

TFM - Máster Data Science

Análisis Sentimiento Twitter

Índice

[TFM - Máster Data Science 0](file:///C:\Users\andominguez\Desktop\KSchool\Documento%20tfm\TFM%20-%20KSchool.docx#_Toc48586347)

[Análisis Sentimiento Twitter 0](file:///C:\Users\andominguez\Desktop\KSchool\Documento%20tfm\TFM%20-%20KSchool.docx#_Toc48586348)

[1 Introducción 2](#_Toc48586349)

[1.1. Estado del Arte 2](#_Toc48586350)

[El Corte Inglés 2](#_Toc48586351)

[Twitter 2](#_Toc48586352)

[2 Alcance TFM 4](#_Toc48586353)

[2.1. Data Engineering 5](#_Toc48586354)

[Volumetría 5](#_Toc48586355)

[Información Tweets 5](#_Toc48586356)

[2.2. Data Science 7](#_Toc48586357)

[2.3. Reporting 9](#_Toc48586358)

[2.4. Ejecución del Programa 11](#_Toc48586359)

[3 Investigación 11](#_Toc48586360)

[3.1. Data Engineering 11](#_Toc48586361)

[3.2. Data Science 11](#_Toc48586362)

[4 Resultados 12](#_Toc48586363)

[Estado general de enero a junio del 2020 12](#_Toc48586364)

[Opinión cuentas Following 12](#_Toc48586365)

[Opinión sobre COVID-19 13](#_Toc48586366)

[5 Conclusión 15](#_Toc48586367)

[6 Interfaz Gráfica 16](#_Toc48586368)

[7 Anexo 17](#_Toc48586369)

[7.1. El Corte Inglés 17](#_Toc48586370)

[7.2. Twitter 17](#_Toc48586371)

[7.3. Repositorio 17](#_Toc48586372)

[7.4. Limitaciones Twiiter 17](#_Toc48586373)

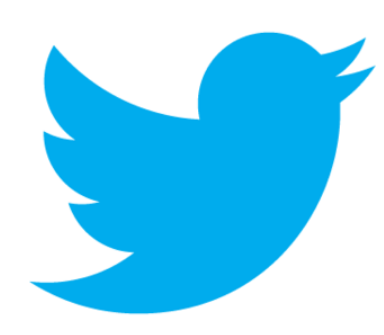
[7.5. MicroStrategy 17](#_Toc48586374)

# Introducción

El objetivo del proyecto fin de máster es averiguar la opinión que tiene los usuarios de una determinada red social sobre una empresa, concepto, tendencia social, etc.

El proyecto se ha orientado sobre la empresa de retail *El Corte Inglés* [7.1] y a la plataforma de *Twitter*[7.2]. Por tanto, se va a realizar un análisis de sentimiento de los tweets generados por los usuarios en un determinado periodo de tiempo haciendo énfasis en:

* Cuentas que me siguen
* Covid-19.

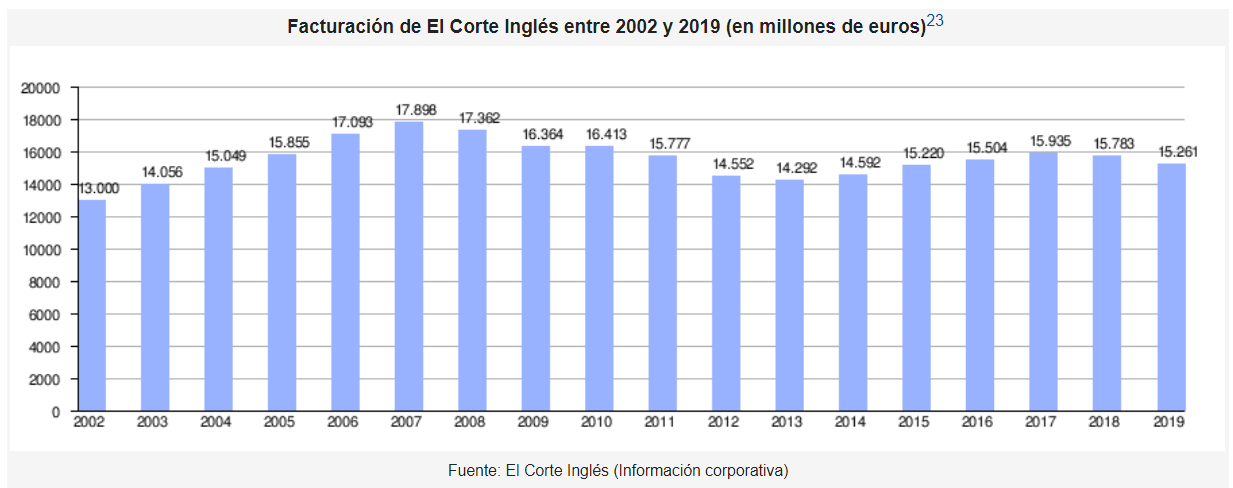
 

## Estado del Arte

### El Corte Inglés

El Corte Inglés es un grupo de [distribución](https://es.wikipedia.org/wiki/Distribuci%C3%B3n_(negocios)) del sector Retail española compuesto por [empresas](https://es.wikipedia.org/wiki/Empresa) de distintos formatos, siendo el principal el de grandes almacenes. Algunas de ellas son HiperCor, SuperCor, Sfera, etc.

