de contenido AlphaVantage.co. Elaboración propia.

Campo	Tipo	Valor
title	Título	Futures: 3 AI Giants Fall On Earnings; Musk's $56 Bil Pay Deal Voided
url	Enlace	https://www.investors.com/market-trend/stock-market-today/dow-jones-futures-tech-giants-microsoft-google-amd-earnings-elon-musk-56-billion-pay-plan/
time_published	Fecha publicación	2024-01-30
authors	Autores	{'Investor's Business Daily', 'ED CARSON'}
summary	Resumen	Dow Jones Futures: Microsoft, Google, AMD Fall On Earnings. Elon Musk Loses $56 Billion Pay Plan Investor's Business Daily ...
banner_image	Imagen	https://www.investors.com/wp-content/uploads/2023/12/Stock-EyeAIgen-adobe.jpg
source	Fuente	Investors Business Daily
source_domain	Página web de la fuente	www.investors.com
overall_sentiment_score	Puntuación general sentimiento	-0.089
overall_sentiment_label	Sentimiento general	Neutral
topic_1	Tema	Life Sciences
topic_1_relevance_score	Puntuación de tema detectado	0.2
topic_2	Tema	Energy & Transportation
topic_2_relevance_score	Puntuación de tema detectado	0.2
topic_3	Tema	Technology
topic_3_relevance_score	Puntuación de tema detectado	0.2
ticker_1	Ticker	MSFT
ticker_1_relevance_score	Puntuación de ticker detectado	0.410697
ticker_1_sentiment_score	Puntuación de sentimiento del ticker	-0.09547

ticker_1_sentiment_label	Etiqueta de sentimiento de ticker	Neutral
ticker_2	Ticker	GOOG
ticker_2_relevance_score	Puntuación de ticker detectado	0.30021
ticker_2_sentiment_score	Puntuación de sentimiento del ticker	-0.080922
ticker_2_sentiment_label	Etiqueta de sentimiento de ticker	Neutral
ticker_3	Ticker	MPC
ticker_3_relevance_score	Puntuación de ticker detectado	0.061473
ticker_3_sentiment_score	Puntuación de sentimiento del ticker	0.0
ticker_3_sentiment_label	Etiqueta de sentimiento de ticker	Neutral
ticker_4	Ticker	NVDA
ticker_4_relevance_score	Puntuación de ticker detectado	0.182967
ticker_4_sentiment_score	Puntuación de sentimiento del ticker	-0.136553
ticker_4_sentiment_label	Etiqueta de sentimiento de ticker	Neutral
ticker_5	Ticker	AAPL
ticker_5_relevance_score	Puntuación de ticker detectado	0.182967
ticker_5_sentiment_score	Puntuación de sentimiento del ticker	-0.138782
ticker_5_sentiment_label	Etiqueta de sentimiento de ticker	Neutral

- Top 10 tickers:

Tabla 5.7 – Tabla de contenido Top 10 tickers. Elaboración propia

Campo	Tipo	Valor
Ticker	Indiciador busátil	NVDA
Number_of_Mentions	Número de menciones	7436

5.4.5 Almacenamiento de los datos

En la presentación de las fases de una metodología CRISP-DM, realizada en el Estado del Arte, se puede observar que la fase de almacenamiento de los datos no existe. Esto se debe a que

el almacenamiento está incluido en el proceso de preparación porque la obtención y limpieza de datos se realizan al mismo tiempo, en el mismo programa diseñado. Lo correcto sería realizar el almacenamiento una primera vez tras la obtención de los datos y otra segunda después de la limpieza de éstos.

A continuación, en la siguiente Figura 5.1 se presenta el esquema que se define para almacenar los datos en PostgreSQL. Se ha utilizado un tipo de base de datos relacional debido a que la información se encuentra estructurada. También, se ha definido este modelo de datos para poder conectar los distintos conjuntos de datos mediante el ticker y, en algunos casos, también con la fecha. Se utiliza la tabla de *'reddit_comments'* como la principal.

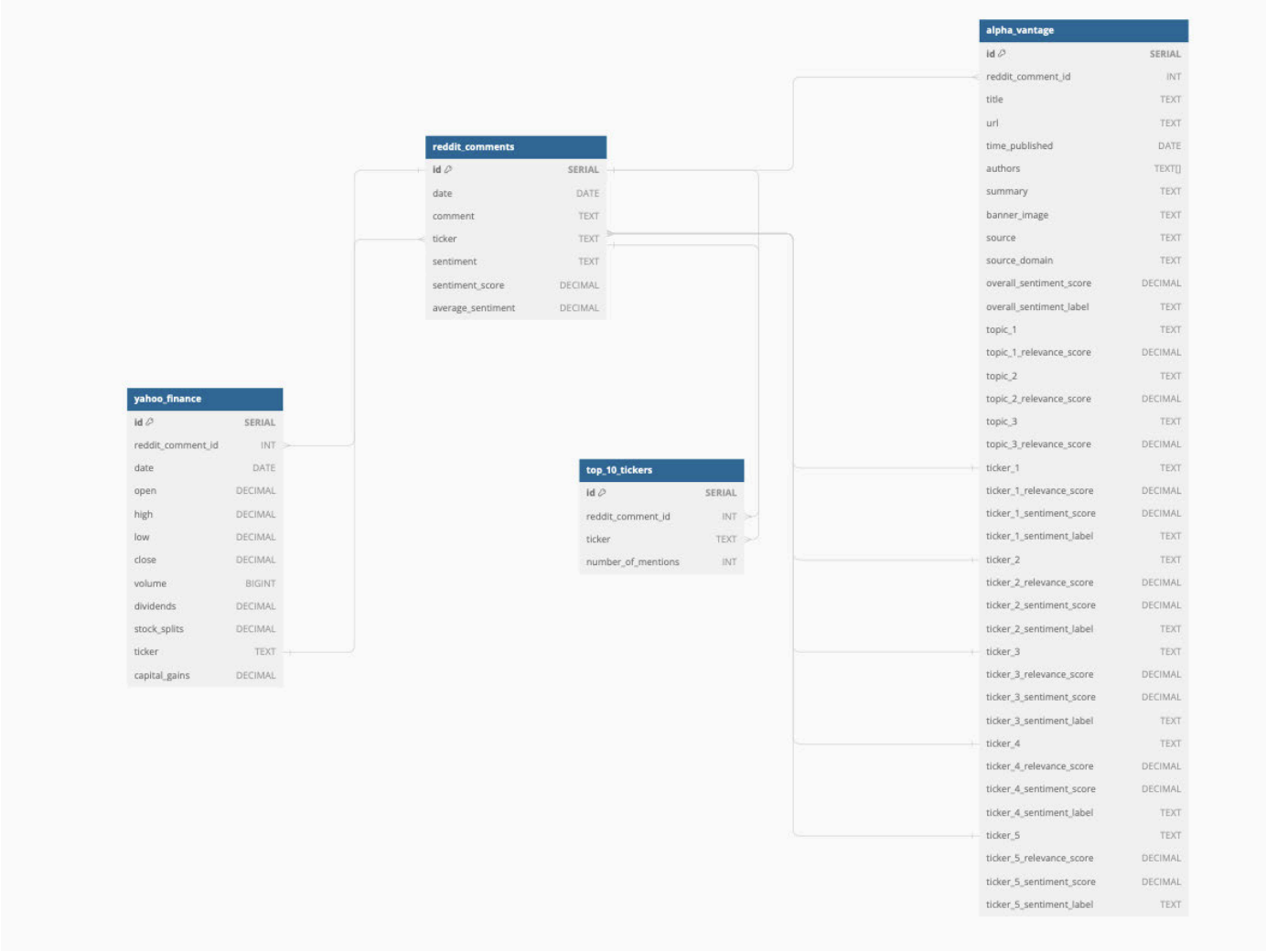

Figura 5.1 – Esquema de la base de datos. Elaboración propia.

5.5 Modelado

En este apartado se van a desarrollar y visualizar KPIs (Indicadores Clave de Desempeño) en Power BI. Con este modelado con KPIs no se obtiene un modelo predictivo tradicional, sino que dará una visión cuantificable y clara de la relación de los comentarios de los usuarios de

Reddit, con el precio de las acciones u otros indicadores de mercado financiero. Se logrará a través de la creación de gráficos interactivos y paneles en la herramienta que ya se ha mencionado, Power BI. También, se realizará la búsqueda de diversos patrones en los datos para mejorar la toma de decisiones.

5.5.1 Selección de los KPIs

El objetivo de la selección de los KPIS es entender la influencia de los comentarios de Reddit, más concretamente del subgrupo WallStreetBets, en el mercado financiero. Para ello, los KPIs más relevantes son [81] [82]:

- **Sentimiento promedio por Ticker**: Mide el sentimiento general, pudiendo ser negativo, positivo o neutro, de los comentarios de los usuarios por cada ticker. El sentimiento de los comentarios en redes sociales es un indicador del comportamiento futuro de mercado.
- **Volumen de comentarios por Ticker:** Un volumen alto puede relacionarse con movimientos importantes en el mercado.
- **Tendencia de precio de cierre**: Monitorea el precio de las acciones al cierre del mercado financiero a lo largo del tiempo. Es fundamental ya que tiene una relación directa a como las acciones responden al sentimiento y al volumen de comentarios.
- **Relación entre el precio de cierre y el sentimiento**: Compara el precio de mercado al cierre con el sentimiento diario encontrado en los comentarios de Reddit. Permite entender la influencia del sentimiento de los usuarios en las variaciones de precio de las acciones.
- **Número de noticias por Ticker:** Contiene el número de noticias que hay en el mismo período en el que se obtienen los comentarios de Reddit, además del sentimiento de dichas noticias. Algunas organizaciones estudiadas no contarán con este gráfico al no haber noticias suyas en el período establecido.

5.5.2 Desarrollo de los KPIs

En este punto este punto se realiza la obtención de los KPIs presentados en el apartado anterior para poder realizar la posterior evaluación de modelado.

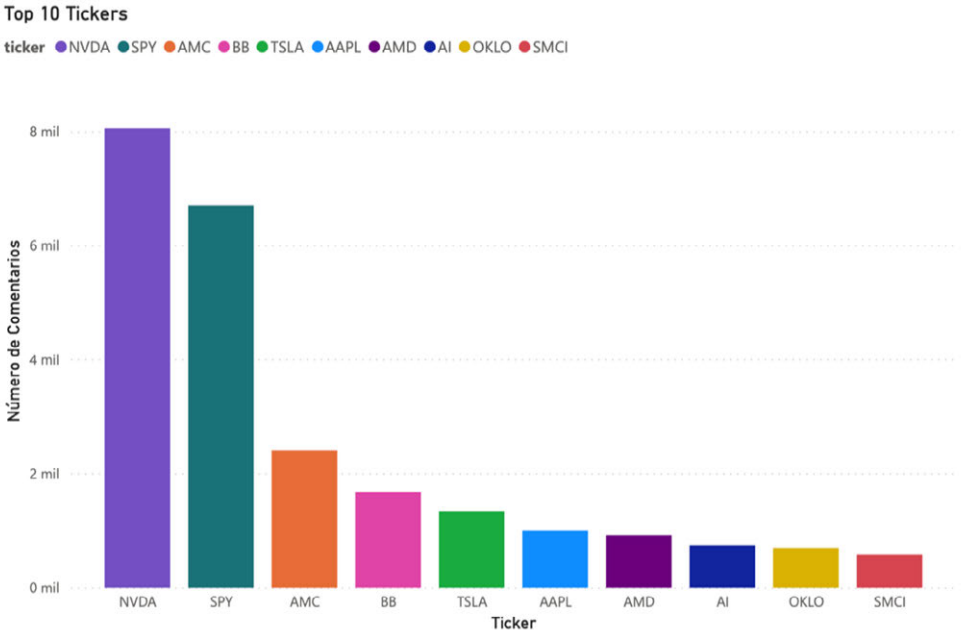

Figura 5.2 - Top 10 tickers. Elaboración propia.

Esta primera Figura 5.2 muestra los diez tickers más mencionados en el grupo de Reddit, son sobre los cuales se van a obtener los KPIs, mencionados anteriormente, para realizar el análisis. En la gráfica se pueden observar los siguientes tickers:

- NVDA: Se corresponde con la empresa Nvidia.
- SPY: Es el fondo indexado SPDR S&P 500.
- AMC: Es el ticker de AMC Entertainment Holdings Inc.
- BB: Se trata de la empresa BlackBerry.
- TSLA: Hace referencia a Tesla.
- AAPL: Representa a Apple.
- AMD: Se refiere a Advanced Micro Devices.
- AI: Se corresponde a C3.ai Inc.
- OKLO: Es Oklo Inc.
- SMCI: Hace referencia a Supermicro.

Por lo tanto, **con relación a las gráficas de NVDA**, se consigue lo siguiente:

- **Volumen de comentarios por Ticker**. Como se observa en Figura 5.3, hay un número alto de comentarios, con una distribución importante entre comentarios negativos, positivos y neutros.

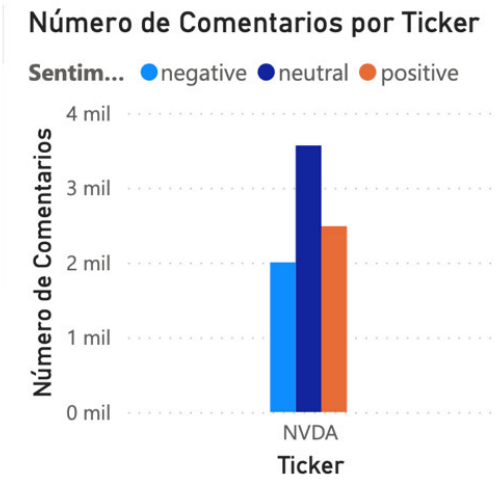

Figura 5.3 - Volumen de comentarios sobre NVDA. Elaboración propia.

- **Relación entre el precio de cierre y el sentimiento.** En la Figura 5.4 se puede apreciar hay un sentimiento positivo debido al aumento del precio de las acciones.

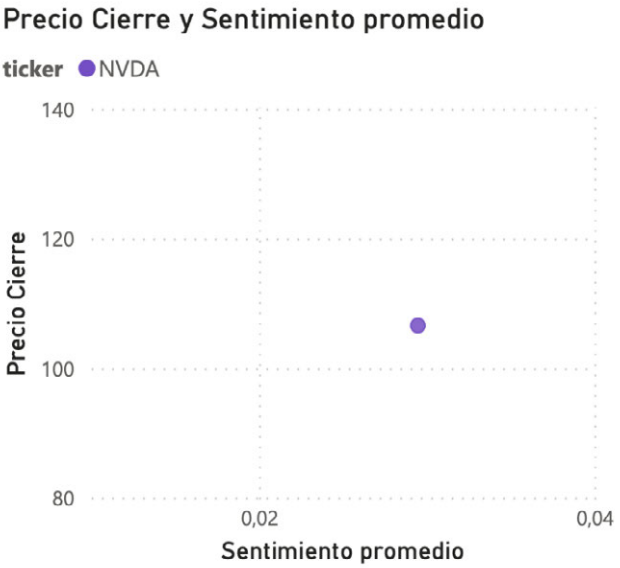

Figura 5.4 - Precio de cierre por sentimiento promedio de NVDA. Elaboración propia.

- **Número de noticias por Ticker.** En esta Figura 5.5 también se muestra variedad, con noticias neutras y de carácter alcista o positivo.

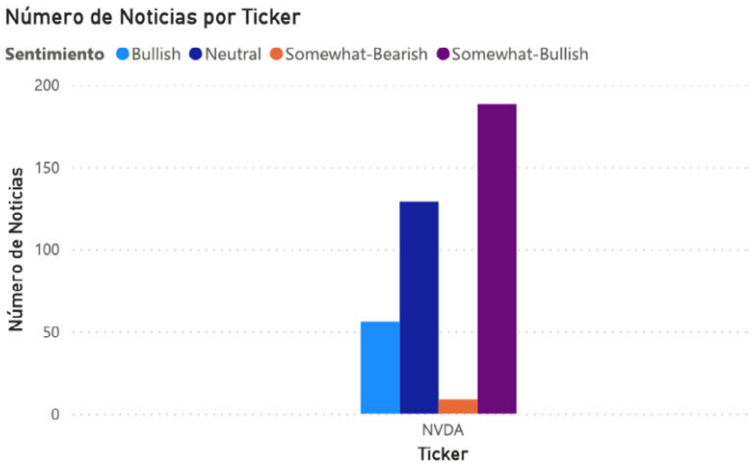

Figura 5.5 - Número de noticias sobre NVDA. Elaboración propia.

- **Tendencia de precio de cierre**. En la Figura 5.6 se muestra una tendencia de aumento de precio en el período analizado.

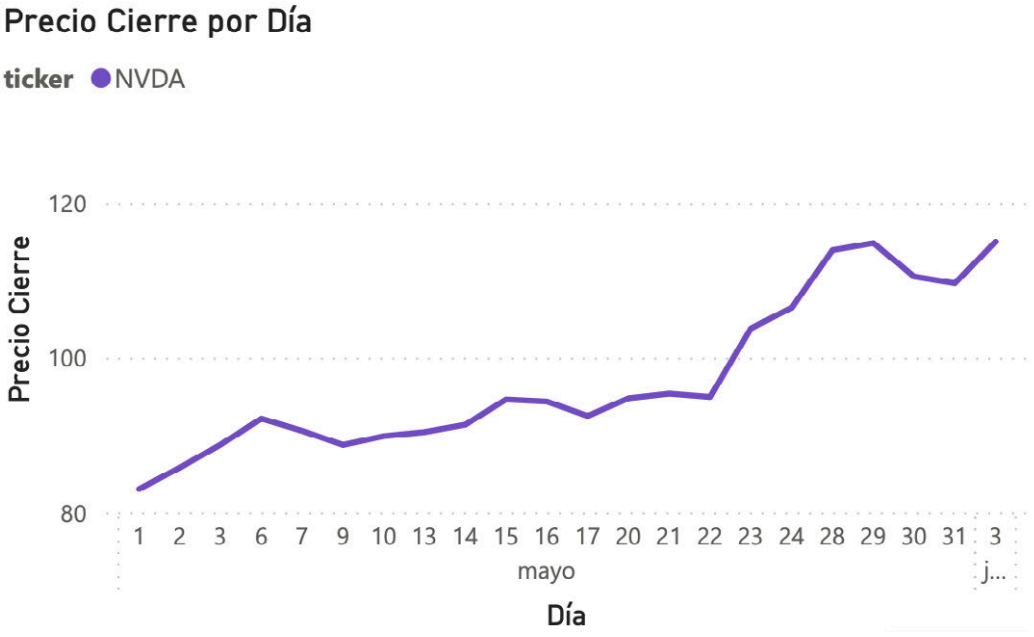

Figura 5.6 - Precio de cierre de NVDA. Elaboración propia.

- **Sentimiento promedio por Ticker.** El sentimiento promedio que se observa en la Figura 5.7 muestra períodos de percepción negativa y otros con percepción positiva entre los usuarios de Reddit.

