Obligatorio Big Data en Inversiones

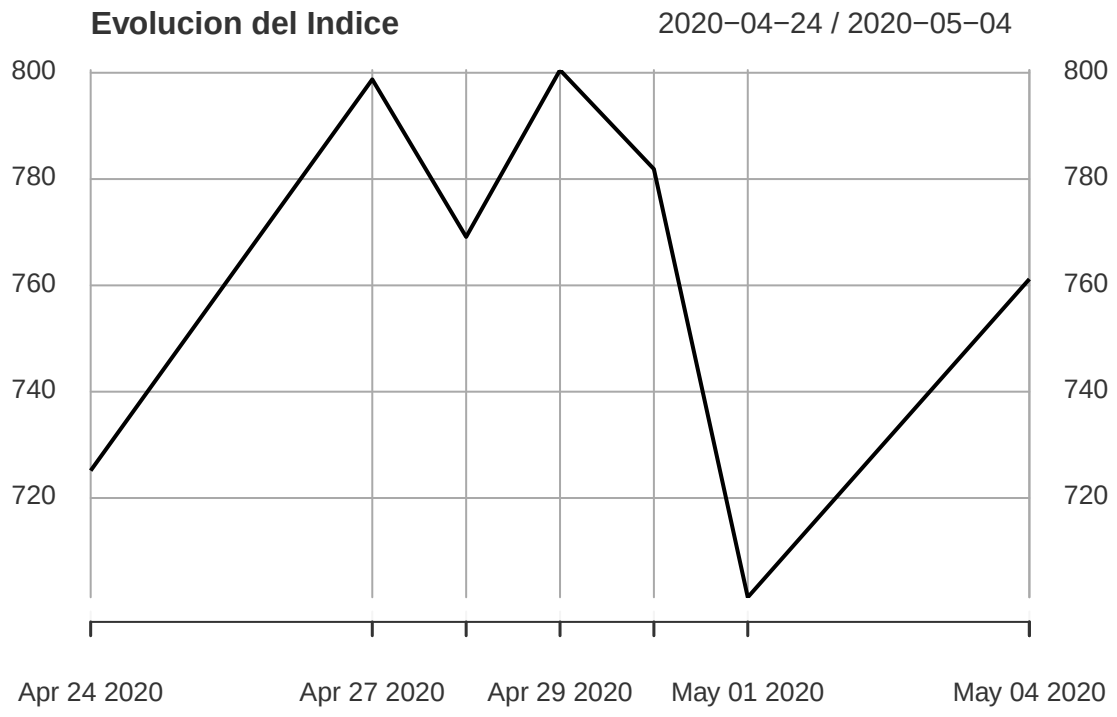
Camblor - González

5/5/2020

Contents

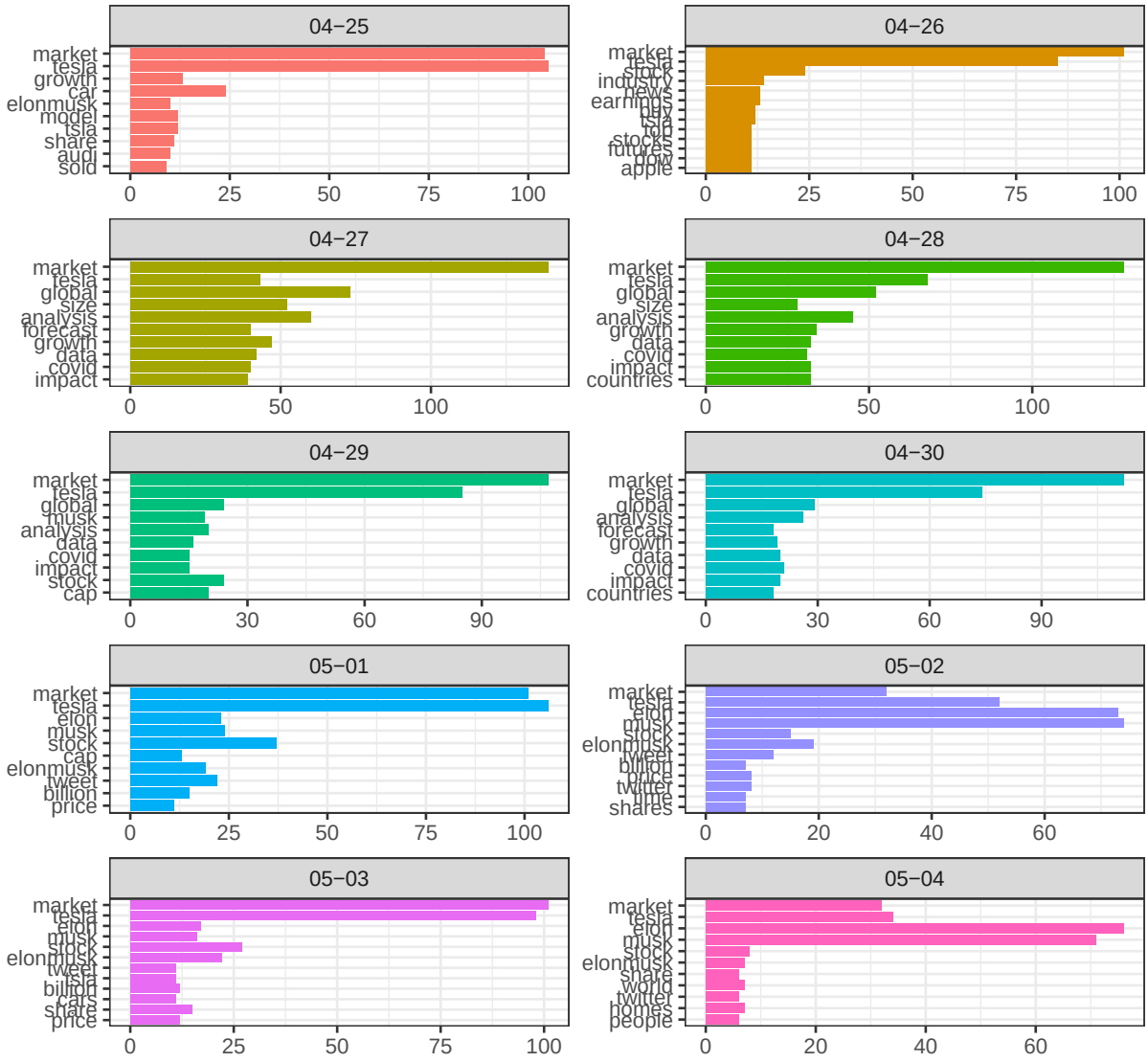
Introducción

Twitter es actualmente una dinámica y gran fuente de contenidos que, dada su popularidad e impacto, se ha convertido en la principal fuente de información para estudios de Social Media Analytics. Su utilización para el análisis de reputación de empresas, productos o personalidades, estudios de impacto relacionados con marketing, extracción de opiniones y predicción de tendencias son sólo algunos ejemplos de aplicaciones. En este contexto, decidimos realizar un análisis partiendo de Twitter tomando como foco tanto a la empresa Tesla como a su CEO Elon Musk. Musk presenta una grado de participación más que significativo en la red social, teniendo casi 4 millones de seguidores y mencionando activamente a la empresa. Nos motivó este caso a raíz de un tweet realizado por Musk el día 1 de Mayo, donde afirmaba que "las acciones de Tesla estaban demasiado altas". Se sospecha que sus declaraciones podrían ser una de las causas de que las propias acciones de Tesla sufrieran una baja en la bolsa de valores del 9,3%. Es de nuestro interés obtener información sobre la sensibilidad del público y su repercusión en la empresa ante la interacción en la mencionada red social.



Obtención, limpieza y estandarización de datos