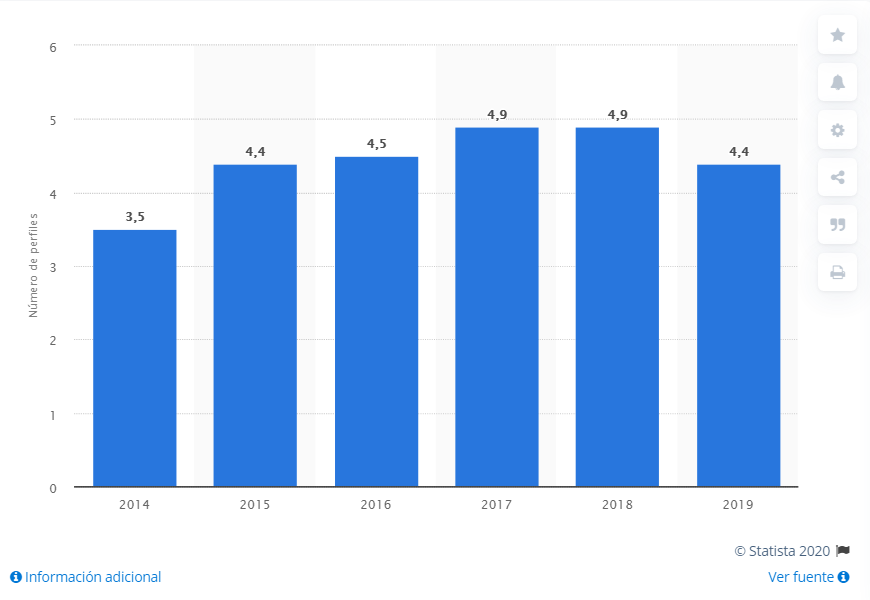
Con más de 90 centros comerciales distribuidos entre España y Portugal, es considerado la empresa de Retail más importante del país, con una facturación de más 15.000 millones de euros.



### Twitter

Twitter es actualmente una de las redes sociales y plataforma de comunicación más populares del mundo, presente en todo el planeta y con más de 150 millones de usuarios activos registrados.

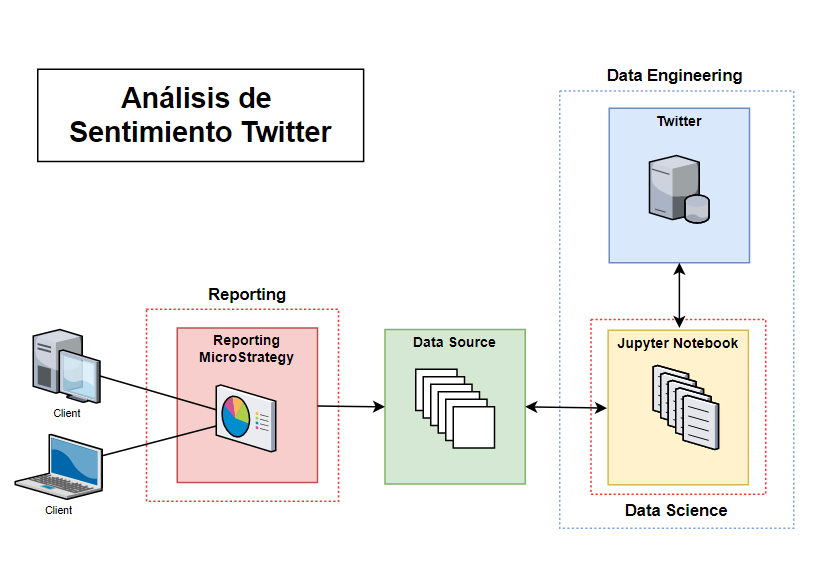
En España, son más de 4 millones de personas registradas en la plataforma.



Actualmente, existen muchos tipos de proyectos que se dediquen al análisis del sentimiento de las redes sociales en la actualidad. Sin embargo, en este TFM, se propone el extraer toda esa información y representarla en un cuadro de mando donde el usuario puede, de manera rápida, sencilla y dinámica profundizar sobre los datos.

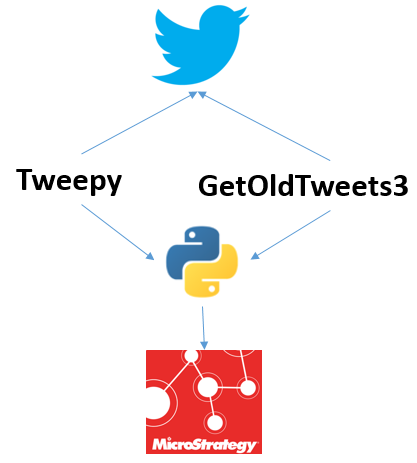
# Alcance TFM

A continuación, se presenta un esquema de los principales componentes del proyecto:



El proyecto consta de tres partes bien diferenciadas.

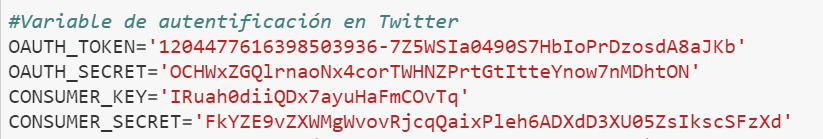
* Data Engineering: Extracción de tweets de un determinado periodo de tiempo
* Data Science: Medición del sentimiento de los tweets
* Reporting: Visualización de los datos



A continuación, se detalla cada uno de los niveles:

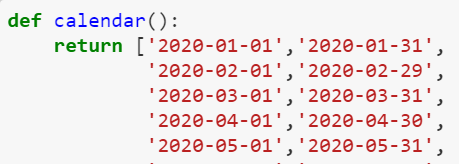
## Data Engineering

Para poder acceder a Twitter, es necesario acceder con unas credenciales que nos facilita Twitter una vez que nos hemos registrado correctamente.



Dada la limitación que tienen las cuentas gratuitas de desarrollador de Twitter, se ha utilizado la librería Python – **GetOldTweets3** [7.3] para poder recuperar los tweets con una profundidad histórica de más de una semana.