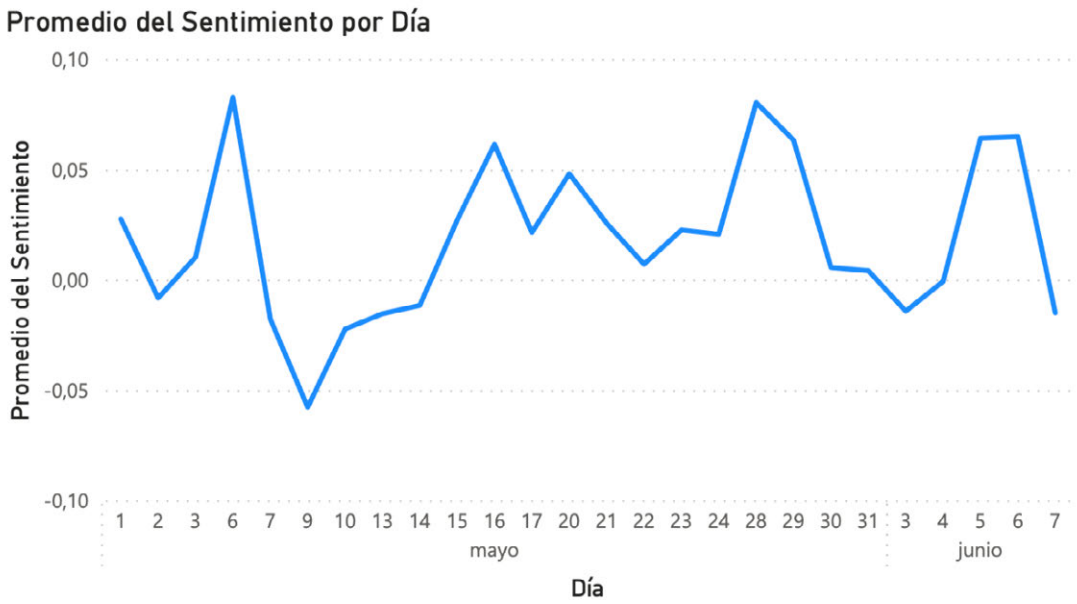

Figura 5.7 - Promedio de sentimiento de NVDA por día. Elaboración propia.

Con relación a las gráficas de SPY, se obtiene lo siguiente:

- **Volumen de comentarios por Ticker.** Como se aprecia en la Figura 5.8, hay una gran cantidad de comentarios, mayoritariamente positivos o neutros.

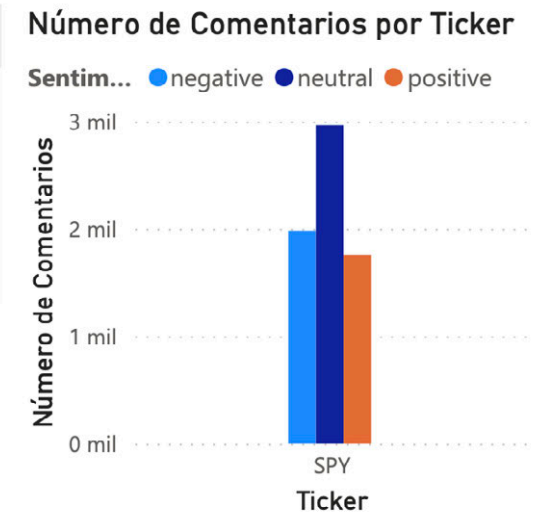

Figura 5.8 - Volumen de comentarios de SPY. Elaboración propia

- **Relación entre el precio de cierre y el sentimiento.** En esta Figura 5.9 se puede ver como la correlación es menos evidente que en el caso de NVDA.

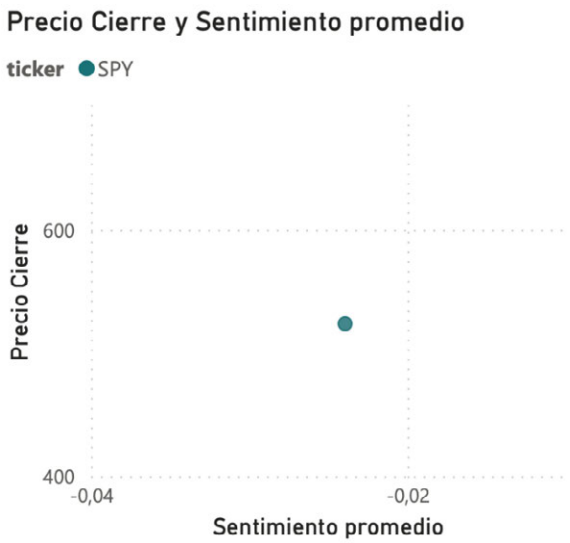

Figura 5.9 – Precio de cierre por sentimiento promedio de SPY. Elaboración propia.

- **Tendencia de precio de cierre**. La Figura 5.10 muestra como la tendencia es constante, con pequeñas variaciones.

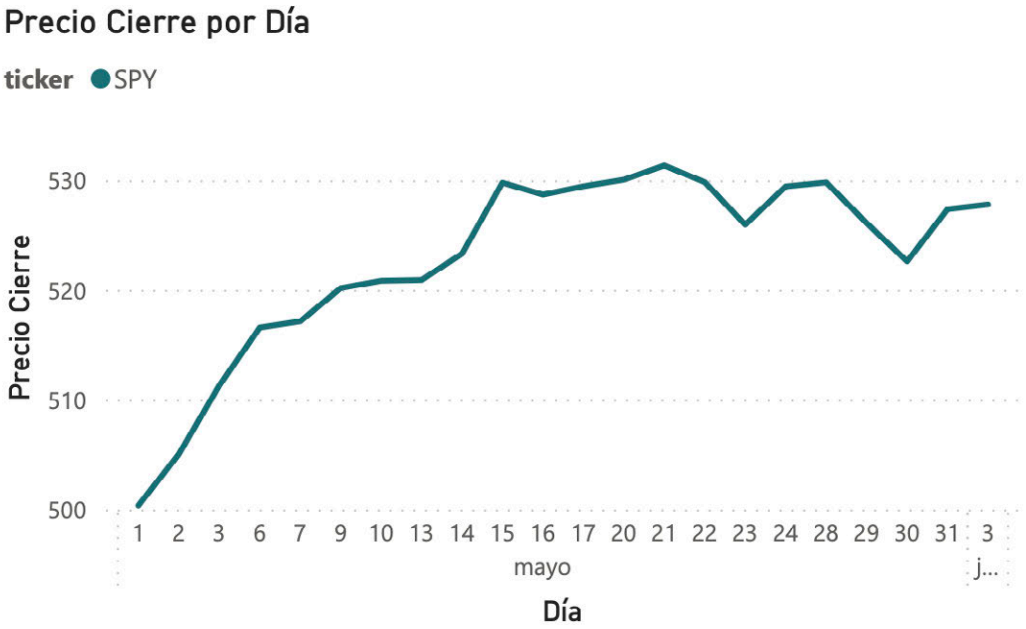

Figura 5.10 – Precio de cierre de SPY. Elaboración propia.

- **Sentimiento promedio por Ticker.** El sentimiento promedio de la Figura 5.11 tiende a la neutralidad, pero es ligeramente positivo.

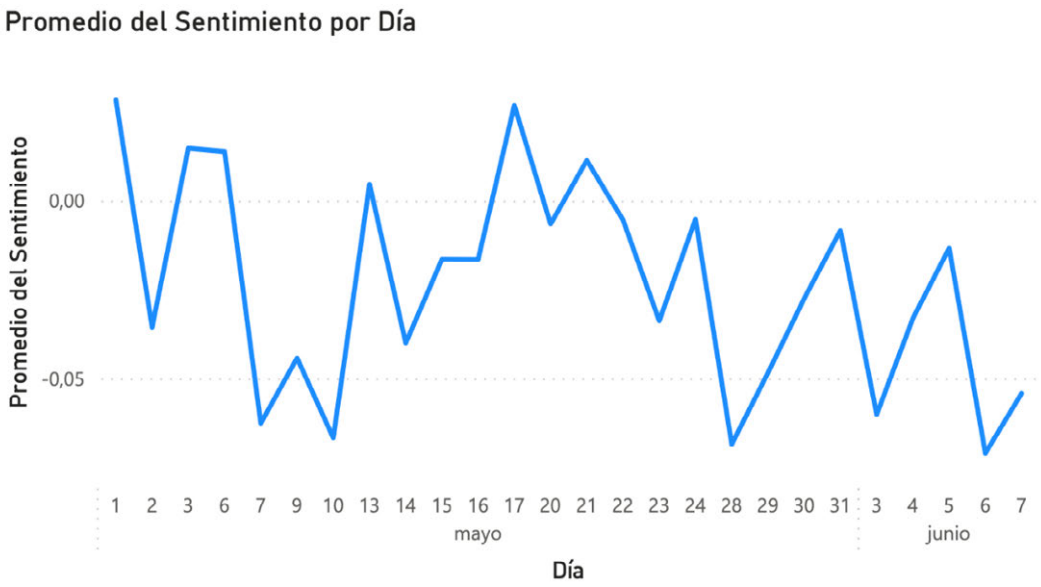

Figura 5.11 - Promedio de sentimiento de SPY por día. Elaboración propia

Con respecto a las gráficas de AMC:

- **Volumen de comentarios por Ticker.** En la Figura 5.12 se observa su popularidad en Reddit, debido al alto volumen de comentarios.

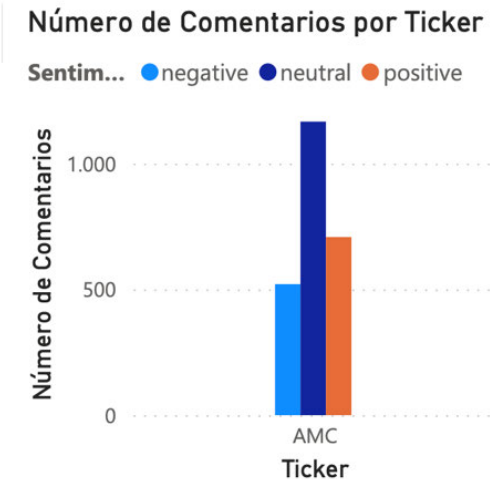

Figura 5.12 - Volumen de comentarios sobre AMC. Elaboración propia.

- **Relación entre el precio de cierre y el sentimiento.** En esta Figura 5.13 se puede apreciar una fuerte correlación entre los picos de precio de cierre de mercado y los de sentimiento positivo.

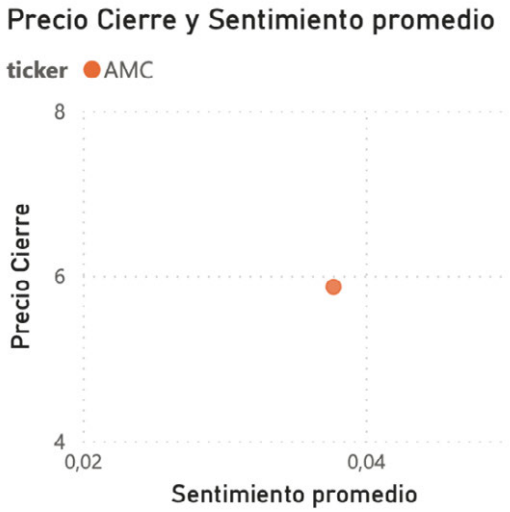

Figura 5.13 – Precio de cierre por sentimiento promedio de AMC. Elaboración propia

- **Número de noticias por Ticker.** En este caso, en la Figura 5.14 se presentan noticias principalmente alcistas y neutras.

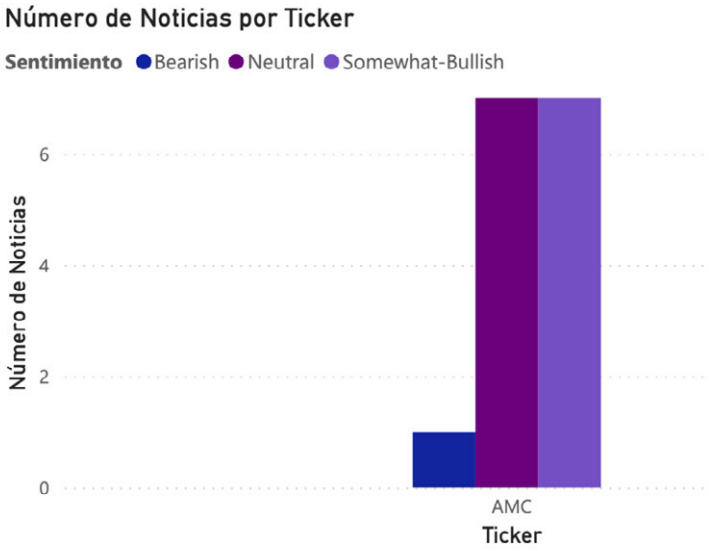

Figura 5.14 – Número de noticias sobre AMC. Elaboración propia.

- **Tendencia de precio de cierre.** La Figura 5.15 representa una tendencia de precio de cierre con grandes fluctuaciones.

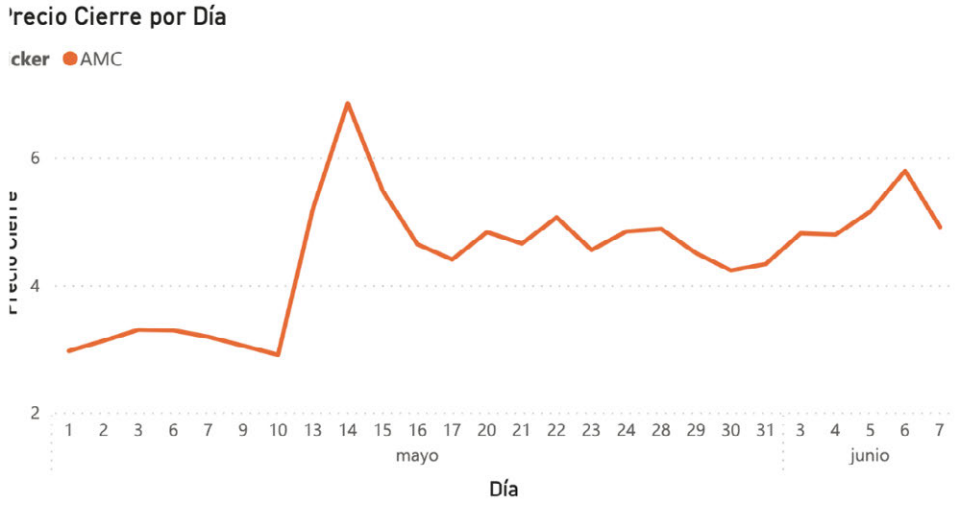

Figura 5.15 - Precio de cierre de AMC. Elaboración propia.

- **Sentimiento promedio por Ticker.** En la Figura 5.16 se puede distinguir un sentimiento promedio variante, pero que tiende a lo positivo.

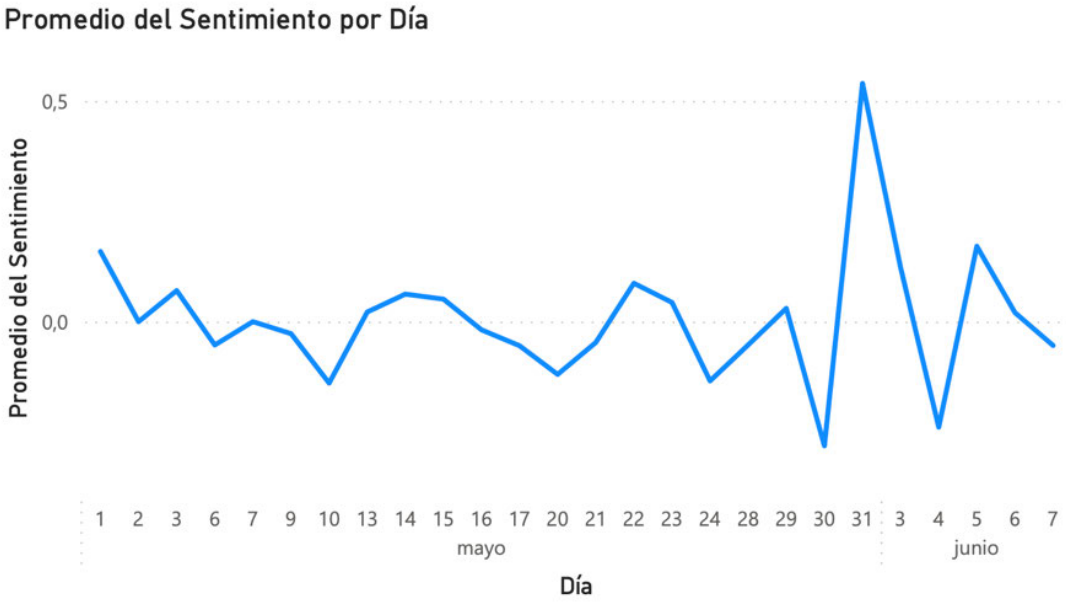

Figura 5.16 - Promedio de sentimiento de AMC por día. Elaboración propia.

Sobre BB, se obtienen las siguientes gráficas:

- **Volumen de comentarios por Ticker.** En la Figura 5.17 se presenta un volumen moderado de datos, comparado con los tickers anteriores, con unos sentimientos muy balanceados.

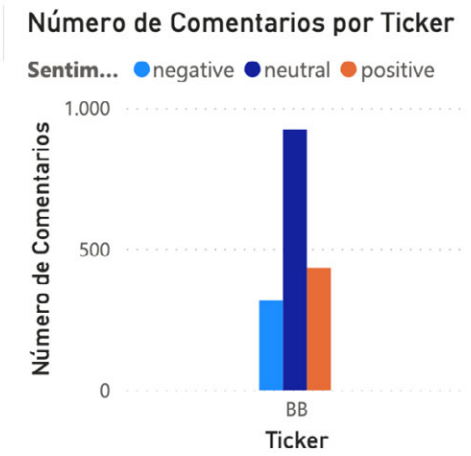

Figura 5.17 - Volumen de comentarios sobre BB. Elaboración propia.

- **Relación entre el precio de cierre y el sentimiento.** La relación entre el precio de cierre y el sentimiento que se representa en la Figura 5.18 muestra una correlación más débil que en el caso de NVDA y AMC.

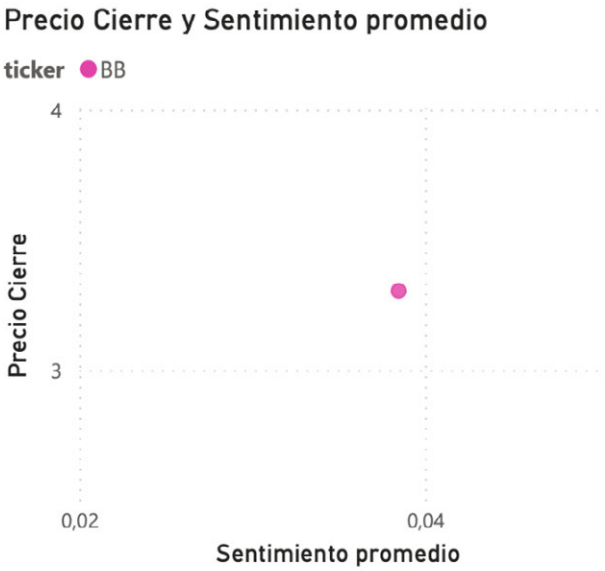

Figura 5.18 – Precio de cierre por sentimiento promedio de BB. Elaboración propia.

- **Tendencia de precio de cierre.** Se observa una tendencia ligeramente ascendente en la Figura 5.19.

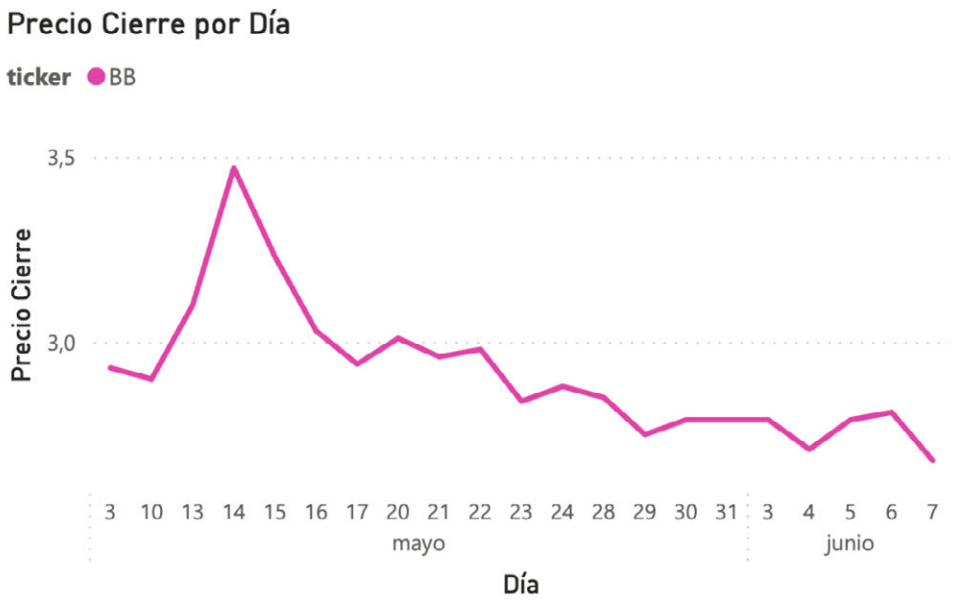

Figura 5.19 - Precio de cierre de BB. Elaboración propia.

- **Sentimiento promedio por Ticker.** La Figura 5.20 muestra una tendencia hacia lo positivo.

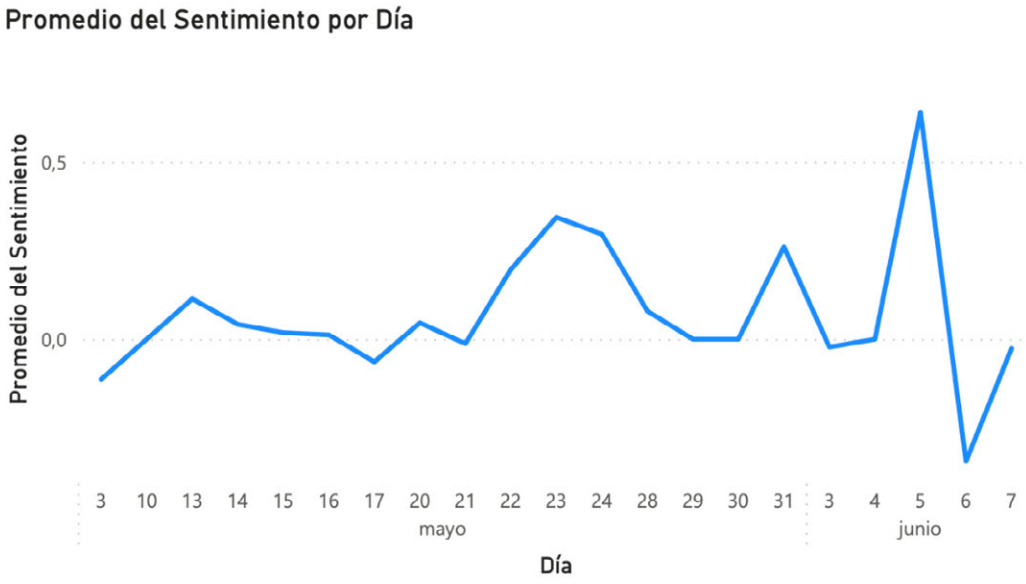

Figura 5.20 - Promedio de sentimiento de BB por día. Elaboración propia.

En referencia a las gráficas de TSLA, se consigue lo siguiente:

- **Volumen de comentarios por Ticker.** En la Figura 5.21 se aprecia que TSLA tiene un número elevado de comentarios, reflejando su relevancia en Reddit.

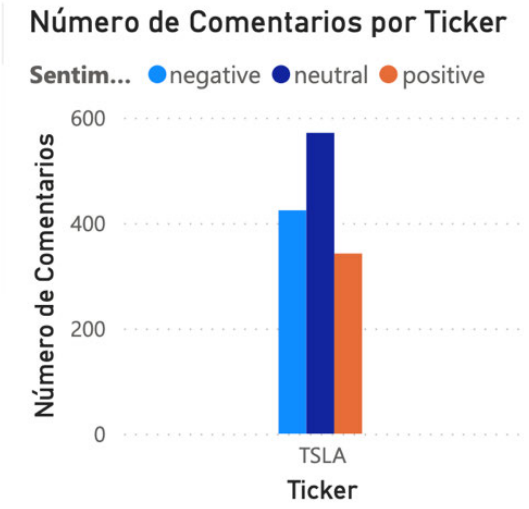

Figura 5.21 - Volumen de comentarios sobre TSLA. Elaboración propia.

- **Relación entre el precio de cierre y el sentimiento.** En este caso, en la Figura 5.22 se observa una correlación moderada.

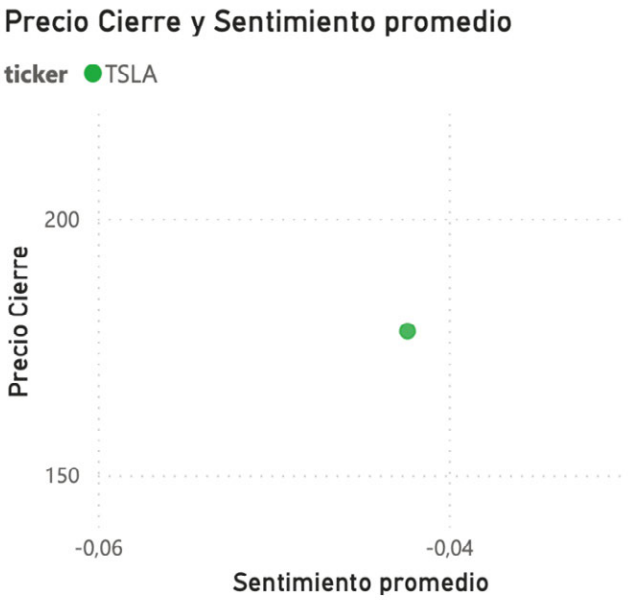

Figura 5.22 – Precio de cierre por sentimiento promedio de TSLA. Elaboración propia.

- **Número de noticias por Ticker.** La Figura 5.23 muestra que las noticias relacionadas con Tesla son principalmente neutras.

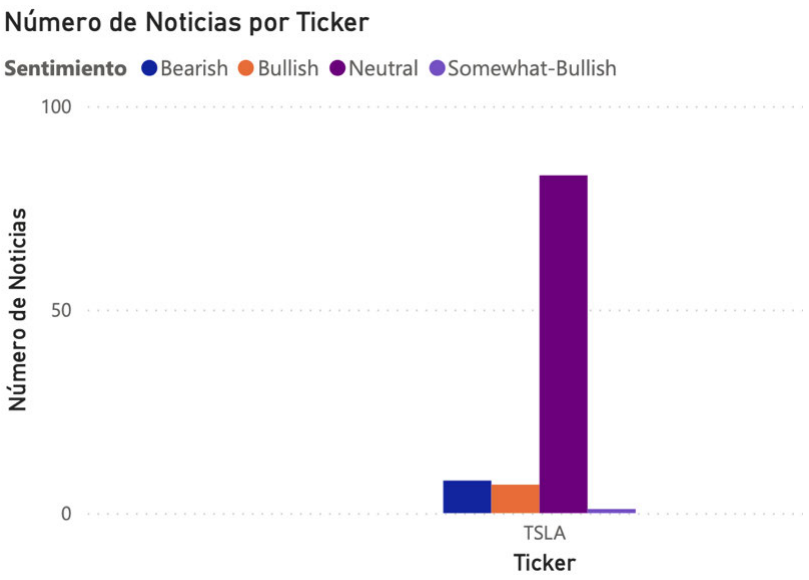

Figura 5.23 – Número de noticias sobre TSLA. Elaboración propia.

- **Tendencia de precio de cierre.** La Figura 5.24 representa que el precio de cierre tiene una tendencia claramente ascendente.

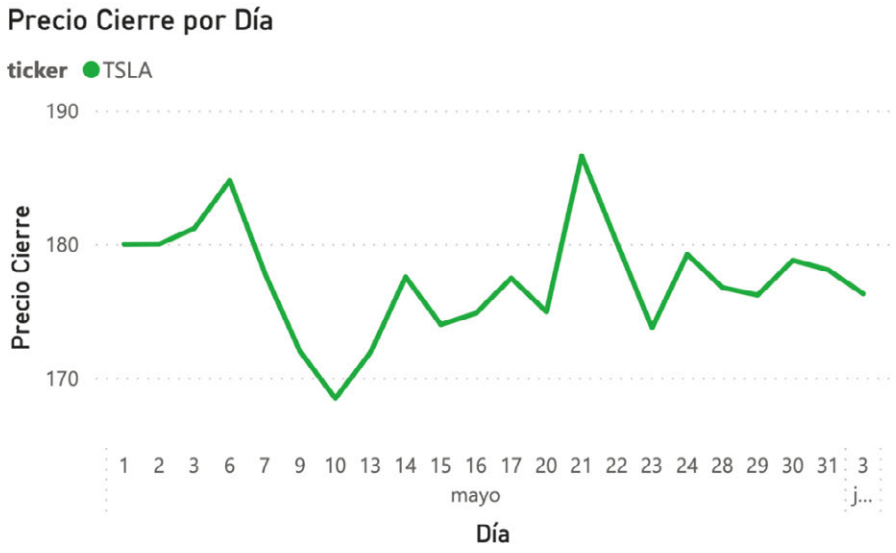

Figura 5.24 - Precio de cierre de TSLA. Elaboración propia

- **Sentimiento promedio por Ticker**. En este caso, como se puede apreciar en la Figura 5.25, el sentimiento promedio es mixto, con variaciones muy importantes a lo largo del tiempo.

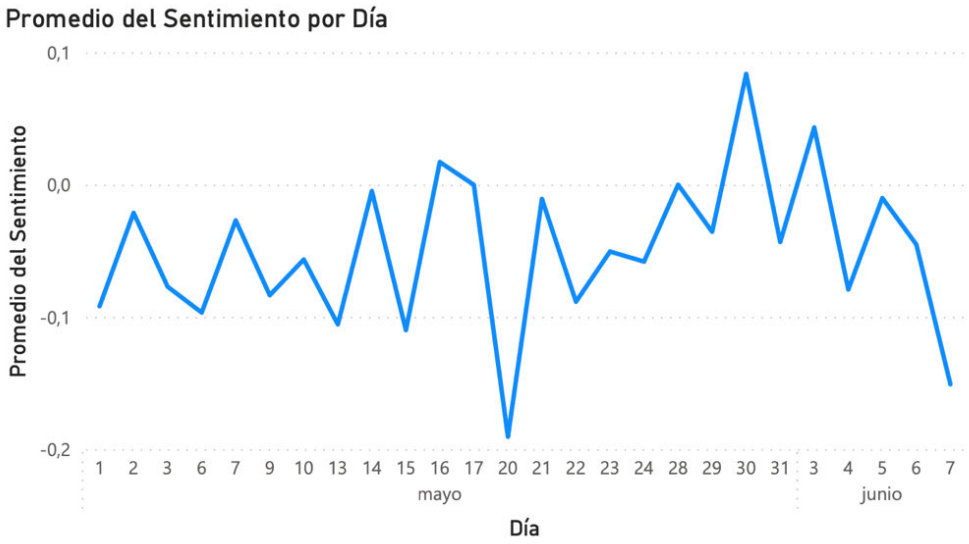

Figura 5.25 - Promedio de sentimiento de TSLA por día. Elaboración propia.

Por último, sobre las gráficas de AAPL, AI, AMD, OKLO y SMCI:

- **Volumen de comentarios por Ticker.** Como se puede observar en la Figura 5.26, varía mucho entre los tickers, pero AAPL y AMD muestran un alto volumen de comentarios.

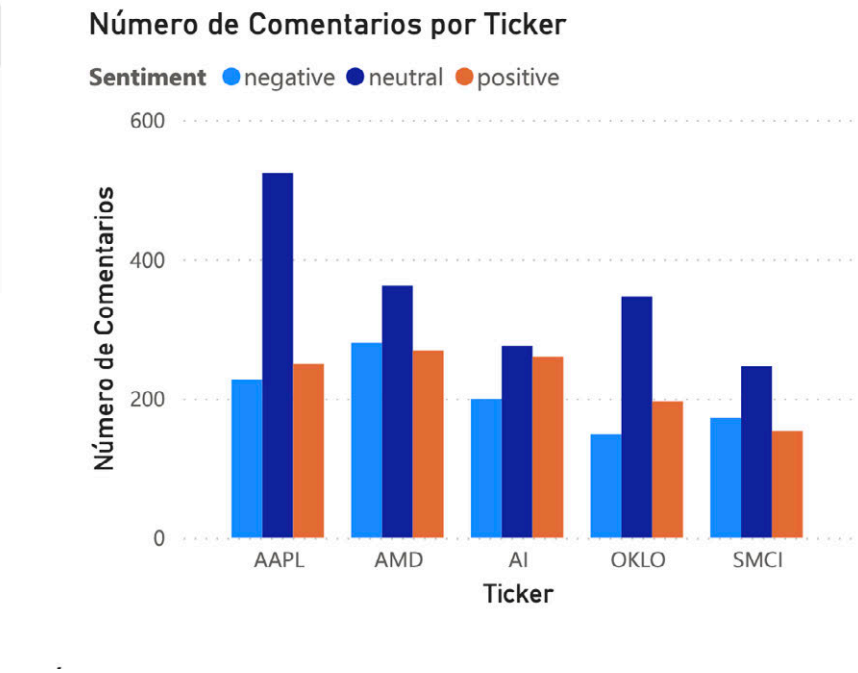

Figura 5.26 - Volumen de comentarios sobre AAPL, AI, AMD, OKLO y SMCI. Elaboración propia.

- **Relación entre el precio de cierre y el sentimiento.** Como se puede apreciar en la Figura 5.27, la correlación es más fuerte en los casos de AMD y AAPL.

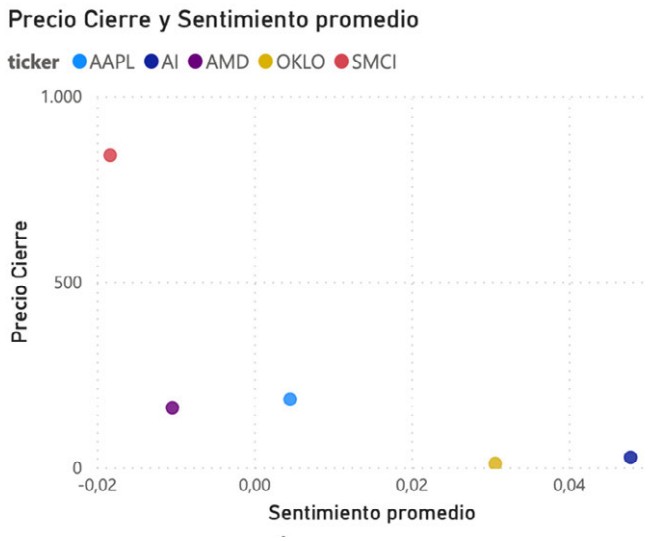

Figura 5.27 – Precio de cierre por sentimiento promedio de AAPL, AI, AMD, OKLO y SMCI. Elaboración propia.

- **Número de noticias por Ticker.** En la Figura 5.28 se observa que las noticias son principalmente alcistas y neutras.

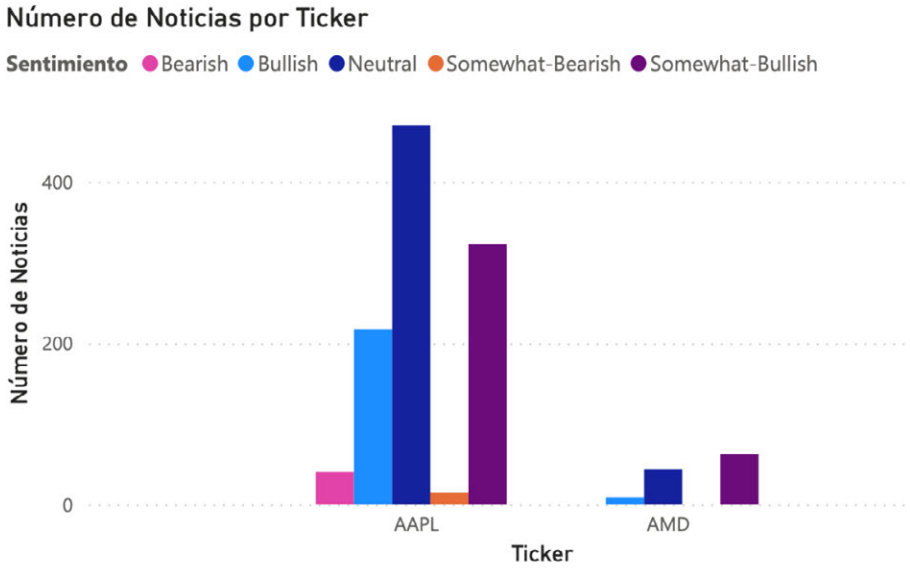

Figura 5.28 – Número de noticias sobre AAPL, AI, AMD, OKLO y SMCI. Elaboración propia.

- **Tendencia de precio de cierre.** La tendencia del precio del cierre de las acciones que se aprecia en la Figura 5.29 es generalmente positiva.