Una vez instalada y para poder extraer los datos, se ha utilizado los parámetros “desde-hasta” mediante un calendario junto con la palabra a buscar, que en este ha sido “el corte ingles”.



### Volumetría

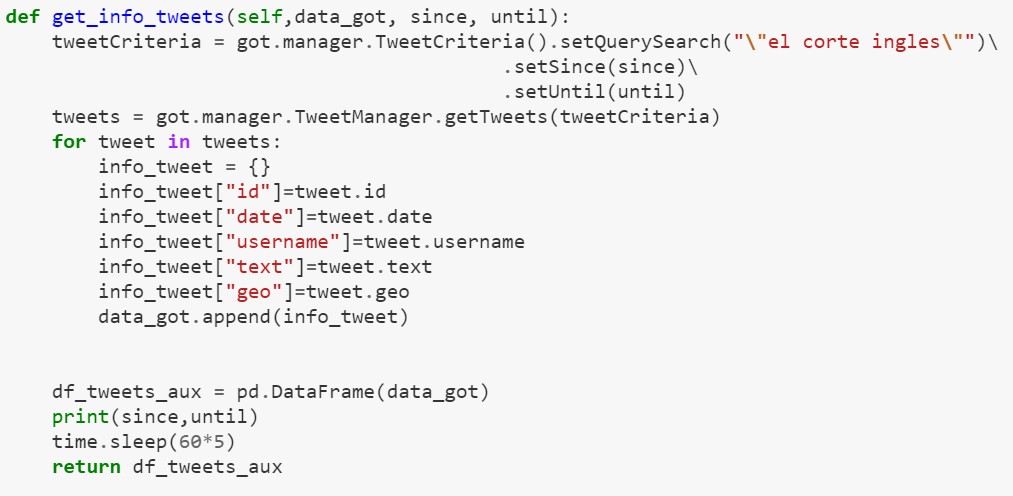
El periodo a tratar en el proyecto de TFM ha sido de enero 2020 a junio 2020 con la siguiente volumetría por mes:

|  |  |
| --- | --- |
| **Mes** | **Volumetría** |
| Enero | 10.089 |
| Febrero | 6.931 |
| Marzo | 10.486 |
| Abril | 2.277 |
| Mayo | 8.681 |
| Junio | 9.215 |

### Información Tweets

Los datos que se han recuperado han sido:

* ID del tweet
* Fecha de Creación del tweet
* Usuario que ha creado el tweet
* Tweet
* Geolocalización (para trabajo futuro)



Dado que posee una limitación de conexión abierta en GetOldTweets3, por cada iteración, se deja al procedimiento un tiempo de 5min en reposo. El resultado se almacena en un Dataframe donde se irá añadiendo los resultados de las iteraciones posteriores.

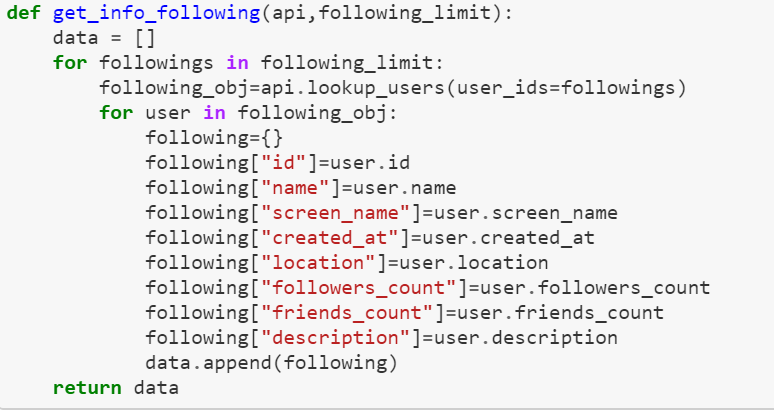
Para poder recuperar la información relativa al *El Corte Inglés*, en la actualidad, tiene 292K followers y 1.628 following.



Para extraer la información de las cuentas y realizar un análisis más exhaustivo de las cuentas followings, se utilizado la librería Tweepy [7.4] de Python.

A continuación, se expone el código para la extracción de la información las cuentas que sigue la cuenta El Corte Inglés. Para ello, primero se extrae la información de los IDs de las cuentas y posteriormente, se agrupa por grupos y se extrae la información de la cada una de las cuentas.





Los Dataframes generados con la información de los Tweets y de los followings son:

* *dt\_tweets*
* *Info\_following\_el\_corte\_ingles*

## Data Science

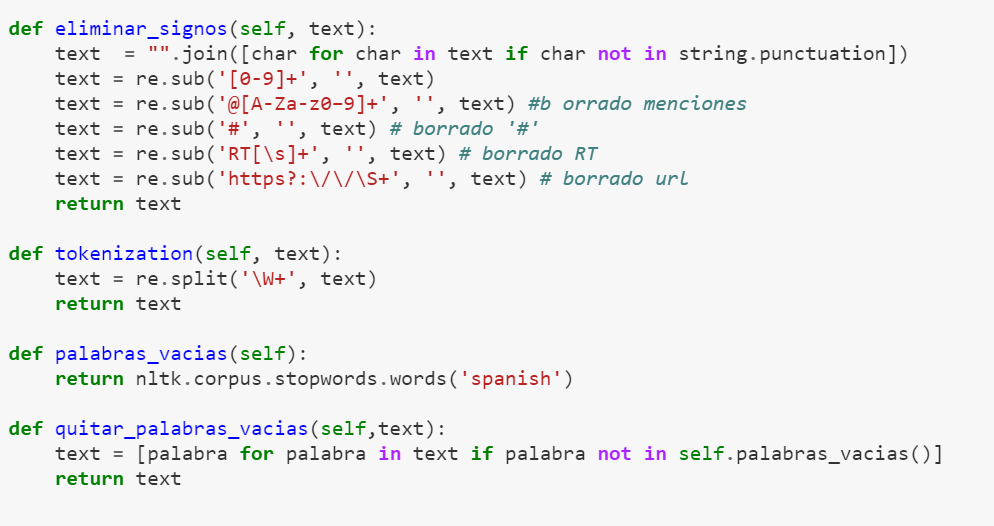
Una vez extraída la información de todos los tweets generados en el periodo de tiempo fijado y la información de las cuentas que sigue *El Corte Inglés*, se ha utilizado para medir el sentimiento de los tweets la librería TextBlob y NLTK.