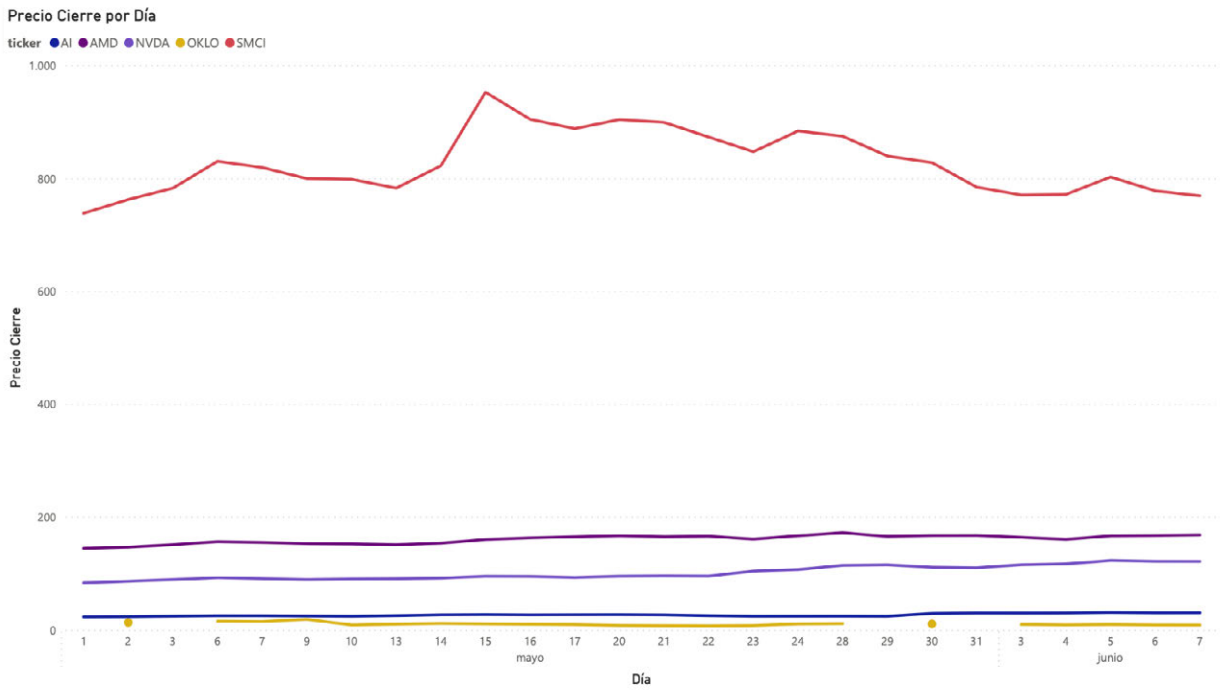

Figura 5.29 - Precio de cierre de AAPL, AI, AMD, OKLO y SMCI. Elaboración propia.

- **Sentimiento promedio por Ticker.** Como se puede apreciar en la Figura 5.30, el sentimiento promedio fluctúa mucho entre los distintos tickers, pero generalmente tiende hacia lo positivo.

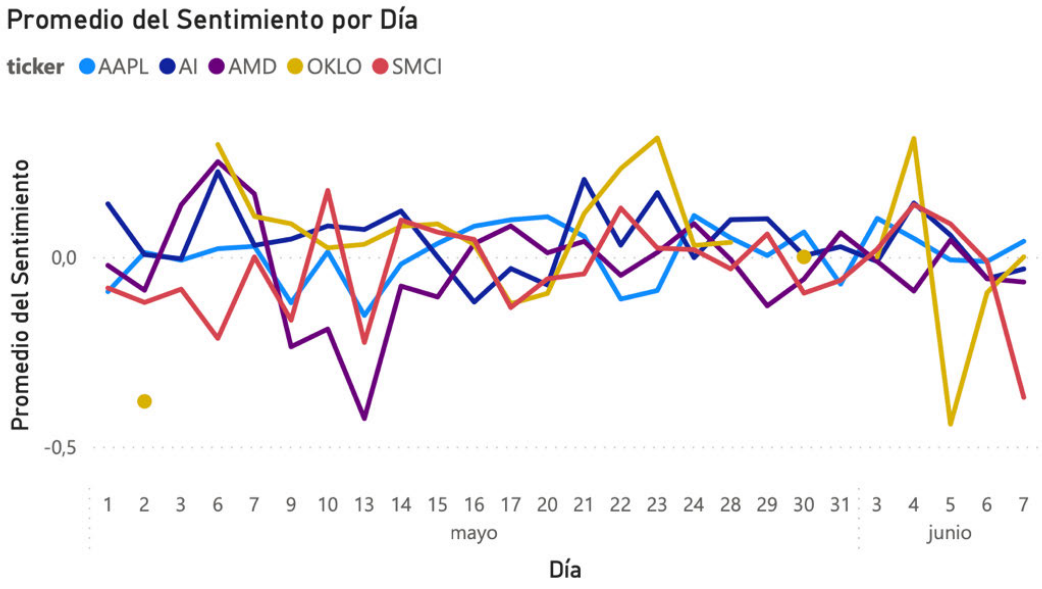

Figura 5.30 - Promedio de sentimiento de AAPL, AI, AMD, OKLO y SMCI por día. Elaboración propia.

5.5.3 Evaluación de los KPIs

Los KPIs de NVDA muestran que Nvidia es una empresa con mucho interés en la comunidad de Reddit, como se ha podido ver en el gran volumen de comentarios en los que se menciona este ticker. El sentimiento promedio es muy poco variante, lo que se puede deber a eventos específicos o noticias. La correlación entre el sentimiento de los usuarios y el precio de cierre

de las acciones sugiere que ambos están ligados, es decir, que un sentimiento negativo puede llevar a caídas en el precio y uno positivo puede provocar subidas. Además, el gran número de noticias y su sentimiento, principalmente alcista, ayuda a mantener el interés en esta organización.

En cuanto a SPY, un indicador bursátil que también es muy mencionado en la comunidad analizada de Reddit. Como se ha podido apreciar, el precio de cierre no representa grandes fluctuaciones, lo que indica su estabilidad en el mercado. En este caso, la correlación entre el precio de cierre y el sentimiento es menor que el visto en NVDA, lo que indica que hay otros factores que puedan influir en su comportamiento en bolsa.

Continuando con AMC, también hay un gran interés entre los usuarios de la comunidad por esta organización. El precio de cierre muestra que es muy cambiante, indicando que es de naturaleza especulativa. Al igual que con NVDA, la correlación entre el precio de cierre y el sentimiento promedio es significativa. El número de noticias y su sentimiento, puede indicar el comportamiento de los usuarios.

En el caso de BB, se puede apreciar que despierta menos interés que el resto de las organizaciones mencionadas. A pesar de su precio de cierre alcista, el resto de KPIs indican una débil correlación y un sentimiento neutro con respecto a la organización.

TSLA, al contrario que BB, recibe un gran número de cometarios que reflejan el constante interés en la empresa. El precio de cierre es muy volátil y lo que hace que el sentimiento varie con frecuencia. Tiene una fuerte correlación entre el sentimiento y el precio de cierre de las acciones, por lo tanto, se pueden sacar las mismas conclusiones que con NVDA y AMC.

Por último, del resto de empresas analizadas, a través de los KPIs, solo destacan AMD y AI. En ellas se observa que, en la comparación de precio y sentimiento, el cambio de sentimiento de los usuarios puede provocar cambios de precios.

Como conclusión, **el análisis de estos KPIs ofrece una compresión de cómo los comentarios en Reddit, su sentimiento y las noticias afectan al precio de las acciones.** El sentimiento y el volumen de comentarios son indicadores clave para poder apreciar el interés por esa organización y las expectativas de los usuarios. La correlación entre el precio de cierre y el sentimiento refleja la importancia de tener que observar con constancia los comentarios de la comunidad para anticipar movimientos en el mercado. Finalmente, **el análisis de las noticias financieras ofrece más contexto que puede provocar cambios en los usuarios y en el mercado financiero**. En resumen, estos KPIs ayudan a realizar una toma de decisiones más informadas.

5.6 Evaluación del modelo

En este apartado se analizan los resultados obtenidos con relación a los objetivos de negocio, los cuales buscan medir el impacto de los comentarios de los usuarios de Reddit en el mercado financiero.

5.6.1 Revisión del proceso

Durante el análisis, se identificaron como puntos clave:

- Recolección de datos: Se observó que tanto la cantidad como la calidad de los datos es fundamental para la precisión de los KPIs.
- Análisis de sentimientos: La metodología utilizada es de vital importancia para asegurar un análisis sentimental correcto.
- Visualización de los datos: Debe de ser clara y concisa. Añadir nuevas visualizaciones mejorará la comprensión de los datos.

5.6.2 Próximas etapas de actuación

Teniendo en cuenta los puntos anteriores de evaluación del modelo y su revisión, se presentan los siguientes pasos a seguir:

- Implementación del modelo: Como se ha podido comprobar, los KPIs han resultado ser válidos y útiles, por lo tanto, se puede proceder a la implementación del modelo dentro de una organización.
- Mejora continua: Se establece un ciclo continuo de recolección de datos y su análisis para la mejora de los KPIs.
- Formación: Llevar a cabo la formación del personal para poder hacer un uso eficaz del modelo.
- Monitoreo y ajustes: Implementar un sistema de monitoreo para verificar la validez del modelo a lo largo del tiempo y realizar posibles ajustes.

5.7 Implementación del modelo

En este punto se presentan los pasos que hay que llevar a cabo para implementar el modelo en un entorno productivo, asegurando su rendimiento futuro, así como su monitoreo y la comunicación de los resultados.

5.7.1 Implementación

El proceso de implementación estará formado por su integración en aplicaciones, es decir, será incorporado en una aplicación web para que los usuarios finales tengan un acceso fácil a los KPIs y puedan hacer un análisis a tiempo real. También, se creará una API para que otros sistemas puedan conectarse al modelo y poder utilizar la información que ofrece. Todo ello en entornos finales de producción, realizando las pruebas de despliegue oportunas para verificar que todo funciona perfectamente.

5.7.2 Monitoreo

Este proceso se basa en implementar sistemas de monitoreo en tiempo real, ya sea a través de logs, utilizando otras plataformas que se integren con el entorno productivo, para poder

ver la actividad en la API, u otras técnicas. Además de detectar cambios en los datos de entrada para poder hacer ajustes rápidos en el modelo y asegurar su precisión y veracidad.

5.7.3 Comunicación de los resultados

Esta etapa consiste en la preparación de informes detallados y presentaciones, para comunicar los resultados a las personas interesadas. Es muy importante la utilización de paneles interactivos que proporcionen una visión a tiempo real de los resultados, algo que Power BI ofrece, por ejemplo.

5.8 Conclusiones de la validación

La validación del modelo propuesto, llevada a cabo en este PFG, ha permitido evaluar con eficiencia y eficacia el sistema de vigilancia e inteligencia en un caso práctico. La metodología CRISP-DM ha demostrado ser un marco eficiente y estructurado para guiar todas las etapas que forman el sistema de vigilancia e inteligencia. Cada una de las fases ha sido crucial para la transformación de los datos en información útil.

La fase de la recolección de datos mediante técnicas de web scraping ha conseguido obtener una gran cantidad de datos de la comunidad WallStreetBets de Reddit. El procesado de estos datos, mediante técnicas de limpieza y normalización, ha sido esencial para garantizar la calidad de ellos y del análisis realizado en la siguiente fase. En el análisis se han identificado correlaciones entre los movimientos del mercado y los comentarios de Reddit. El uso de herramientas como PostgreSQL y Power BI, sobre todo, han facilitado el análisis ayudando a identificar indicadores clave para la toma de decisiones.

Los resultados obtenidos han demostrado que el sistema de vigilancia e inteligencia propuesto puede proporcionar información de calidad, que influye directamente en las tomas de decisiones de las organizaciones. Además, el modelo validado es flexible y se puede adaptar a empresas de diferentes tamaños, lo que permite que todas se puedan beneficiar de la información obtenida con este sistema de vigilancia e inteligencia.

Durante el proceso de validación, se han identificado diferentes mejoras que pueden hacer que el sistema sea más útil y preciso. En resumen, la validación llevada a cabo ha confirmado que el sistema de vigilancia e inteligencia cumple con los objetivos planteados y es capaz de transformar un gran volumen de datos inicial en información valiosa para la toma de decisiones. Por lo tanto, se puede decir que el modelo propuesto ha quedado validado, demostrando su eficacia y aplicación en entornos reales.

Capítulo 6. Presupuesto y planificación

6.1 Introducción

En este capítulo se presentan los costes asociados a la realización del proyecto, entre los cuales se distinguen costes de software, hardware y de recursos humanos.

6.2 Costes de hardware

Para la realización de este proyecto ha sido necesario el uso de un ordenador portátil gaming MSI con un procesador Intel Core i7-11800H, 32GB de RAM y una memoria SSD de 1TB, desde el que se han realizado todas las tareas de programación, obtención de KPIs y redacción de esta memoria. El precio se puede observar en Tabla 6.1.

6.3 Costes de software

Para este proyecto el único gasto de software que se ha tenido es el pago de Microsoft Office, para la realización de la memoria. El resto de los programas utilizados son gratuitos. El coste asociado al software se puede apreciar en la Tabla 6.1.

6.4 Recursos humanos

Para la realización de este proyecto, se ha necesitado la implicación de varios profesionales, entre los cuales se encuentran:

- Analista de datos: Es el encargado de analizar los datos utilizados en el proyecto y garantizar que son de calidad para obtener los objetivos del sistema de vigilancia e inteligencia.
- Gestor técnico: Es la persona encargada del diseño y desarrollo del sistema de vigilancia e inteligencia.
- Director: Se trata de la persona encarga de que todo el personal del proyecto trabaje adecuadamente y se cumplan los objetivos que el cliente quiere conseguir.
- Gestor de fuentes: Encargado de la búsqueda de fuentes de datos y de su mantenimiento.

El desglose del coste de estos profesionales se puede apreciar en la Tabla 6.1.

6.5 Desglose de costes

A continuación, se presenta el desglose total de los costes del proyecto en formato tabla [83] [84]:

Tabla 6.1 – Presupuesto del proyecto. Elaboración propia.

Categoría	Concepto	Coste por unidad	Unidades	Coste total
Personal	Analista de datos	18 €/hora	75	1.350 €
	Gestor técnico	22 €/hora	30	660 €
	Director	25 €/hora	60	1.500 €
	Gestor de fuentes	17 €/hora	35	595 €
Material Hardware	Portátil Gaming MSI GS66 Stealth 11UE-076XES	1.899 €	1	1.899 €
Material Software	Microsoft Office 365 Personal	69 €	1	69 €
		Total		6.073 €

6.6 Planificación y alcance del proyecto

Las tareas que se han realizado para llevar a cabo el sistema de vigilancia e inteligencia presentado en este PFG son las siguientes:

- Conocer la metodología CRISP-DM, estudiando diferentes aplicaciones para poder realizar la óptima para nuestro caso práctico.
- Buscar herramientas para desarrollar el sistema y hacer pruebas para seleccionar las herramientas que se van a utilizar.
- Diseñar el programa para la extracción de datos, analizarlos y obtener información de interés.
- Generar informes.
- Comprobar el correcto funcionamiento del sistema comparándolo con información de fuentes fiables.
- Documentar de manera escrita la memoria del PFG.

En cuanto al alcance del proyecto, se ha quedado fuera la implementación del sistema en un entorno productivo real

A continuación, se muestran las tareas mencionadas en la Tabla 6.2 y su planificación en el tiempo en un gráfico de Gantt en la Figura 6.1:

Tabla 6.2 - Planificación del proyecto. Elaboración propia.

FECHA DE COMIENZO	ACTIVIDAD	DESCRIPCIÓN	DURACIÓN DÍAS
20/9/23	ESTUDIO	Conocer la metodlogía CRISP-DM	28
18/10/23	BÚSQUEDA DE HERRAMIENTAS Y FAMILARIZACIÓN CON LAS SELECCIONADAS	Búsqueda de herramientas para desarrollar el sistema y hacer pruebas para seleccionar las herramientas que se van a utilizar.	14
1/11/23	DESARROLLO DEL SISTEMA	interés.	42
13/12/23	IMPLEMENTACIÓN DEL SISTEMA DE VIGILANCIA E INTELIGENCIA	Generar informes.	21
3/1/24	VALIDACIÓN	Comprobar el correcto funcionamiento del sistema de vigilancia e inteligencia a través de pruebas.	7
10/1/24	DOCUMENTACIÓN DEL TRABAJO DESARROLLADO	Documentar de manera escrita la memoria del PFG.	28

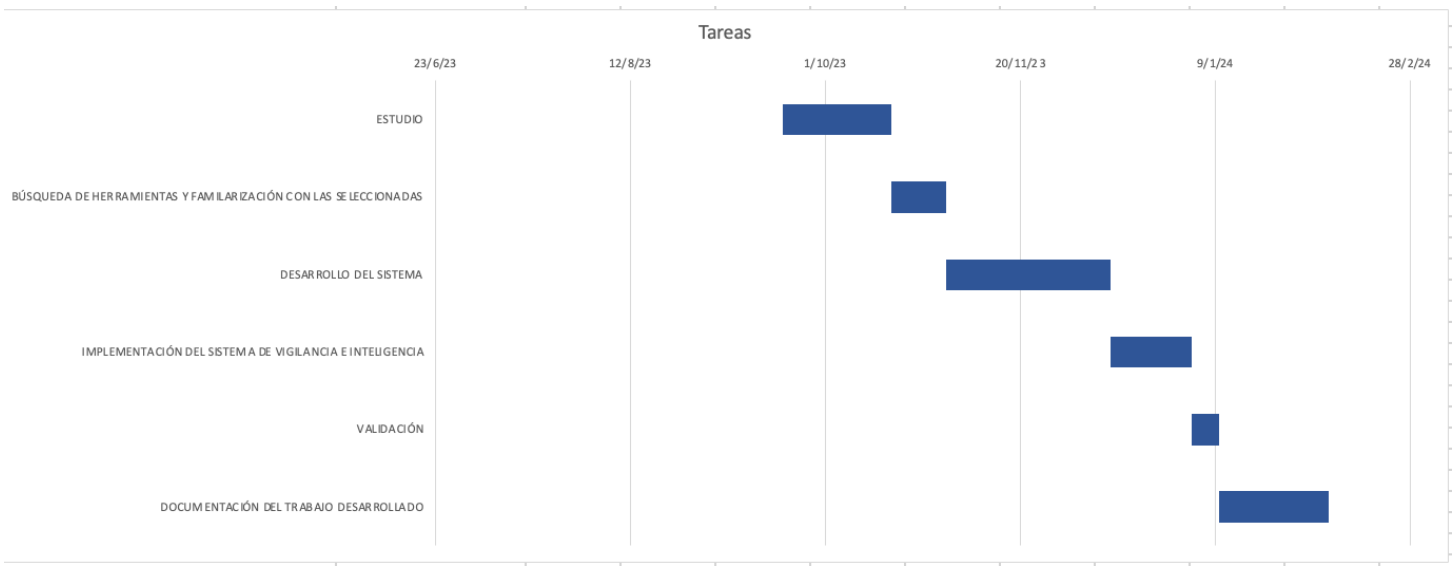

Figura 6.1 - Planificación temporal. Elaboración propia.

Capítulo 7. Impacto del proyecto

7.1 Introducción

Este capítulo tiene como objetivo exponer las implicaciones ambientales, sociales, tecnológicas y económicas del proyecto, así como su contribución a los Objetivos de Desarrollo Sostenible (ODS) [85].