El procedimiento que mide la subjetividad y objetividad de cada uno de ellos es el siguiente:



A dicho método, se le pasa como parámetro el Dataframe obtenido de la fase anterior que contiene toda la información de los tweets. Pero antes de pasar el dato “puro” se ha procedido a realizar una limpieza del dato. Para ello, se han creado las siguientes funciones:

* Conversión de mayúsculas a minúsculas.
* Traducción de los tweets al español.
  + Se procede a traducir todos los tweets a una única lengua debido a que se aprecia información de diferentes lenguas (inglés, catalán, etc)
* Eliminación de signo de puntuación.
  + Se eliminan todos los puntos de puntuación, pero además se procede a eliminar:
    - Números
    - Palabras claves como RT y #
    - URLs
* Tokenización de los tweets.
  + Se almacena en un objeto de tipo lista las palabras resultantes de la primera fase de limpieza de los datos.
* Eliminación de palabras vacías.
  + Se eliminan aquellas palabras vacías que no aportan valor al análisis del sentimiento de los tweets. Para ello, se ha utilizado la librería NLTK.



Para poder extraer la valoración que representa el sentimiento de un tweet, se ha utilizado la función *sentiment.polarity* incorporada en la librería TextBlob. Para ello, para cada una de las filas del dataset se ha calculado la objetividad de cada uno de ellos, clasificando el resultado en:

* -1: Valoración Negativa
* 0: Valoración Neutra
* 1: Valoración Positiva.

Estos valores, luego serán ponderados en el cuadro de mando de MicroStartegy representando el valor real.

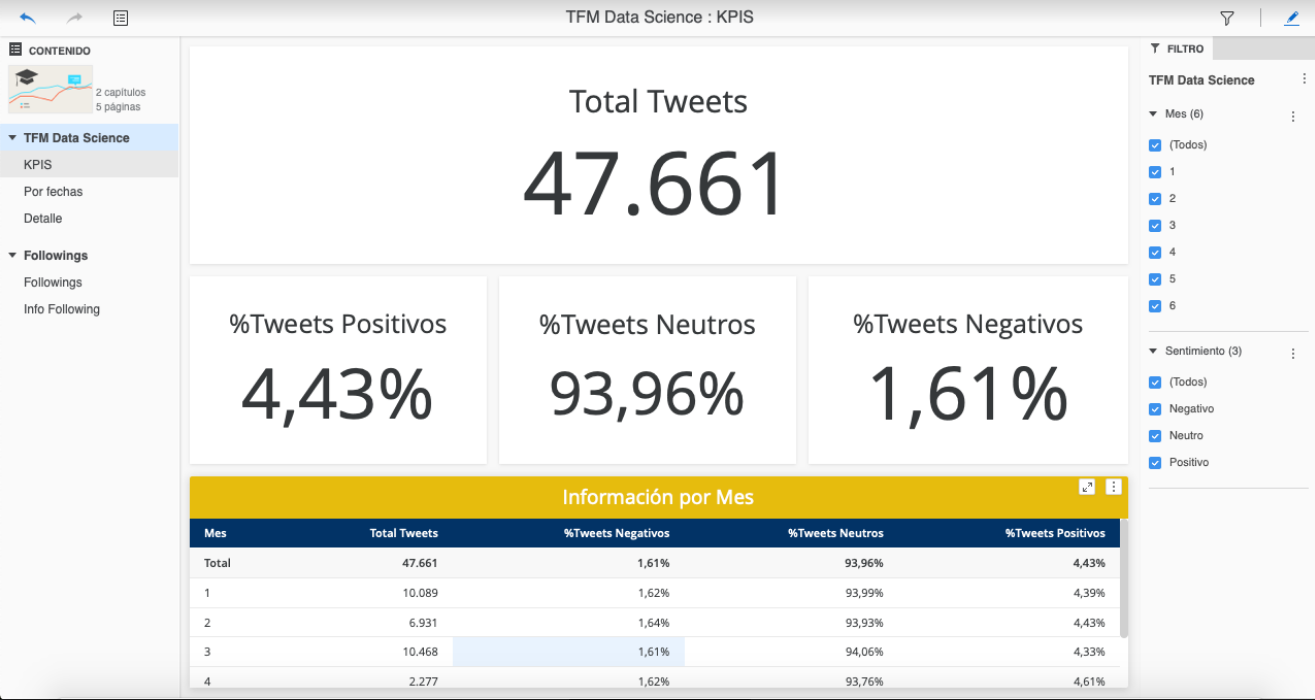
## Reporting

La herramienta que se ha utilizado para presentar los datos ha sido MicroStrategy [7.7]. Este software empresarial te permite crear de forma rápida y sencilla, un cuadro de mando.

El cuadro de mando se compone de dos capítulos:

* Análisis Temporal.
  + KPIS principales
  + Por Fechas
  + Detalle
* Análisis Followings.
  + Followings
  + Info Followings.
* Análisis Covid
  + KPIS principales
  + Por Fechas
  + Detalle

El origen del cuadro de mando son ficheros CSV generados por los notebooks de Python. Dichos ficheros se encuentran disponibles en el repo del proyecto [7.5]



## Ejecución del Programa

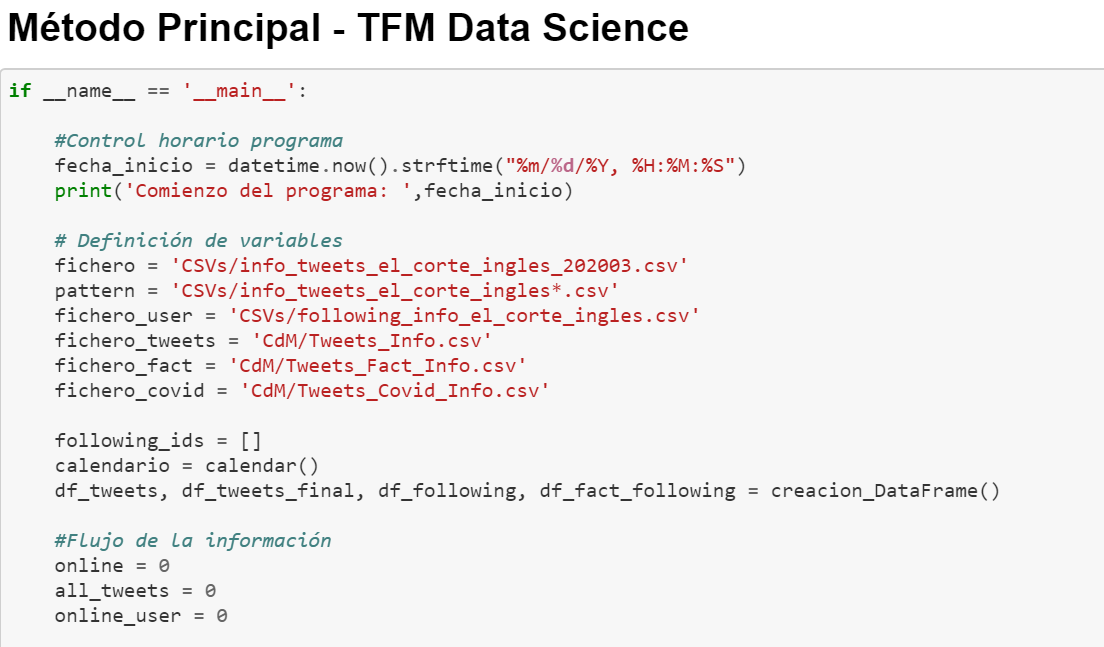
Para poder ejecutar el programa, el único notebook que se tiene que ejecutar es:

tfm\_Data\_Science.ipynb

Este a su vez, irá convocando al resto de notebooks que contienen los métodos necesarios para su correcta ejecución.

Consta de varios flujos de ejecución:

* *Online*=1
  + Conexión mediante GetOldTweets3 y decarga de las fechas disponibles en el calendario.
* *Online*=0 y all\_tweets=1
  + Lectura de los ficheros CSVs ya almacenados en el directorio CSVs sin conexión a Twiiter.
* *Online*=0 y *all\_tweets*=0
  + Lectura de un único fichero.
* *Online\_user* = 1
  + Conexión mediante Tweepy a Twitter para recuperar la información de las cuentas following.



# Investigación

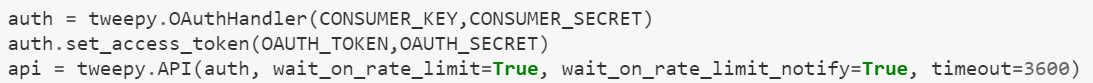
A continuación, se detallan cada uno de los componentes de investigación del TFM:

## Data Engineering

Al conectarnos a Twitter mediante la librería de Python Tweepy, observamos algunas limitaciones de la API de Twitter. Algunas de éstas restricciones [7.6] son las siguientes:

* Acceso a Tweet de más de 7 días.
* Limitación de recuperar información de cuentas following/followers.
* Timeout de la conexión.

Para solventar algunos de estos problemas, nos hemos apoyado en las siguientes soluciones:



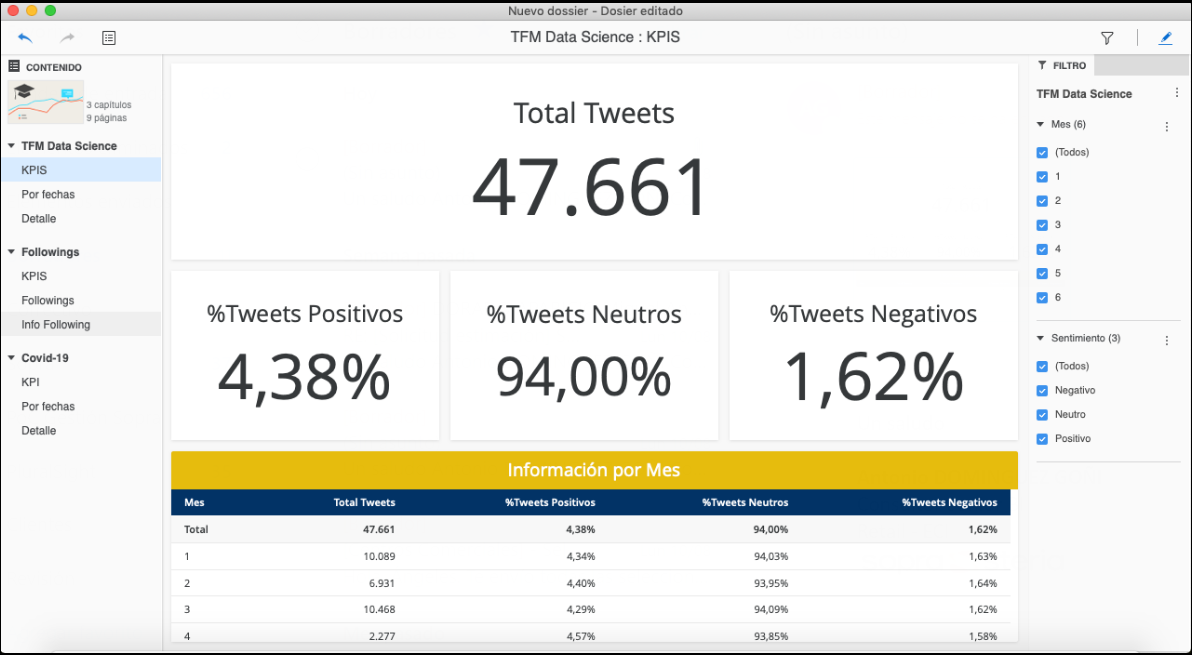
Inicializando estos parámetros, nos ayuda a que sea la propia API quien gestione los posibles errores de ejecución. Sin embargo, algunos códigos los hemos incluido dentro de una Try/Catch dejando un tiempo de reposo.