7.2 Implicaciones

A continuación, se presentan las diferentes implicaciones que puede tener el proyecto:

- **Ambientales**: En los proyectos de explotación de datos en los que se implementan la metodología CRISP-DM, generalmente se fomenta el uso de infraestructuras tecnológicas eficientes, lo que puede ayudar a reducir los consumos energéticos y la huella de carbono asociada al análisis de datos [86].
- **Sociales**: El análisis de la influencia de las redes sociales, más en concreto de Reddit, en el mundo financiero tiene un gran impacto en la sociedad. Como se ha podido observar, ha permitido entender cómo las opiniones de los usuarios pueden afectar en la toma de decisiones.
- **Tecnológicas**: Al implementar la metodología CRISP-DM, se ha utilizado una estructura de trabajo que facilita la implementación de proyectos de explotación de datos, dando lugar a una mejor forma de tratar grandes volúmenes de datos [87].
- **Económicas**: Es uno de los más destacados, pues mejora la eficiencia del mercado financiero. Al comprender la influencia de las redes sociales en el mundo financiero, las organizaciones o inversores individuales pueden tomar decisiones con más información, lo que provoca más oportunidades y menos riesgos.

7.3 Contribución a los Objetivos de Desarrollo Sostenible (ODS)

El proyecto contribuye a varios de objetivos de desarrollo sostenible definidos por las Naciones Unidad, entre ellos se pueden encontrar [85]:

- Industria, Innovación e infraestructura (ODS 9): Al mejorar las infraestructuras de análisis de datos y el promover la innovación tecnológica.
- Producción y Consumo Responsables (ODS 12): Debido a la optimización de los recursos y al reducir el consumo energético por el uso de infraestructuras sostenibles.
- Trabajo Decente y Crecimiento Económico (ODS 8): Al fomentar el crecimiento económico.

Capítulo 8. Conclusiones y desarrollo de trabajos futuros

8.1 Conclusiones

En la actualidad, se manejan enormes volúmenes de datos, y el adecuado tratamiento y análisis de estos datos proporcionan ventajas a las organizaciones al tomar decisiones anticipadas y detectar cambios en el entorno, lo que les otorga una mejora competitiva.

Como se ha observado, la CRISP-DM propone una metodología por fases para el tratamiento de estos datos, presentando todas las ventajas que conlleva el uso de estas técnicas. Cada vez más países y organizaciones están adoptando estas metodologías como una parte fundamental y necesaria para posicionarse de manera ventajosa en ciencia, innovación y desarrollo. En base a ello, surge el sistema de vigilancia e inteligencia presentado en este trabajo de grado para cubrir todas las fases de CRISP-DM, ordenar y seleccionar la información accesible y obtener resultados de calidad.

La implementación de este sistema ha logrado unificar todas las herramientas utilizadas mediante diversos conectores, creando una arquitectura robusta capaz de manejar grandes volúmenes de datos y analizarlos en tiempos óptimos. Se han utilizado lenguajes de programación como Python y sus módulos para realizar transformaciones en los datos y en las tablas, los cuales han modelado y estructurado la información para agilizar el proceso de consulta en las fases de análisis e inteligencia estratégica. En esta memoria se ha presentado el ciclo completo de vida de un dato, desde su obtención hasta su máximo aprovechamiento.

El desarrollo de este sistema ha sido de gran utilidad para ampliar los conocimientos sobre las metodologías para proyectos de explotación de datos, así como las herramientas existentes para implementar estos sistemas. Gracias al estudio exhaustivo llevado a cabo para realizar el proyecto, se ha mejorado considerablemente la capacidad técnica y analítica en la manipulación de datos y el modelado de la información.

Trabajar con una de las metodologías más utilizadas ha servido para aprender a interpretar este tipo de documentos, tan importantes al diseñar e implementar proyectos, garantizando la calidad necesaria en el ámbito de la ingeniería. Además, el conocimiento adquirido de manejo de datos será de gran valor para futuros trabajos e investigaciones.

El funcionamiento del sistema y sus fases se ha validado y explicado mediante un caso práctico utilizando como base los comentarios escritos por los usuarios de la red social Reddit, realizando un análisis exhaustivo de toda la información obtenida. Se eligió este tipo de información debido a su globalidad y gran uso por parte de la Sociedad, donde se comentan multitud de temas y en incluido el que hemos tratado en este proyecto.

En cuanto a las herramientas utilizadas para implementar sistemas, existe una amplia variedad de opciones, como se ha visto en esta memoria. Se ha considerado el uso de software de código abierto, lo cual se ha logrado.

8.2 Trabajos futuros

En el contexto del proyecto que se ha desarrollado se pueden realizar varias mejoras o ampliaciones. Una de ellas es la integración con otras redes sociales, es decir, utilizar datos provenientes de Twitter o Facebook, lo que permitiría un análisis más exhaustivo. También se puede mejorar las técnicas de análisis sentimental, utilizando algoritmos más avanzados que podrían mejorar el detalle y la precisión del análisis de los comentarios.

Otra de las posibles mejoras es la de añadir modelos predictivos para predecir los movimientos del mercado financiero basándose en los datos extraídos de las redes sociales. También se puede mejorar la recolección de datos para que se recojan automáticamente cada cierto tiempo, permitiendo un análisis en tiempo real.

Por último, como se ha mencionado anteriormente, se pueden crear aplicaciones finales o APIs que permitan a los usuarios finales acceder a la información.

Capítulo 9. Referencias

[1] itUser, «La cantidad de datos que manejan las empresas crecerá rápidamente hasta 2025,» 9 Septiembre 2020. [En línea]. Available: https://almacenamientoit.ituser.es/noticias-y-actualidad/2020/09/la-cantidad-de-datos-que-manejan-las-empresas-crecera-rapidamente-hasta-2025. [Último acceso: 30 Junio 2024].

[2] Atico34, «Datos externos: información clave para la empresa,» [En línea]. Available: https://protecciondatos-lopd.com/empresas/datos-externos/. [Último acceso: 30 Junio 2024].

[3] Atico34, «Datos internos en la empresa. ¿Qué son y para qué se usan?,» [En línea]. Available: https://protecciondatos-lopd.com/empresas/datos-internos/. [Último acceso: 30 Junio 2024].

[4] M. B. Pacheco, «Fuentes y manejo de la información en las Organizaciones,» [En línea]. Available: https://www.gestiopolis.com/fuentes-y-manejo-de-la-informacion-en-las-organizaciones/. [Último acceso: 30 Junio 2024].

[5] O. Fuentes, «Organización inteligente: información, aprendizaje y conocimiento,» [En línea]. Available: https://www.gestiopolis.com/organizacion-inteligente-informacion-aprendizaje-y-conocimiento/. [Último acceso: 30 Junio 2024].

[6] T. M. M., «La economía del dato puede salvar la brecha regional del negocio empresarial en España,» 4 Julio 2024. [En línea]. Available: https://www.elconfidencial.com/espana/comunidad-valenciana/2024-07-04/big-data-tecnologia-empresas-valencia-bra_3916363/. [Último acceso: 5 Julio 2024].

[7] Lucidchart, «Método híbrido Agile-Waterfall: ¿Es adecuado para tu equipo?,» [En línea]. Available: https://www.lucidchart.com/blog/es/hibrido-metodologia-agile-waterfall-para-tu-equipo. [Último acceso: 9 Julio 2024].

[8] D. Bongardt, «Data volume expected to explode until 2025,» 9 Septiembre 2019. [En línea]. Available: https://www.hannovermesse.de/en/news/news-articles/data-volume-expected-to-explode-until-2025. [Último acceso: 13 Junio 2023].

[9] A. Woodie, «Global DataSphere to Hit 175 Zettabytes by 2025, IDC Says,» 27 Noviembre 2018. [En línea]. Available: https://www.datanami.com/2018/11/27/global-datasphere-to-hit-175-zettabytes-by-2025-idc-says/. [Último acceso: 13 Junio 2023].

[10] P. Orte, «esPublico Gestiona,» 1 Octubre 2021. [En línea]. Available: https://espublicogestiona.com/la-nueva-era-del-dato/. [Último acceso: 2022 Octubre 18].

[11] «What is Data Transformation? Importance and Best Practices,» 24 Abril 2024. [En línea]. Available: https://edgedelta.com/company/blog/what-is-data-transformation. [Último acceso: 8 Julio 2024].

[12] «computing,» 13 Agosto 2019. [En línea]. Available: https://www.computing.es/analytics/el-volumen-de-datos-en-las-empresas-crece-un-569-en-dos-anos/. [Último acceso: 24 Septiembre 2023].

[13] E. Bello, «ebschool,» 23 Osctubre 2023. [En línea]. Available: https://www.iebschool.com/blog/data-mining-mineria-datos-big-data/. [Último acceso: 10 Noviembre 2023].

[14] ucc, «ediciones.ucc.edu.co,» [En línea]. Available: https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230?inline=1. [Último acceso: 11 Noviembre 2023].

[15] A. S. H. Juan Miguel Moine, «Universidad Tecnológica Nacional, Universidad Nacional de Buenos Aires,» [En línea]. Available: https://core.ac.uk/download/pdf/296383792.pdf. [Último acceso: 12 Noviembre 2023].

[16] N. HOTZ, «What is SEMMA?,» 31 Marzo 2024. [En línea]. Available: https://www.datascience-pm.com/semma/. [Último acceso: 8 Julio 2024].

[17] d. e. i. (. Comité técnico CTN 166 Actividades de investigación, *Gestión de la I+D+i: Sistema de vigilancia e inteligencia,* Madrid, 2018.

[18] «ediciones.ucc.edu.co,» [En línea]. Available: https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230?inline=1. [Último acceso: 8 Agosto 2023].

[19] S. R. H.-A. I. C.-Z. S. J. H.-T. A. y. A.-. P. J. C. Timarán-Pereira, «El proceso de descubrimiento de conocimiento en bases de datos,» 2016. [En línea]. Available: https://ediciones.ucc.edu.co/index.php/ucc/catalog/download/36/40/230?inline=1. [Último acceso: 14 Marzo 2023].

[20] J. Landa, «fcojlanda,» [En línea]. Available: https://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/. [Último acceso: 8 Agosto 2023].

[21] javatpoint, «KDD- Knowledge Discovery in Databases,» [En línea]. Available: https://www.javatpoint.com/kdd-process-in-data-mining. [Último acceso: 8 Julio 2024].

[22] M. X. D. R. Claudia L. Hernández G. [En línea]. Available: http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/982/COMTEL-2009-80-

96.pdf?sequence=1#:~:text=La%20metodolog%C3%ADa%20SEMMA%20(Sample%2C%20Explore,de%20apoyo%20para%20el%20negocio.. [Último acceso: 10 Agosto 2023].

[23] J. M. Moine, Abril 2013. [En línea]. Available: https://core.ac.uk/download/pdf/16703288.pdf. [Último acceso: 10 Agosto 2023].

[24] F. C. Peralta, «Proceso de Conceptualización del Entendimiento del Negocio para Proyectos de Explotación de Información,» Octubre 2014. [En línea]. Available: https://www.researchgate.net/figure/Fases-de-la-metodologia-P-3-TQ-y-sus-componentes-7_fig3_284215308. [Último acceso: 9 Julio 2024].

[25] A. J. J. A. y. G. C. B. Edgar Corona Organiche, «Principales Metodologías en el Desarrollo de Proyectos de Minería de Datos,» [En línea]. Available: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://tecnocultura.org/index.php/Tecnocultura/article/download/9/9/27&ved=2ahUKEwjjv4DF45mHAxWvVqQEHb0VCCQQFnoECBkQAQ&usg=AOvVaw2UAbJQd59lFGy4B6Xm8McG. [Último acceso: 9 Julio 2024].

[26] F. Utrilla, «UNE,» 3 Mayo 2018. [En línea]. Available: https://revista.une.org/3/sistemas-de-vigilancia-e-inteligencia-en-la-gestion-de-la-id.html. [Último acceso: 2022 Octubre 25].

[27] UFV, «¿Cuáles son las 7 V del Big Data?,» 15 Agosto 2023. [En línea]. Available: https://www.ufv.es/cuales-son-las-7-v-del-big-data-preguntas-grados/#:~:text=Las%207%20V%20son%3A%20Volumen,ser%20procesados%20con%20herramientas%20convencionales.. [Último acceso: 9 Julio 2024].

[28] Geekflare, «7 Best News Scraper Tools and APIs for Data Collection,» 24 Enero 2024. [En línea]. Available: https://geekflare.com/news-scraper-tools-apis/. [Último acceso: 25 Junio 2024].

[29] R. Nead, «Top 7 Web Scraping Tools for Data Extraction,» 5 Mayo 2023. [En línea]. Available: https://dev.co/web-scrapers. [Último acceso: 25 Junio 2024].

[30] X. Chu, «Data Cleaning,» 1 Enero 2019. [En línea]. Available: https://link.springer.com/referenceworkentry/10.1007/978-3-319-77525-8_3. [Último acceso: 25 Junio 2024].

[31] datacamp, «A List of The 20 Best ETL Tools And Why To Choose Them,» Junio 2024. [En línea]. Available: https://www.datacamp.com/blog/a-list-of-the-16-best-etl-tools-and-why-to-choose-them. [Último acceso: 25 Junio 2024].

[32] L. Rámirez, «IEBS,» 19 Abirl 2022. [En línea]. Available: https://www.iebschool.com/blog/mejores-herramientas-big-data/. [Último acceso: 23 Agosto 2023].

[33] M. Duò, «Kinsta,» 20 Agosto 2023. [En línea]. Available: https://kinsta.com/es/blog/herramientas-de-visualizacion-de-datos/. [Último acceso: 28 Agosto 2023].

[34] Oracle, «What is a Relational Database (RDBMS)?,» [En línea]. Available: https://www.oracle.com/database/what-is-a-relational-database/#:~:text=In%20a%20relational%20database%2C%20each,the%20relationships%20among%20data%20points.. [Último acceso: 25 Junio 2024].

[35] MongoDB, «What is NoSQL?,» [En línea]. Available: https://www.mongodb.com/resources/basics/databases/nosql-explained. [Último acceso: 25 Junio 2024].

[36] IBM, «What is OLAP (online analytical processing)?,» [En línea]. Available: https://www.ibm.com/topics/olap. [Último acceso: 25 Junio 2024].

[37] Gartner Inc., «Positioning technology players within a specific market,» [En línea]. Available: https://www.gartner.com/en/research/methodologies/magic-quadrants-research. [Último acceso: 9 Julio 2024].

[38] Gartner, «Microsoft Power BI, líder del Cuadrante Mágico de Gartner 2023,» Enero 2023. [En línea]. Available: https://blog.bismart.com/microsoft-power-bi-lider-cuadrante-magico-gartner-2023. [Último acceso: 26 Junio 2024].

[39] S. López, «Ruta Financiera,» [En línea]. Available: https://rutafinanciera.es/el-impacto-de-las-redes-sociales-en-tus-finanzas-personales/. [Último acceso: 11 Abril 2024].

[40] M. J. Aznar, 30 Septiembre 2019. [En línea]. Available: https://blogs.ugr.es/almenara/5-las-fuentes-para-el-estudio-de-la-economia/. [Último acceso: 11 Abril 2024].

[41] finanboo, «www.finanboo.com,» [En línea]. Available: https://www.finanboo.com/es/blog/cuales-son-las-fuentes-de-informacion-financiera/. [Último acceso: 11 Abril 2024].

[42] «blog cuales son tus metas,» 31 Mayo 2023. [En línea]. Available: https://blog.cualessontusmetas.com/cuales-son-los-organismos-reguladores-del-sistema-financiero/. [Último acceso: 12 Abril 2024].

[43] [En línea]. Available: https://www.studyin-uk.com/spain/profiles/university/london-school-of-economics/. [Último acceso: 12 Abril 2024].

[44] A. R. Córcoles, 2022. [En línea]. Available: https://www.acta.es/medios/informes/2022003.pdf. [Último acceso: 12 Abril 2024].

[45] Finanacial Times, [En línea]. Available: https://aboutus.ft.com. [Último acceso: 12 Abril 2024].

[46] Wall Street Journal, [En línea]. Available: https://www.wsj.com/. [Último acceso: 12 Abril 2024].

[47] [En línea]. Available: https://www.uv.es/cibisoc/documentos/Yahoo-Finance_OscarCarchano.pdf. [Último acceso: 12 Abril 2024].

[48] Bloomberg, «https://www.bloomberg.com/,» [En línea]. [Último acceso: 12 Abril 2024].

[49] Alpha Vantage Inc, [En línea]. Available: https://es.linkedin.com/company/alpha-vantage-inc. [Último acceso: 12 Abril 2024].

[50] [En línea]. Available: https://elasesorfinanciero.com/las-principales-asociaciones-profesionales-en-la-industria-del-asesoramiento-financiero/. [Último acceso: 12 Abril 2024].

[51] turnitin, 27 Julio 2020. [En línea]. Available: https://latam.turnitin.com/blog/como-identificar-sitios-publican-informacion-falsa. [Último acceso: 11 Abril 2024].

[52] IBM, [En línea]. Available: https://www.ibm.com/es-es/topics/data-reliability. [Último acceso: 15 Abril 2024].

[53] LinkedIn, «LinkedIn,» [En línea]. Available: https://es.linkedin.com/advice/1/what-most-effective-data-sources-investment?lang=es#:~:text=Algunas%20de%20las%20bases%20de,pueden%20ahorrarle%20tiempo%20y%20esfuerzo.. [Último acceso: 14 Abril 2024].

[54] Á. S. &. P. Móricz, «Towards data-driven decision making: the role of analytical culture and centralization efforts,» 16 Septiembre 2023. [En línea]. Available: https://link.springer.com/article/10.1007/s11846-023-00694-1. [Último acceso: 8 Julio 2024].

[55] bismart, «Business Intelligence for Business Decision Making,» [En línea]. Available: https://blog.bismart.com/en/business-intelligence-data-driven-decisions. [Último acceso: 8 Julio 2024].

[56] T. J. M. M. N. Shimon Kogan, «Social Media and Financial News Manipulation,» 4 Julio 2023. [En línea]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3237763. [Último acceso: 25 Junio 2024].

[57] D. V. G. Alton M. K. Chew, «Social Media Big Data: The Good, The Bad, and the Ugly (Un)truths,» 1 Junio 2021. [En línea]. Available: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2021.623794/full. [Último acceso: 25 Junio 2024].

[58] E. I. N. P. R. &. R. R. Yogesh K. Dwivedi, «Social Media Adoption, Usage And Impact In Business-To-Business (B2B) Context: A State-Of-The-Art Literature Review,» 2 Febrero 2021. [En línea]. Available: https://link.springer.com/article/10.1007/s10796-021-10106-y. [Último acceso: 25 Junio 2024].

[59] rdstation, «rdstation,» [En línea]. Available: https://www.rdstation.com/es/redes-sociales/. [Último acceso: 30 Noviembre 2022].

[60] U. J. E. O. A. J. S. &. S. L. Femi Olan, «Fake news on Social Media: the Impact on Society,» 19 Enero 2022. [En línea]. Available: https://link.springer.com/article/10.1007/s10796-022-10242-z. [Último acceso: 25 Junio 2024].

[61] Á. Z. &. M. E. S. Reyes, «Pros & cons: impacts of social media on mental health,» 6 Julio 2023. [En línea]. Available: https://bmcpsychology.biomedcentral.com/articles/10.1186/s40359-023-01243-x. [Último acceso: 25 Junio 2024].

[62] R. Pomeroy, «A Scientific Analysis of the GameStop Short Squeeze,» 15 Abril 2022. [En línea]. Available: https://www.realclearscience.com/articles/2022/04/15/study_gamestop_short_squeeze.html. [Último acceso: 25 Junio 2024].

[63] S. M. Martin Smits, «The Impact Of Social Media On Business Performance,» 1 Julio 20213. [En línea]. Available: https://ciencia.ucp.pt/ws/portalfiles/portal/27211288/The_Impact_Of_Social_Media_On_Business_Performance.pdf. [Último acceso: 25 Junio 2024].

[64] V. R. &. G. D. Romas Vijeikis, «Towards Automated Surveillance: A Review of Intelligent Video Surveillance,» 6 Julio 2021. [En línea]. Available: https://link.springer.com/chapter/10.1007/978-3-030-80129-8_53. [Último acceso: 25 Junio 2024].

[65] M. D. F. M. Veronika Plotnikova, «Adaptations of data mining methodologies: a systematic literature review,» 25 Mayo 2020. [En línea]. Available: https://peerj.com/articles/cs-267/. [Último acceso: 26 Junio 2024].

[66] B. Katz, «The Intelligence Edge: Opportunities and Challenges from Emerging Technologies for U.S. Intelligence,» 17 Abril 2020. [En línea]. Available: https://www.csis.org/analysis/intelligence-edge-opportunities-and-challenges-emerging-technologies-us-intelligence. [Último acceso: 26 Junio 2024].

[67] D. T. J.-G. S. M. A. Amani Ibrahim, «The Challenges of Leveraging Threat Intelligence to Stop Data Breaches,» 28 Agosto 2020. [En línea]. Available: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2020.00036/full. [Último acceso: 26 Junio 2024].

[68] K. B. M. M. G. S. M. S. M. Asma Dhaouadi, «Data Warehousing Process Modeling from Classical Approaches to New Trends: Main Features and Comparisons,» 12 Agosto 2022. [En línea]. Available: https://www.mdpi.com/2306-5729/7/8/113. [Último acceso: 26 Junio 2024].

[69] profile, «profile,» 22 Marzo 2021. [En línea]. Available: https://profile.es/blog/pandas-python/. [Último acceso: 24 Enero 2023].

[70] EDB, «7 Best Practice Tips for PostgreSQL Bulk Data Loading,» 19 Febrero 2023. [En línea]. Available: https://www.enterprisedb.com/blog/7-best-practice-tips-postgresql-bulk-data-loading. [Último acceso: 29 Junio 2024].

[71] E. Bello, «iebschool,» 4 febrero 2022. [En línea]. Available: https://www.iebschool.com/blog/microsoft-power-bi-analitica-usabilidad/. [Último acceso: 20 febrero 2023].

[72] deloitte, «deloitte,» [En línea]. Available: https://www2.deloitte.com/es/es/pages/technology/articles/que-es-power-bi.html. [Último acceso: 2023 febrero 19].

[73] F. V. T. M. K. N. S. R. W. &. E. M. Maryam M Najafabadi, «Deep learning applications and challenges in big data analytics,» 24 Febrero 2015. [En línea]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-014-0007-7. [Último acceso: 25 Junio 2024].

[74] P. M. R. B. &. A. G. T. Ramalingeswara Rao, «The big data system, components, tools, and technologies: a survey,» 18 Septiembre 2018. [En línea]. Available: https://link.springer.com/article/10.1007/s10115-018-1248-0. [Último acceso: 25 Junio 2024].

[75] estrategiastrading, «estrategiastrading.com,» [En línea]. Available: https://estrategiastrading.com/datos-financieros-para-python-alpha-vantage. [Último acceso: 8 Marzo 2024].

[76] fastercapital, «https://fastercapital.com,» 1 Febero 2024. [En línea]. Available: https://fastercapital.com/es/contenido/WallStreetBets--como-una-comunidad-de-Reddit-convirtio-a-GameStop-en-una-accion-de-culto.html. [Último acceso: 5 Marzo 2024].

[77] yahoo, «Yahoo Finance,» [En línea]. Available: https://es.finance.yahoo.com/. [Último acceso: 10 Marzo 2024].

[78] BolsaZone, «BolsaZone,» [En línea]. Available: https://bolsazone.com/news/yahoo-finance-portal-informacion-financiera-gratuito/). . [Último acceso: 13 Marzo 2024].

[79] M. Suarez, «datosmaestros.com,» 10 Abril 2023. [En línea]. Available: https://datosmaestros.com/tecnicas-de-limpieza-de-datos/. [Último acceso: 3 Junio 2024].

[80] M. K. Barai, «Sentiment Analysis with TextBlob and Vader,» 21 Febrero 2024. [En línea]. Available: https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/. [Último acceso: 27 Junio 2024].

[81] E. A. &. K. D. Lahtinen, «Social media sentiment and market behavior,» 26 Mayo 2018. [En línea]. Available: https://link.springer.com/article/10.1007/s00181-018-1430-y. [Último acceso: 26 Junio 2024].

[82] H. X. Kaifeng Guo, «Deep learning in finance assessing twitter sentiment impact and prediction on stocks,» 24 Mayo 2024. [En línea]. Available: https://peerj.com/articles/cs-2018/. [Último acceso: 26 Junio 2024].

[83] Worten, [En línea]. Available: https://www.worten.es/productos/portatil-gaming-msi-gs66-stealth-11ue-076xes-intel-core-i7-11800h-nvidia-geforce-rtx-3060-ram-32-gb-1-tb-ssd-15-6-7413696. [Último acceso: 6 Mayo 2023].

[84] Microsoft, [En línea]. Available: https://www.microsoft.com/es-ES/microsoft-365/buy/compare-all-microsoft-365-products. [Último acceso: 6 Mayo 2023].

[85] Naciones Unidas, «www.un.org,» [En línea]. Available: https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/. [Último acceso: 20 Junio 2024].

[86] J. Roundy, «techtarget,» 7 Junio 2023. [En línea]. Available: https://www.techtarget.com/searchdatacenter/tip/Strategies-to-work-toward-data-center-decarbonization. [Último acceso: 20 Junio 2024].

[87] iic.uam, «iic.uam,» [En línea]. Available: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/. [Último acceso: 20 Junio 2024].

Manual de usuario

En este manual de usuario se exponen todos los pasos que hay que seguir para la obtención de los datos y la generación de los KPIs presentados en el punto 5.5.2. Para el primer apartado, se genera una guía para la ejecución de los programas en un entorno de Python y para el segundo, se define una guía en el que se detalla paso a paso la creación de cada uno de los gráficos presentados.

A.1 Instalación del entorno

En primer lugar, para llegar a tener los datos necesarios para realizar el caso práctico presentado, hay que llevar a cabo la instalación del entorno. Como se ha mencionado, se ha utiliza Python, por tanto, hay que realizar los siguientes pasos:

- **Instalación de Python**: Ir a su página web (https://www.python.org) y seguir los pasos de instalación.
- **Verificar la instalación**: Abrir el terminal y escribir el comando: '*python –version',* tendrá que aparecer la versión previamente instalada.
- **Instalación de las librerías necesarias:** Se utilizará '*pip'* para la instalación de las librerías, se trata de un programa de instalación que ya viene por defecto en Python. Si no está instalado habría que ejecutar el comando '*python get-pip.py*' en el terminal, en la ruta donde se encuentra instalado Python. Por último, se instalarán las librerías con el comando '*pip install pandas vaderSentiment playwright datetime*'.

A.2 Obtención de los datos

Se deben ejecutar en el siguiente orden para obtener con éxito los datos necesarios:

- wallstreetbets_scraper.py: En cualquier entorno para Python o con el comando '*python wallstreetbets_scraper.py'*
- tickers_sentimental_analysis.py: En cualquier entorno para Python o con el comando '*python tickers_sentimental_analysis.py'*
- getAlpha&YahooInfo.py: En cualquier entorno para Python o con el comando '*python getAlpha&YahooInfo.py'*
- limpieza_datos.py: En cualquier entorno para Python o con el comando '*python limpieza_datos.py*'

  A continuación, se presentan el código de los programas:

A.2.1. wallstreetbets_scraper.py

Es el encargado de hacer el proceso de web scraping del subgrupo WallStreetBets de Reddit.

```
from playwright.sync_api import sync_playwright
import time
import pandas as pd
```

```
from datetime import datetime

# Indicar el mes y el año para obtener sus datos
month = "june"
year = "2024"

if(month == "january" or month == "march" or month == "may" or month == "july" or
month == "august" or month == "october" or month == "december"):
    daysMonth = 31
else:
    daysMonth = 30
# Lista de los links de cada día encontrados
daily_links = {}

# Inicializa Playwright
with sync_playwright() as p:
    browser = p.firefox.launch()

    # Hace la búsqueda de los 'Daily Discussion Thread' para el mes y el año
especificados
    page = browser.new_page()

page.goto(f"https://www.reddit.com/r/wallstreetbets/search/?q=daily+discussion+{mo
nth}+{year}")

    # Espera para que la página cargue
    time.sleep(5)

    # Lista de las URLs obtenidas antes de empezar a hacer scroll
    prev_urls = []

    # Encontrar y almacener las URLs que tengan el año y mes especificados
    while True:
        # Scroll hasta abajo para obtener más resultados
        page.evaluate("window.scrollTo(0, document.body.scrollHeight);")
        time.sleep(2)

        # Obtener todas las URLs en la página
        links = page.query_selector_all("//a[contains(@href, '/comments/')]")
        urls = [link.get_attribute("href") for link in links]
        if urls == prev_urls:
            for day in range(1, daysMonth + 1):
                found_day = False
                for url in urls:
                    if f"daily_discussion_thread_for_{month}_{day}_{year}" in url:
                        found_day = True
                        daily_links.setdefault(day, []).append(url)
                if not found_day:
                    continue  # Si no se encuentra el día, seguimos con el
siguiente
            break

        prev_urls = urls
```

```
    # Función para procesar los comentarios
    def process_comments(comments, day):
        print("Found " + str(len(comments)) + " comments")
        data = []
        date_obj = datetime(2024, 6, day)
        formatted_date = date_obj.strftime("%d/%m/%Y")

        for comment in comments:
            data.append({"Date": formatted_date, "Comment": comment})

        df = pd.DataFrame(data)
        df.to_csv(f"{month}_{day}_{year}_reddit_comments.csv", index=False)


    # Función para obtener los datos de cada día, se buscan comentarios que no se
hayan procesado con anterioridad y se sigue cargando la página hasta que no se
pueda más
    def process_day(day_links):
        comments = set()
        processed_links = set()
        for link in day_links:
            if link in processed_links:
                break
            processed_links.add(link)

            page.goto("https://www.reddit.com" + link + "?sort=top")

            page.evaluate("window.scrollTo(0, document.body.scrollHeight);")
            time.sleep(5)

            while True:
                try:
                    page.evaluate("document.querySelector('button.button-
small.button-brand').click()")
                    time.sleep(3)
                except:
                    print('Button not found. End of the page')
                    break


            comment_elements = page.query_selector_all("//shreddit-comment-
tree[@id='comment-tree']//div[@id='-post-rtjson-content']")


            for element in comment_elements:
                element_text = str(element.inner_text().replace('\n', ' '))
                if element_text not in comments:
                    comments.add(element_text)
        return comments

    # Procesar los comentarios para cada día
    for day, links in daily_links.items():
        print(f"Procesando comentarios de {month} {day} {year}:")
```

```
        comments = process_day(links)
        process_comments(comments, day)

    browser.close()
```

Como se puede observar en el código, se obtiene un conjunto de datos por cada día analizado, se guardará cada conjunto en la misma carpeta.

A.2.2. tickers_sentimental_analysis.py

Se encarga de procesar los datos obtenidos del programa anterior y asignar los indicadores de bolsa, si lo hubiera, y de analizar sentimentalmente cada comentario.

```
import os
import re
import pandas as pd
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer

vader_analyzer = SentimentIntensityAnalyzer()
ticker_pattern = re.compile(r'\b[A-Z]{2,5}\b|\$\b[A-Za-z]{1,5}\b')

def extract_tickers(text):
    matches = ticker_pattern.findall(text)
    tickers = set()

    for match in matches:
        ticker = match[1:].upper() if match.startswith('$') else match.upper()
    return list(tickers)

def analyze_sentiment(text):
    scores = vader_analyzer.polarity_scores(text)
    sentiment = "positive" if scores['compound'] >= 0.05 else "negative" if
scores['compound'] <= -0.05 else "neutral"
    return sentiment, scores['compound']


def process_comments(file_path):
    df = pd.read_csv(file_path)
    df = df.dropna(subset=['Comment'
    df['Tickers'] = df['Comment'].apply(extract_tickers)
    df[['Sentiment', 'Sentiment_Score']] = df['Comment'].apply(lambda x:
pd.Series(analyze_sentiment(x)))
    return df

# Ruta donde se encuentrarn los .csv obtenidos del script anterior
input_folder = '/WallStreetBetsComments'


combined_df = pd.DataFrame()

for file_name in os.listdir(input_folder):
    if file_name.endswith('.csv'):
        file_path = os.path.join(input_folder, file_name)
        daily_df = process_comments(file_path)
```

```
        combined_df = pd.concat([combined_df, daily_df], ignore_index=True)


output_file = 'Reddit_Comments_Sentiment_Analysis.csv'
combined_df.to_csv(output_file, index=False)

print(f"Archivo guardado ")
```

A.2.3. getAlpha&YahooInfo.py

Obtiene la información de las APIs de AlphaVantage.co y Yahoo Finance.

```
import pandas as pd
import requests
import time
from datetime import datetime, timedelta
from collections import Counter
import yfinance as yf

def getTopTickers():
# Ruta al archive obtenido del programa anterior

    df = pd.read_csv('/Reddit_Comments_Sentiment_Analysis.csv')

    df_exploded = df.explode('Tickers')

    ticker_counts = df_exploded['Tickers'].value_counts()

    top_10_tickers = ticker_counts.head(11)

    top_10_df = pd.DataFrame({'Ticker': top_10_tickers.index,
'Number_of_Mentions': top_10_tickers.values})

    return top_10_df

def yahooFinance(top_10_df):
    financial_info_list = []
    top_10_df = top_10_df[top_10_df['Ticker'].apply(lambda x: x.strip() != '[]')]

        for ticker in top_10_df['Ticker']:
        try:
            ticker = ticker.replace("'", "").strip().upper()
            ticker = ticker.replace("[", "")
            ticker = ticker.replace("]", "")

            stock = yf.Ticker(ticker)

            hist = stock.history(period="3mo")
            hist['ticker'] = ticker

            financial_info_list.append(hist)
        except Exception as e:
            print(f"No se ha podido obtener información de {ticker}: {e}")
```

```
    financial_info_df = pd.concat(financial_info_list)

    return financial_info_df
def alphaNews():
    df_list = []
    year = 2024


    for month in range(1, 13):
        start_date = datetime(year, month, 1)
        if month == 12:
            end_date = datetime(year + 1, 1, 1) - timedelta(days=1)
        else:
            end_date = datetime(year, month + 1, 1) - timedelta(days=1)
        time_from = start_date.strftime('%Y%m01T0130')
        time_to = end_date.strftime('%Y%m%dT0130')

        url =
f'https://www.alphavantage.co/query?function=NEWS_SENTIMENT&apikey=FGU5UHMH1GUUAWK
W&tickers=AAPL&time_from={time_from}&time_to={time_to}&limit=1000'

        r = requests.get(url)
        data = r.json()

        if 'feed' in data:
            flat_data = pd.json_normalize(data['feed'])
            df_list.append(pd.DataFrame(flat_data))

    df = pd.concat(df_list, ignore_index=True)
    return df

if __name__ == '__main__':

    top_10_df = getTopTickers()
    if top_10_df is not None:
        output_file = 'top_10_most_mentioned_tickers.csv'
        top_10_df.to_csv(output_file, index=False)
        financial_info_df = yahooFinance(top_10_df)
        if financial_info_df is not None:
            output_file = 'yahooFinance_info_top_10.csv'
            financial_info_df.to_csv(output_file, index=True)
```

A.2.4. LimpezaDatos.py

Es le programa utilizado para la limpieza de datos.

```
import pandas as pd
import json

# Cargar los datasets
alpha_news = pd.read_csv('/alpha_news.csv')
```

```
top_10_tickers = pd.read_csv('/top_10_most_mentioned_tickers.csv')
yahoo_finance = pd.read_csv('/yahooFinance_info_top_10.csv')
reddit_comments = pd.read_csv(' /Reddit_Comments_Sentiment_Analysis.csv')

# Limpieza de datos
# Convertir fechas
alpha_news['time_published'] =
pd.to_datetime(alpha_news['time_published']).dt.date  # Extraer solo la fecha
yahoo_finance['Date'] = pd.to_datetime(yahoo_finance['Date']).dt.date  # Extraer
solo la fecha
reddit_comments['Date'] = pd.to_datetime(reddit_comments['Date'],
format='%d/%m/%Y').dt.date  # Asegurarse de que el formato de fecha es correcto

# Limpiar tickers
top_10_tickers['Ticker'] =
top_10_tickers['Ticker'].str.strip("[]").str.replace("'", "")
reddit_comments['Tickers'] = reddit_comments['Tickers'].apply(lambda x:
x.strip("[]").replace("'", "").split(','))

# Convertir columnas JSON en alpha_news
alpha_news['topics'] = alpha_news['topics'].apply(lambda x:
json.loads(x.replace("'", '"')) if pd.notnull(x) else [])
alpha_news['ticker_sentiment'] = alpha_news['ticker_sentiment'].apply(lambda x:
json.loads(x.replace("'", '"')) if pd.notnull(x) else [])

# Expandir la columna 'Tickers' en reddit_comments para que cada ticker tenga su
propia fila
reddit_comments = reddit_comments.explode('Tickers')
# Convertir la columna 'authors' a un formato compatible con PostgreSQL
alpha_news['authors'] = alpha_news['authors'].apply(lambda x: '{' +
','.join(x.strip('[]').replace("'", "").split(',')) + '}' if pd.notnull(x) else
None)
# Estandarización de tickers en top_10_tickers
top_10_tickers['Ticker'] =
top_10_tickers['Ticker'].str.strip("[]").str.replace("'", "")

# Extraer y expandir los principales tópicos
def extract_topics(topics_list, top_n=3):
    sorted_topics = sorted(topics_list, key=lambda x: float(x['relevance_score']),
reverse=True)
    return sorted_topics[:top_n]

for i in range(3):
    alpha_news[f'topic_{i+1}'] = alpha_news['topics'].apply(lambda x:
x[i]['topic'] if len(x) > i else None)
    alpha_news[f'topic_{i+1}_relevance_score'] = alpha_news['topics'].apply(lambda
x: x[i]['relevance_score'] if len(x) > i else None)

# Extraer y expandir los principales tickers
def extract_tickers(tickers_list, top_n=5):
    sorted_tickers = sorted(tickers_list, key=lambda x:
float(x['relevance_score']), reverse=True)
    return sorted_tickers[:top_n]
```

```
for i in range(5):
    alpha_news[f'ticker_{i+1}'] = alpha_news['ticker_sentiment'].apply(lambda x:
x[i]['ticker'] if len(x) > i else None)
    alpha_news[f'ticker_{i+1}_relevance_score'] =
alpha_news['ticker_sentiment'].apply(lambda x: x[i]['relevance_score'] if len(x) >
i else None)
    alpha_news[f'ticker_{i+1}_sentiment_score'] =
alpha_news['ticker_sentiment'].apply(lambda x: x[i]['ticker_sentiment_score'] if
len(x) > i else None)
    alpha_news[f'ticker_{i+1}_sentiment_label'] =
alpha_news['ticker_sentiment'].apply(lambda x: x[i]['ticker_sentiment_label'] if
len(x) > i else None)

# Eliminar las columnas originales de JSON ya no necesarias
alpha_news = alpha_news.drop(columns=['topics', 'ticker_sentiment'])

# Eliminar columnas innecesarias
alpha_news =
alpha_news.drop(columns=['category_within_source_Business','category_within_source
_Earnings','category_within_source_Economy','category_within_source_General','cate
gory_within_source_GoogleRSS','category_within_source_Investing','category_within_
source_Markets','category_within_source_Mergers','category_within_source_News','ca
tegory_within_source_RSS','category_within_source_Top
News','category_within_source_Top Stories','category_within_source_Trading'])

# Convertir la columna 'Date'
yahoo_finance['Date'] = pd.to_datetime(yahoo_finance['Date'])
reddit_comments['Date'] = pd.to_datetime(reddit_comments['Date'])

# Guardar datasets modificados
alpha_news.to_csv('alpha_news_cleaned.csv', index=False)
top_10_tickers.to_csv('top_10_tickers_cleaned.csv', index=False)
yahoo_finance.to_csv('yahoo_finance_cleaned.csv', index=False)
reddit_comments.to_csv('reddit_comments_cleaned.csv', index=False)
```

A.3 Almacenamiento de los datos

A continuación, se presentan todos los pasos que hay que seguir para almacenar los datos obtenidos en el punto anterior en PostgreSQL.