En el caso de la limitación de aquellos tweets con más de siete días de antigüedad, como ya se ha explicado en el apartado de Alcance TFM, se ha procedido a utilizar la librería de Python GetOldTweets3.

# Resultados

Una vez terminado el proceso de extracción y análisis de los tweets generados con la etiqueta *El Corte Inglés* para un determinado rango de fechas , el cuadro de mando arroja los siguientes resultados.

### Análisis del estado general de enero a junio del 2020

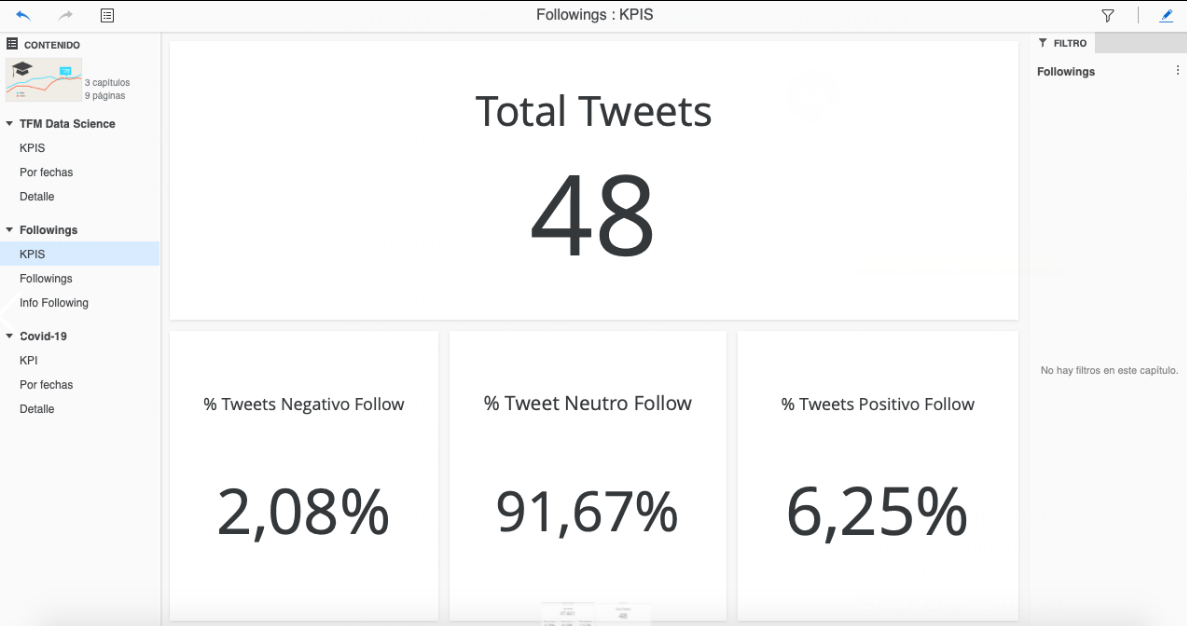


En términos generales, la opinión de la red social ha sido bastante neutra, incluso siendo positiva para algunos usuarios (representando el 5%) y prácticamente sin comentarios negativos.

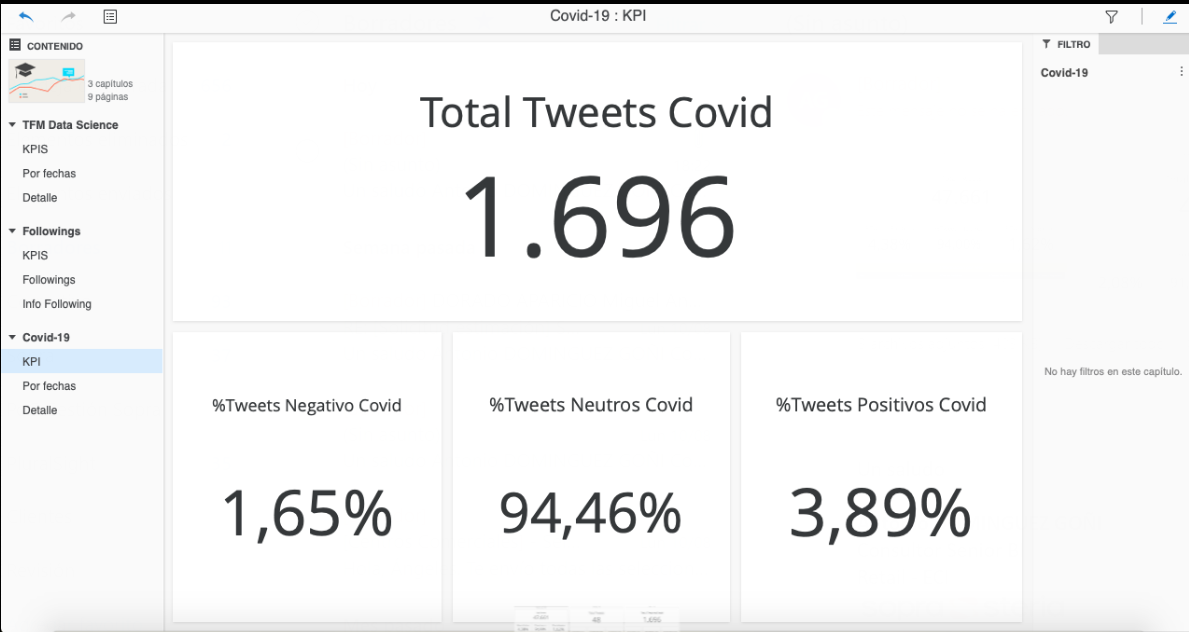
También como se puede apreciar el número de Tweets disminuye a partir de febrero debido a la finalización de la campaña de navidad y las rebajas. Sin embargo, se vuelve a apreciar un incremento de Tweets en el mes de marzo, superando incluso al mes de enero, debido a la crisis sanitaria del Covid-19.

### Análisis opinión cuentas Following

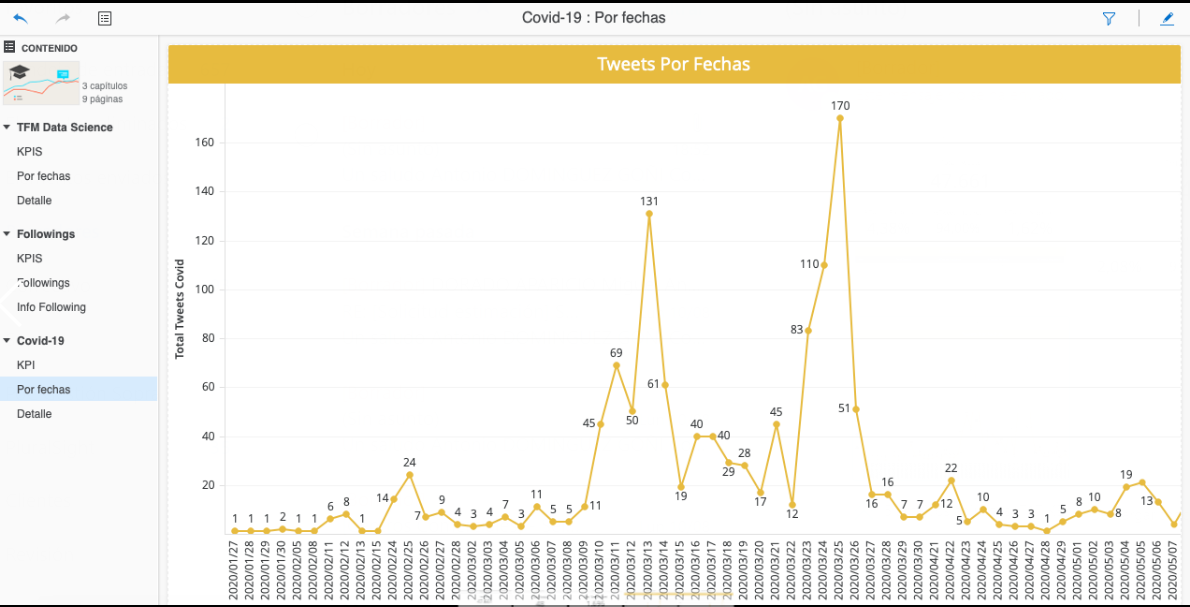
Para mi empresa, puede ser muy interesante la información de las cuentas que hago following para poder destinar recursos a dichas cuentas. La opinión de dichas cuentas, sigue la misma tendencia que el conjunto total de Tweets pero sí se ve un pequeño incremento de comentarios positivos (cerca del 7%)



### Análisis sobre COVID-19

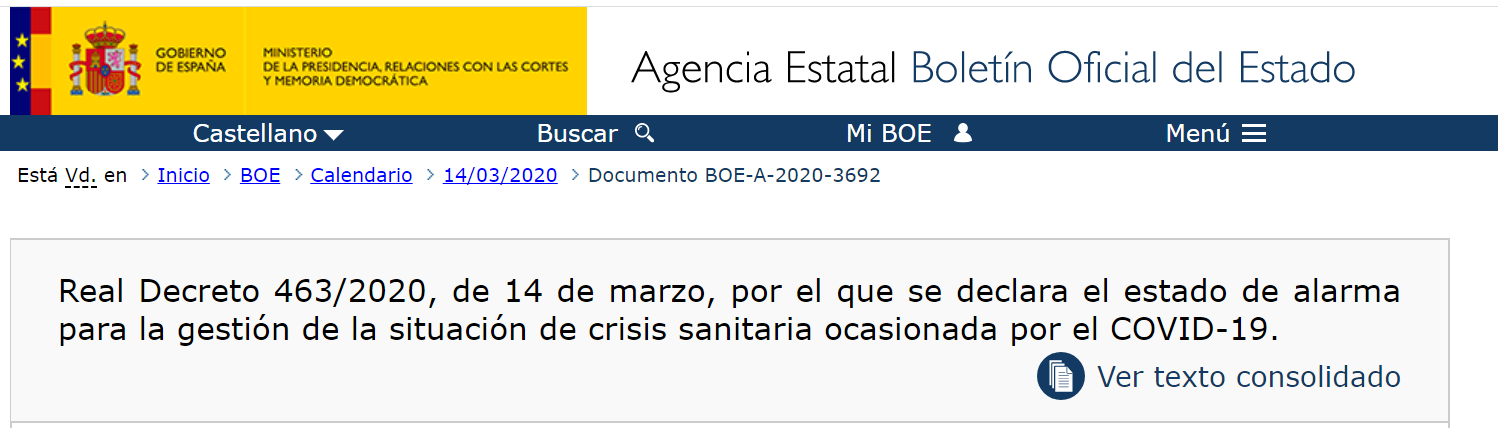


Profundizando en los datos del Covid, vemos un despunte de los tweets generados sobre los días 13/03 al 25/03.