1. **Descargar e instalar PostgreSQL** desde su página web (https://www.postgresql.org/download/). En el proceso de la instalación se pedirá que se defina una contraseña para el usuario administrador, apuntar la que se escriba porque se pedirá en los siguientes puntos.
2. **Abrir el pgAdmin:** Una vez se termina la instalación, abrir este programa, se pedirá la contraseña definida en el punto anterior y la creación de un esquema para la base de datos, se puede indicar el nombre que se prefiere.
3. **Abrir Query Tool:** Es la herramienta necesaria para la ejecución de comandos SQL, hay escribir los siguientes:

```
CREATE TABLE reddit_comments (
  id SERIAL NOT NULL PRIMARY KEY,
  date DATE,
  comment TEXT,
  ticker TEXT UNIQUE,
  sentiment TEXT,
  sentiment_score DECIMAL
);

CREATE TABLE yahoo_finance (
  id SERIAL NOT NULL PRIMARY KEY,
  reddit_comment_id INT,
  date DATE,
  open DECIMAL,
  high DECIMAL,
  low DECIMAL,
  close DECIMAL,
  volume BIGINT,
  dividends DECIMAL,
  stock_splits DECIMAL,
  ticker TEXT,
  capital_gains DECIMAL,
  FOREIGN KEY (reddit_comment_id) REFERENCES reddit_comments(id),
  FOREIGN KEY (ticker) REFERENCES reddit_comments(ticker)
);

CREATE TABLE alpha_vantage (
  id SERIAL NOT NULL PRIMARY KEY,
  reddit_comment_id INT,
  title TEXT,
  url TEXT,
  time_published DATE,
  authors TEXT[],
  summary TEXT,
  banner_image TEXT,
  source TEXT,
  source_domain TEXT,
  overall_sentiment_score DECIMAL,
  overall_sentiment_label TEXT,
  topic_1 TEXT,
  topic_1_relevance_score DECIMAL,
  topic_2 TEXT,
  topic_2_relevance_score DECIMAL,
  topic_3 TEXT,
  topic_3_relevance_score DECIMAL,
  ticker_1 TEXT,
  ticker_1_relevance_score DECIMAL,
  ticker_1_sentiment_score DECIMAL,
  ticker_1_sentiment_label TEXT,
  ticker_2 TEXT,
  ticker_2_relevance_score DECIMAL,
  ticker_2_sentiment_score DECIMAL,
```

```
  ticker_2_sentiment_label TEXT,
  ticker_3 TEXT,
  ticker_3_relevance_score DECIMAL,
  ticker_3_sentiment_score DECIMAL,
  ticker_3_sentiment_label TEXT,
  ticker_4 TEXT,
  ticker_4_relevance_score DECIMAL,
  ticker_4_sentiment_score DECIMAL,
  ticker_4_sentiment_label TEXT,
  ticker_5 TEXT,
  ticker_5_relevance_score DECIMAL,
  ticker_5_sentiment_score DECIMAL,
  ticker_5_sentiment_label TEXT,
  FOREIGN KEY (reddit_comment_id) REFERENCES reddit_comments(id),
  FOREIGN KEY (ticker_1) REFERENCES reddit_comments(ticker),
  FOREIGN KEY (ticker_2) REFERENCES reddit_comments(ticker),
  FOREIGN KEY (ticker_3) REFERENCES reddit_comments(ticker),
  FOREIGN KEY (ticker_4) REFERENCES reddit_comments(ticker),
  FOREIGN KEY (ticker_5) REFERENCES reddit_comments(ticker)
);

CREATE TABLE top_10_tickers (
  id SERIAL NOT NULL PRIMARY KEY,
  reddit_comment_id INT,
  ticker TEXT,
  number_of_mentions INT,
  FOREIGN KEY (reddit_comment_id) REFERENCES reddit_comments(id),
  FOREIGN KEY (ticker) REFERENCES reddit_comments(ticker)
);
```

En la siguiente figura se puede observar la base de datos tras la ejecución de los comandos de creación de tablas en la herramienta, para ello solo hay que copiar este código SQL presentado y darle al botón indicado.

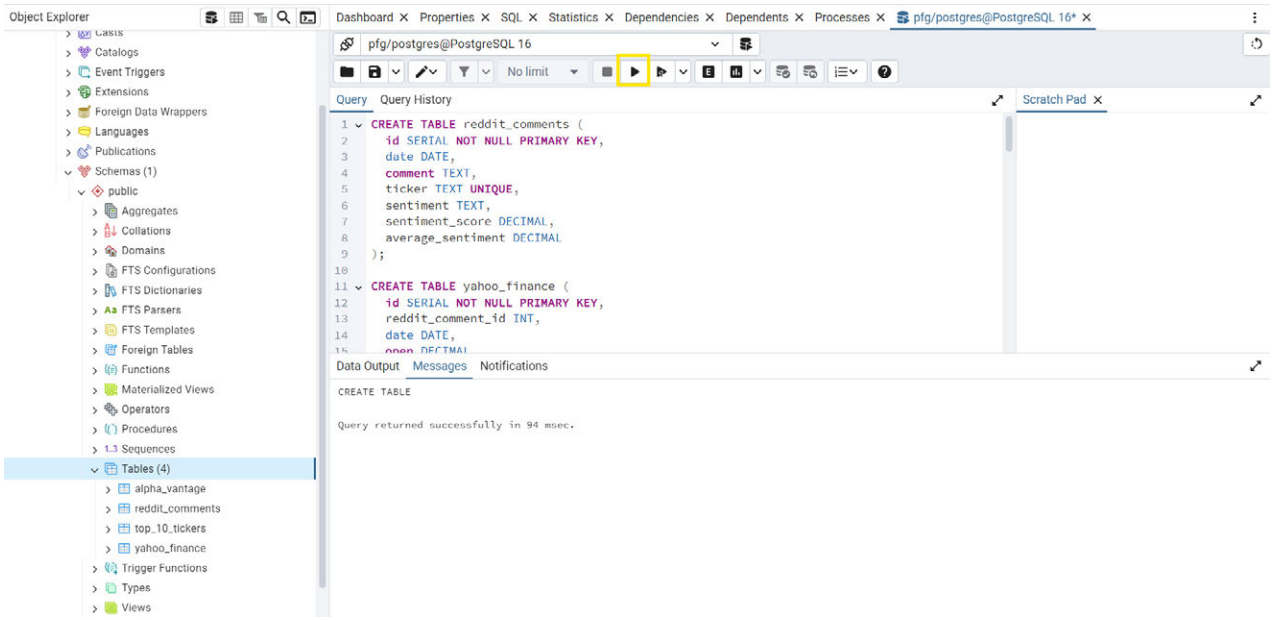

Figura 9.1 - Ejecución de los comandos en Query Tool. Elaboración propia

Por último, se presentan los comandos para añadir todos los datos, que se encuentran en los datasets obtenidos del punto anterior, a las tablas recién creadas:

- Tabla reddit_comments:

```
COPY reddit_comments(date, comment, ticker, sentiment, sentiment_score)
FROM ' \reddit_comments_cleaned.csv'
DELIMITER ','
CSV HEADER;
```

- Tabla yahoo_finance:

```
CREATE TEMP TABLE yahoo_finance_temp (
  date DATE,
  open DECIMAL,
  high DECIMAL,
  low DECIMAL,
  close DECIMAL,
  volume BIGINT,
  dividends DECIMAL,
  stock_splits DECIMAL,
  ticker TEXT,
  capital_gains DECIMAL
);
COPY yahoo_finance_temp(date, open, high, low, close, volume, dividends,
stock_splits, ticker, capital_gains)
FROM '\yahoo_finance.csv'
DELIMITER ','
CSV HEADER;
INSERT INTO yahoo_finance (reddit_comment_id, date, open, high, low, close, volume,
dividends, stock_splits, ticker, capital_gains)
SELECT rc.id, yft.date, yft.open, yft.high, yft.low, yft.close, yft.volume,
yft.dividends, yft.stock_splits, yft.ticker, yft.capital_gains
FROM yahoo_finance_temp yft
JOIN reddit_comments rc ON rc.ticker = yft.ticker;
```

- Tabla alpha_vantage:

```
CREATE TEMP TABLE alpha_vantage_temp (
  title TEXT,
  url TEXT,
  time_published DATE,
  authors TEXT,
  summary TEXT,
  banner_image TEXT,
  source TEXT,
  category_within_source TEXT,
  source_domain TEXT,
  overall_sentiment_score DECIMAL,
  overall_sentiment_label TEXT,
  topic_1 TEXT,
  topic_1_relevance_score DECIMAL,
  topic_2 TEXT,
  topic_2_relevance_score DECIMAL,
```

```
  topic_3 TEXT,
  topic_3_relevance_score DECIMAL,
  ticker_1 TEXT,
  ticker_1_relevance_score DECIMAL,
  ticker_1_sentiment_score DECIMAL,
  ticker_1_sentiment_label TEXT,
  ticker_2 TEXT,
  ticker_2_relevance_score DECIMAL,
  ticker_2_sentiment_score DECIMAL,
  ticker_2_sentiment_label TEXT,
  ticker_3 TEXT,
  ticker_3_relevance_score DECIMAL,
  ticker_3_sentiment_score DECIMAL,
  ticker_3_sentiment_label TEXT,
  ticker_4 TEXT,
  ticker_4_relevance_score DECIMAL,
  ticker_4_sentiment_score DECIMAL,
  ticker_4_sentiment_label TEXT,
  ticker_5 TEXT,
  ticker_5_relevance_score DECIMAL,
  ticker_5_sentiment_score DECIMAL,
  ticker_5_sentiment_label TEXT
);
COPY alpha_vantage_temp(title, url, time_published, authors, summary,
banner_image, source, category_within_source, source_domain,
overall_sentiment_score, overall_sentiment_label, topic_1,
topic_1_relevance_score, topic_2,  topic_2_relevance_score, topic_3,
topic_3_relevance_score, ticker_1, ticker_1_relevance_score,
ticker_1_sentiment_score, ticker_1_sentiment_label, ticker_2,
ticker_2_relevance_score, ticker_2_sentiment_score, ticker_2_sentiment_label,
ticker_3, ticker_3_relevance_score, ticker_3_sentiment_score,
ticker_3_sentiment_label, ticker_4, ticker_4_relevance_score,
ticker_4_sentiment_score, ticker_4_sentiment_label, ticker_5,
ticker_5_relevance_score, ticker_5_sentiment_score, ticker_5_sentiment_label)
FROM ‘\alpha_vantage.csv'
DELIMITER ','
CSV HEADER;
INSERT INTO alpha_vantage (reddit_comment_id, title, url, time_published, authors,
summary, banner_image, source, category_within_source, source_domain,
overall_sentiment_score, overall_sentiment_label, topic_1,
topic_1_relevance_score, topic_2, topic_2_relevance_score, topic_3,
topic_3_relevance_score, ticker_1, ticker_1_relevance_score,
ticker_1_sentiment_score, ticker_1_sentiment_label, ticker_2,
ticker_2_relevance_score, ticker_2_sentiment_score, ticker_2_sentiment_label,
ticker_3, ticker_3_relevance_score, ticker_3_sentiment_score,
ticker_3_sentiment_label, ticker_4, ticker_4_relevance_score,
ticker_4_sentiment_score, ticker_4_sentiment_label, ticker_5,
ticker_5_relevance_score, ticker_5_sentiment_score, ticker_5_sentiment_label)
SELECT rc.id, avt.title, avt.url, avt.time_published,
replace(replace(avt.authors, '{', '{'), '}', '}')::TEXT[], avt.summary,
avt.banner_image, avt.source, avt.category_within_source, avt.source_domain,
avt.overall_sentiment_score, avt.overall_sentiment_label, avt.topic_1,
avt.topic_1_relevance_score, avt.topic_2,  avt.topic_2_relevance_score,
avt.topic_3, avt.topic_3_relevance_score, avt.ticker_1,
```

```
avt.ticker_1_relevance_score,   avt.ticker_1_sentiment_score,
avt.ticker_1_sentiment_label, avt.ticker_2, avt.ticker_2_relevance_score,
avt.ticker_2_sentiment_score,  avt.ticker_2_sentiment_label, avt.ticker_3,
avt.ticker_3_relevance_score, avt.ticker_3_sentiment_score,
avt.ticker_3_sentiment_label, avt.ticker_4, avt.ticker_4_relevance_score,
avt.ticker_4_sentiment_score, avt.ticker_4_sentiment_label, avt.ticker_5,
avt.ticker_5_relevance_score, avt.ticker_5_sentiment_score,
avt.ticker_5_sentiment_label
FROM alpha_vantage_temp avt
JOIN reddit_comments rc ON rc.ticker IN (avt.ticker_1, avt.ticker_2, avt.ticker_3,
avt.ticker_4, avt.ticker_5);
Tabla top_10_tickers:
CREATE TEMP TABLE top_10_tickers_temp (
  ticker TEXT,
  number_of_mentions INT
);
COPY top_10_tickers_temp(ticker, number_of_mentions)
FROM 'C:\Users\Alberto\OneDrive\Escritorio\PFG Reddit\top_10_tickers.csv'
DELIMITER ','
CSV HEADER;
INSERT INTO top_10_tickers (reddit_comment_id, ticker, number_of_mentions)
SELECT rc.id, ttt.ticker, ttt.number_of_mentions
FROM top_10_tickers_temp ttt
JOIN reddit_comments rc ON rc.ticker = ttt.ticker;
```

A.4 Creación de los KPIs

Este apartado incluye desde la conexión con la base de datos hasta la obtención de todos los KPIs. A continuación, se presentan todos los pasos:

1. Abrir Power BI y realizar los pasos que se indican en las siguientes figuras. Los nombres de la base de datos es el que se ha puesto para la realización de este proyecto, va a depender del que le ponga el usuario. El usuario y contraseña son los indicados en la instalación de PostgreSQL del apartado A.3.

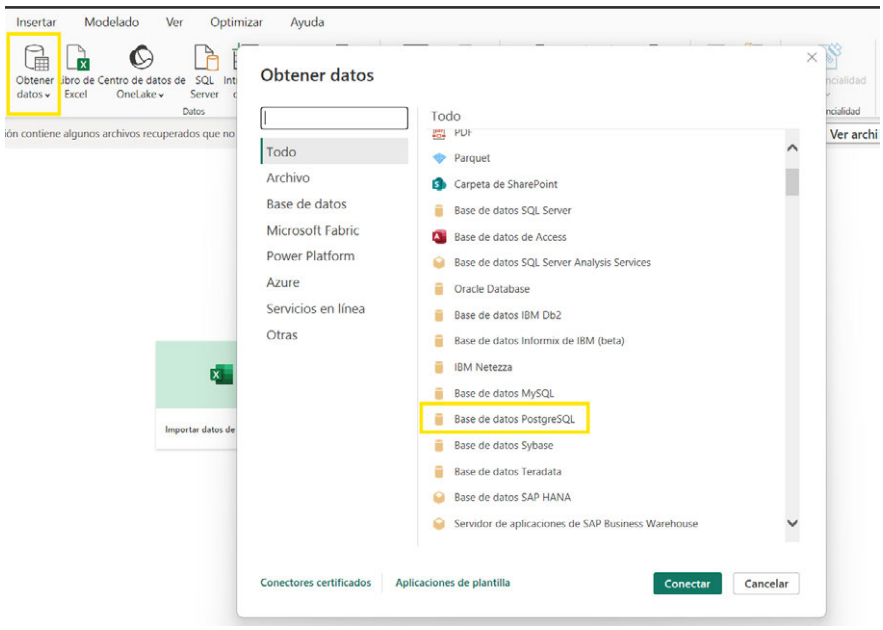

Figura 9.2 - Primer paso para la importación de datos a Power BI. Elaboración propia

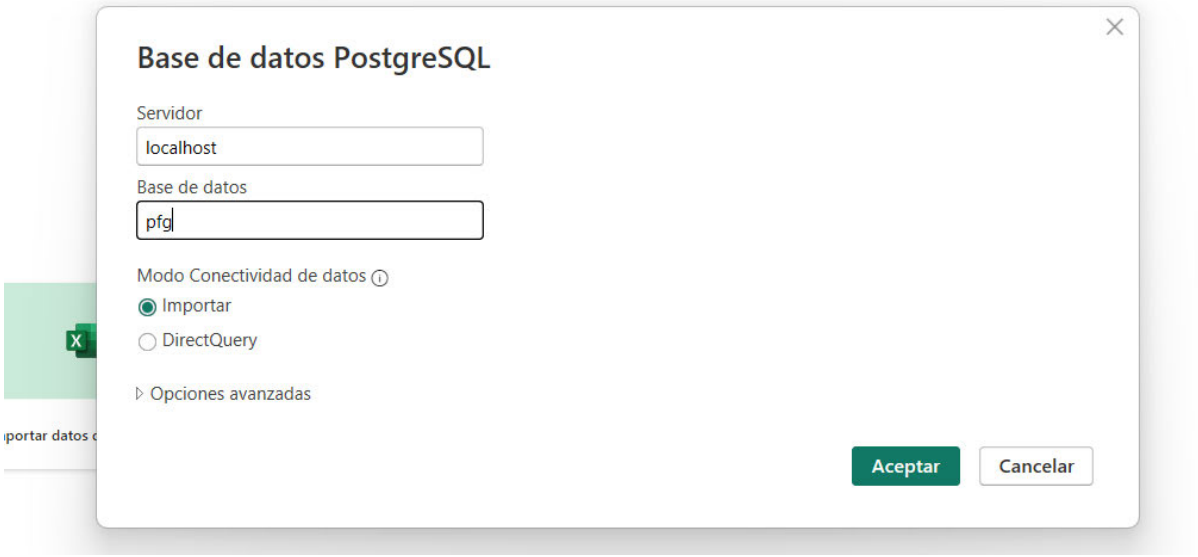

Figura 9.3 - Segundo paso para la importación de datos a Power BI. Elaboración propia

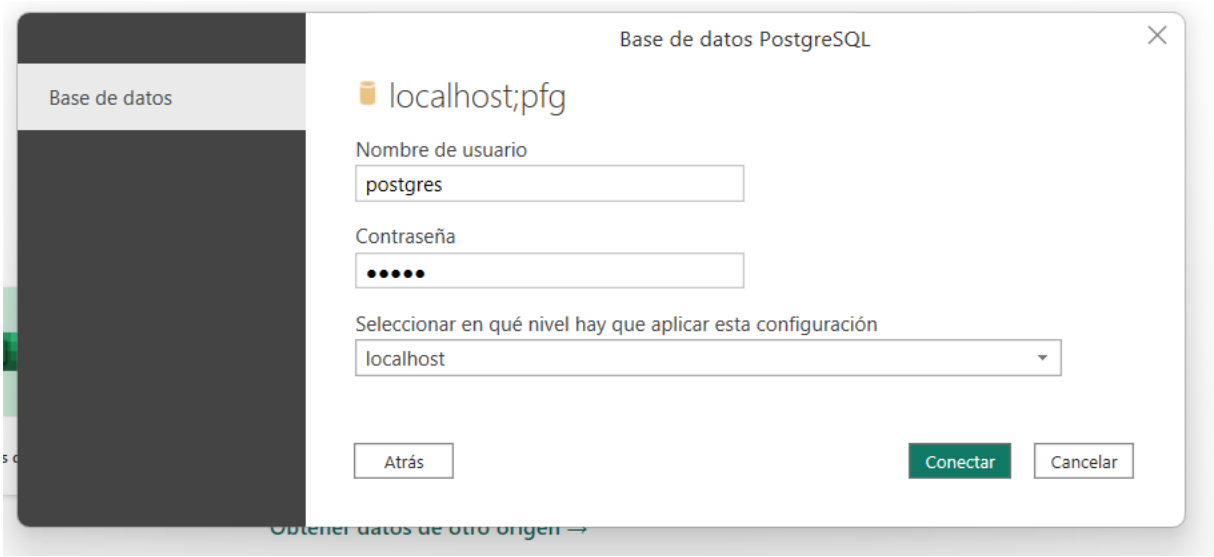

Figura 9.4 - Tercer paso para la importación de datos a Power BI. Elaboración propia

2. Tras estos pasos se puede observar en la siguiente figura, el esquema presentado en la Figura 5.1, en Power BI.

Figura 9.5 - Modelo de datos representado en Power BI. Elaboración propia.

3. Creación del KPI de **Volumen de comentarios por Ticker.**

Para la realización de este KPI hay que seleccionar los valores '*Ticker*' y '*Sentiment*' de la tabla '*reddit_comments*'. Para cada uno de los casos estudiados se selecciona un ticker en específico, en este se puede apreciar en la figura que es NVDA.

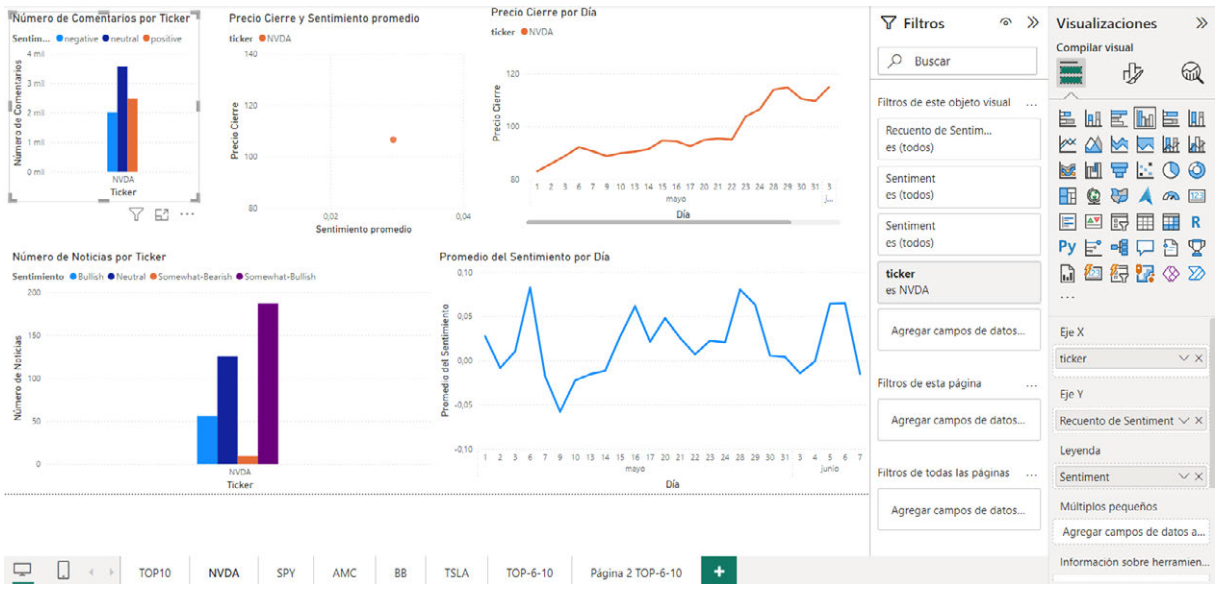

Figura 9.6 - Elaboración del KPI Volumen de comentarios por Ticker. Elaboración propia.

4. Creación del KPI de **Relación entre el precio de cierre y el sentimiento.**

Para este KPI hay que seleccionar el '*Sentiment*' de la tabla '*reddit_comments*' y el '*Close*' de la tabla '*yahoo_finance*', tal como aparece en la siguiente figura. También hay que definir el ticker para cada uno de los casos estudiados.

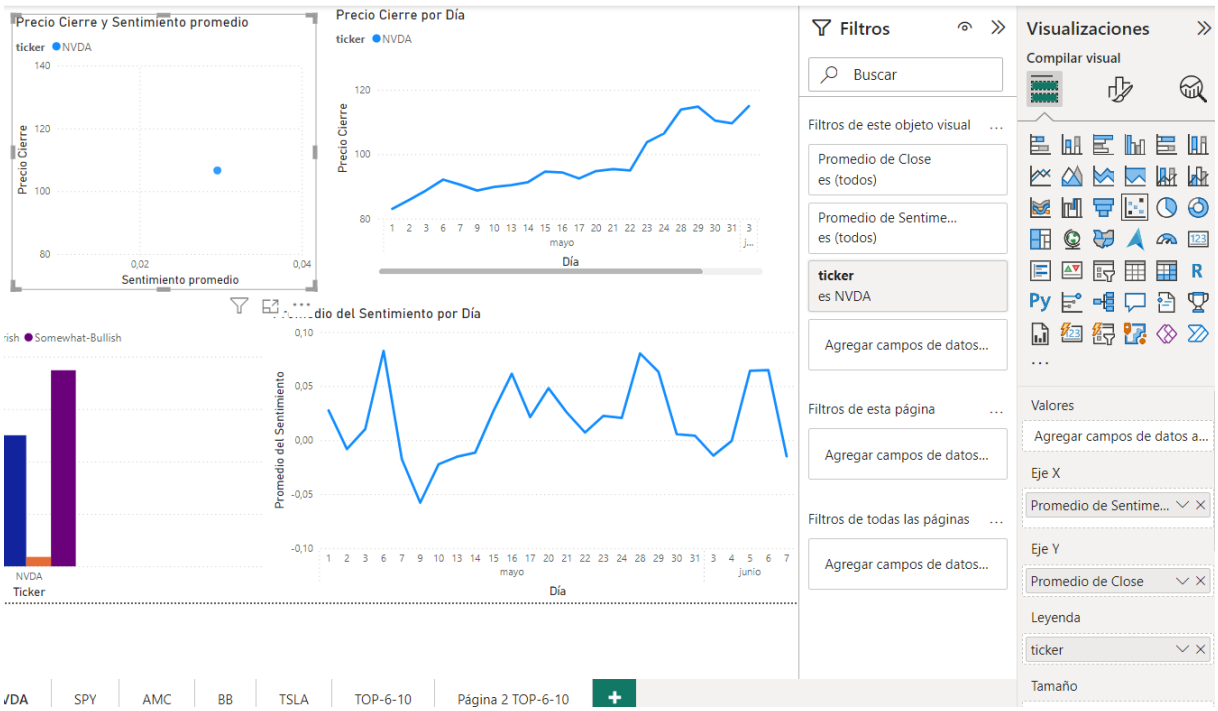

Figura 9.7 - Elaboración del KPI Relación entre el precio de cierre y el sentimiento. Elaboración propia.

5. Creación del KPI de **Tendencia de precio de cierre.**

La generación de este KPI conlleva la utilización de los campos '*Date*' y '*Close*' de la tabla '*yahoo_finance*'. También se tiene que seleccionar el ticker y la fechas entre las cuales se han obtenido datos de Reddit, esto se puede apreciar en la siguiente figura.

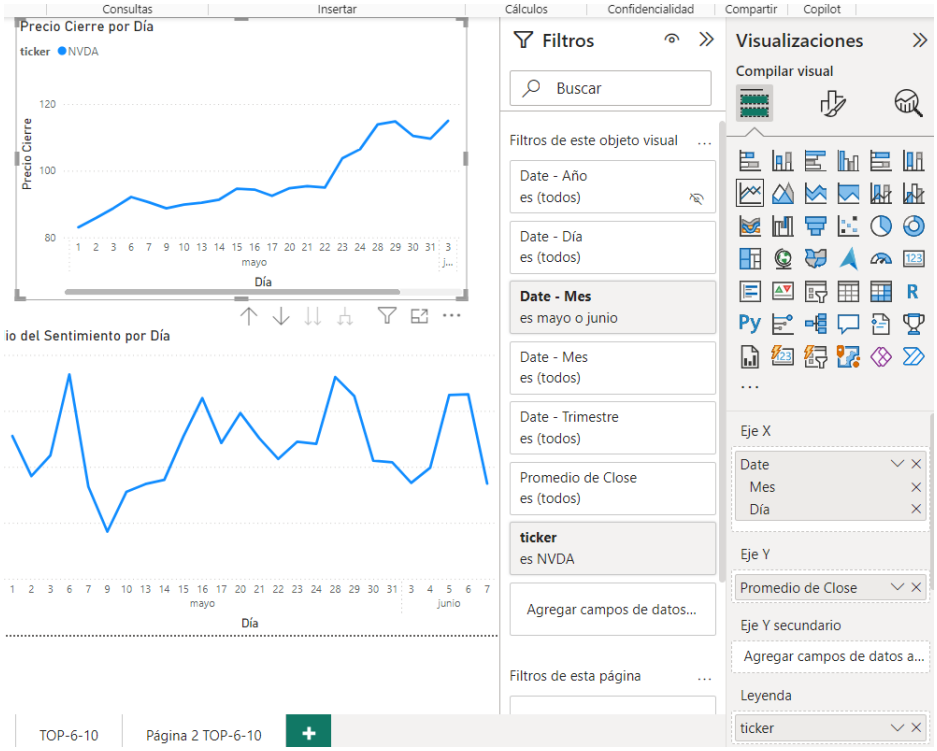

Figura 9.8 - Elaboración del KPI Tendencia de precio de cierre. Elaboración propia.

6. Creación del KPI de **Número de noticias por Ticker.**

Este KPI muestra los datos de los campos '*ticker_1*' y '*ticker_1_sentiment_label*' de la tabla '*alpha_vantage*'. Se tiene que seleccionar el ticker y la fecha de estudio, tal como aparece en la figura siguiente.

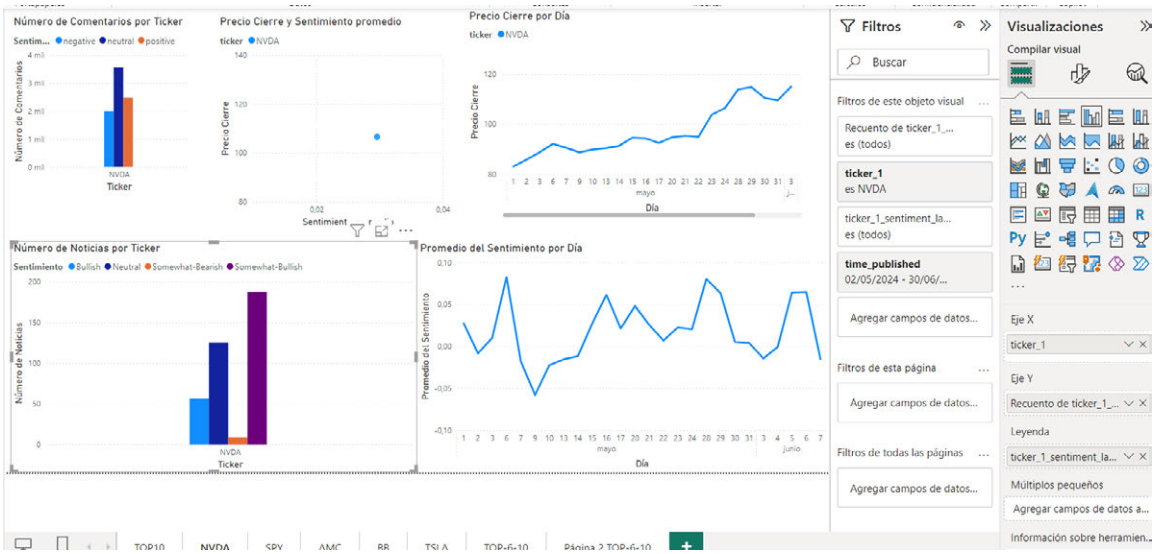

Figura 9.9 - Elaboración del KPI Número de noticias por Ticker. Elaboración propia.

7. Creación del KPI de **Sentimiento promedio por Ticker.**

Para la obtención de este KPI, se seleccionan los campos '*Date*' y '*Sentiment*' de la tabla '*reddit_comments*'. A continuación, se observan los campos seleccionados en la figura.

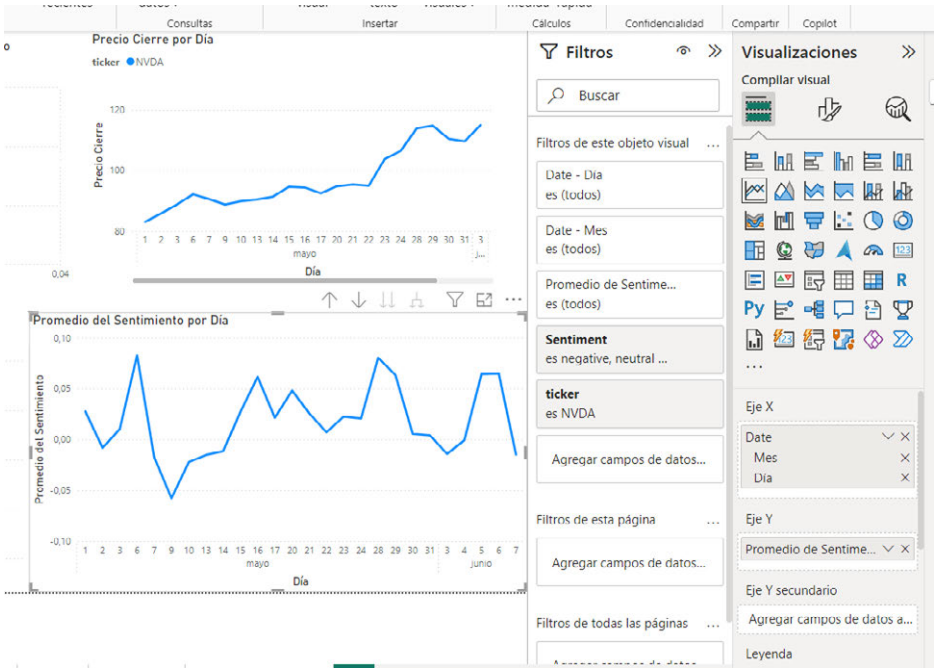

Figura 9.10 Elaboración del KPI Sentimiento promedio por Ticker. Elaboración propia.

8. Creación del KPI de **Top 10 tickers.**

Para obtenerlo se tienen que seleccionar los datos de la tabla '*top_10_tickers*'. Como se puede apreciar en la siguiente figura.

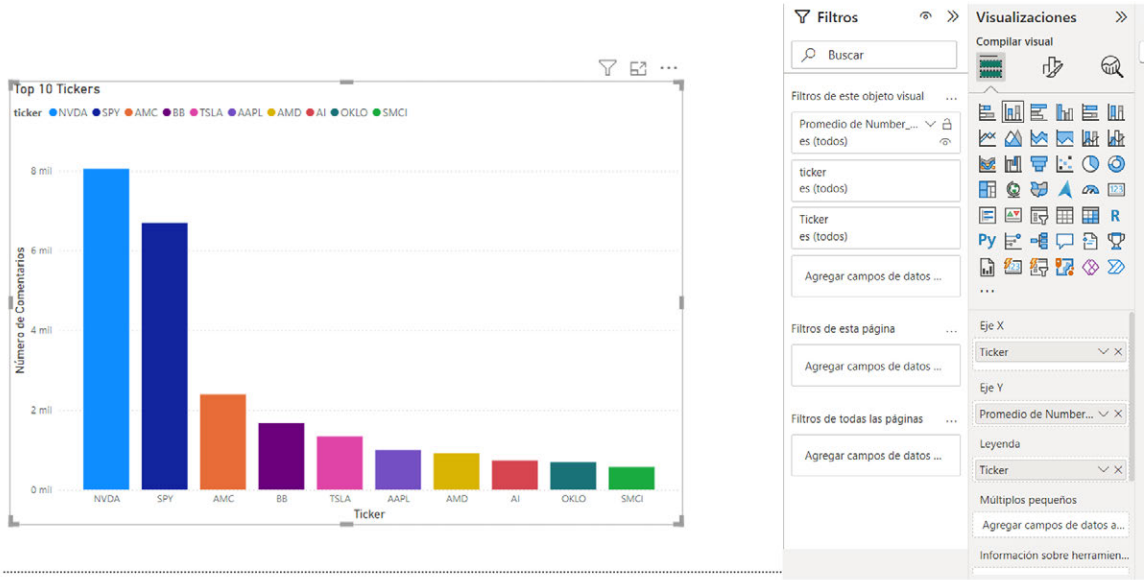

Figura 9.11 - Elaboración del gráfico de Top 10 Tickers. Elaboración propia.