Dichos días coinciden con:

* 13/03/2020: Estado de alerta. Confinamiento.



* 25/03/2020: Publicación por parte de El Corte Inglés que aplica ERTE en su organización



# Conclusión

La conclusión de este TFM presentar de manera clara y sencilla al usuario lo que está opinando las redes sociales sobre su organización, profundizando a nivel de Tweet, cuenta, día y opinión.

Teniendo accesible dicha información, los usuarios que se encarguen del análisis de esta información podrán actuar de manera eficaz para poder identificar la raíz del problema de una manera rápida y concisa y destinar los recursos necesarios para mejorar la opinión pública.

Respecto a los datos que devuelve el algoritmo de TextBlob para obtener el sentimiento de los tweets, es cierto que el volumen de tweets neutros es muy considerable, considerando el aplicar otro algoritmo de machine learning para su tratamiento y comparando resultados.

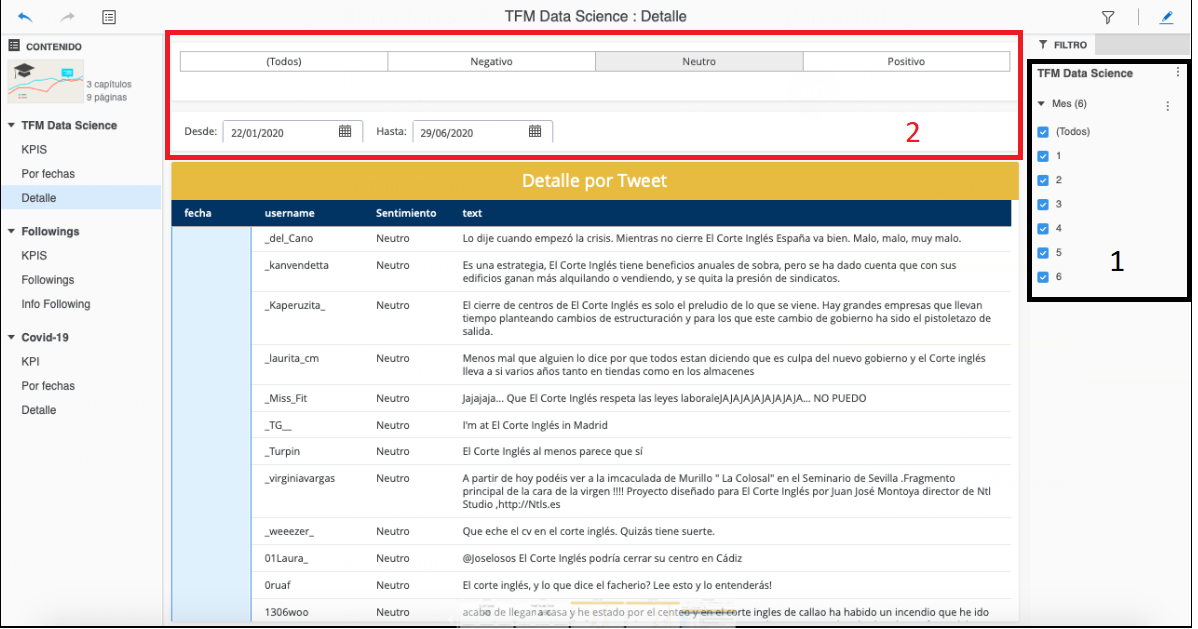
# Interfaz Gráfica

La forma de interactuar con el cuadro de mando es muy sencilla. Una vez ejecutado, consta de 3 capítulos y varias páginas cada uno de ellos.

* Capitulo: Análisis Temporal.
  + Pág: KPIS principales
  + Pág: Por Fechas
  + Pág: Detalle
* Capitulo: Análisis Followings.
  + Pág: Followings
  + Pág: Info Followings.
* Capitulo: Análisis Covid
  + Pág: KPIS principales
  + Pág: Por Fechas
  + Pág: Detalle

Para poder filtrar sobre los datos, consta de dos tipos de filtros.

* 1- Filtros de capítulo: Aplica a todos los objetos de ese capítulo
* 2- Filtro de página: Aplica filtros sobre esa página.



# Anexo

## El Corte Inglés

Para más información <https://es.wikipedia.org/wiki/El_Corte_Ingl%C3%A9s>

## Twitter

Para más información <https://es.wikipedia.org/wiki/Twitter>

## GetOldTweets3

Para más información <https://pypi.org/project/GetOldTweets3/>

## Tweepy

Para más información <http://docs.tweepy.org/en/latest/>

## Repositorio

El repositorio de código se encuentra en la siguiente URL: Github.com/antdomgoni/tfm

## Limitaciones Twiter

Información sobre las limitaciones API Twitter:

<https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits>

## MicroStrategy

MicroStrategy: Software empresarial que permite crear informes, documentos y tableros sobre datos almacenados en diferentes bases de datos, CSVs y otras fuentes para posteriormente realizar análisis sobre los mismos.

Wikipedia: <https://es.wikipedia.org/wiki/MicroStrategy>

Para poder descargar MicroStrategy Desktop se puede realizar desde la siguiente URL: <https://www.microstrategy.com/es/get-started/desktop>