Obtuvimos los tweets generados con las palabras claves “tesla” y “market” utilizadas en un mismo tweet así como “Elon Musk”, para el período entre el 25 de abril y el 04 de Mayo de 2020, y, mediante la combinación de ambas bases llegamos a una única base de datos a partir de la cual realizar nuestro análisis. Vale destacar que, dadas las restircciones de la api de twitter para la descarga de los mismos, se creo una base balanceada de 100 tweets diarios. Los datos obtenidos de twitter, así como de cualquier red social con la que trabajamos, deben atravesar un proceso de limpieza que permitan extraer información útil, con estructura y contenido. Trabajaremos tanto con números, fechas y textos. El manejo de datos tipo cadenas, son complejos y requiere mucho esfuerzo por lo que consideramos como quitar números y puntuación, evitar palabras como “y”, “pero” y “o”; quitar emojis y cómo separar las oraciones en palabras individuales (tokenization). La tokenización nos permite dividir el texto en las unidades que lo conforman, entendiendo por unidad el elemento más sencillo con significado propio para el análisis en cuestión, en este caso, las palabras. Tras realizar la limpieza y tokenización a cada tweet, se modificó la base generando una columna de palabras por tweet. Lo que permitió obtener de manera más sencilla cuáles eran las palabras más utilizadas por día.



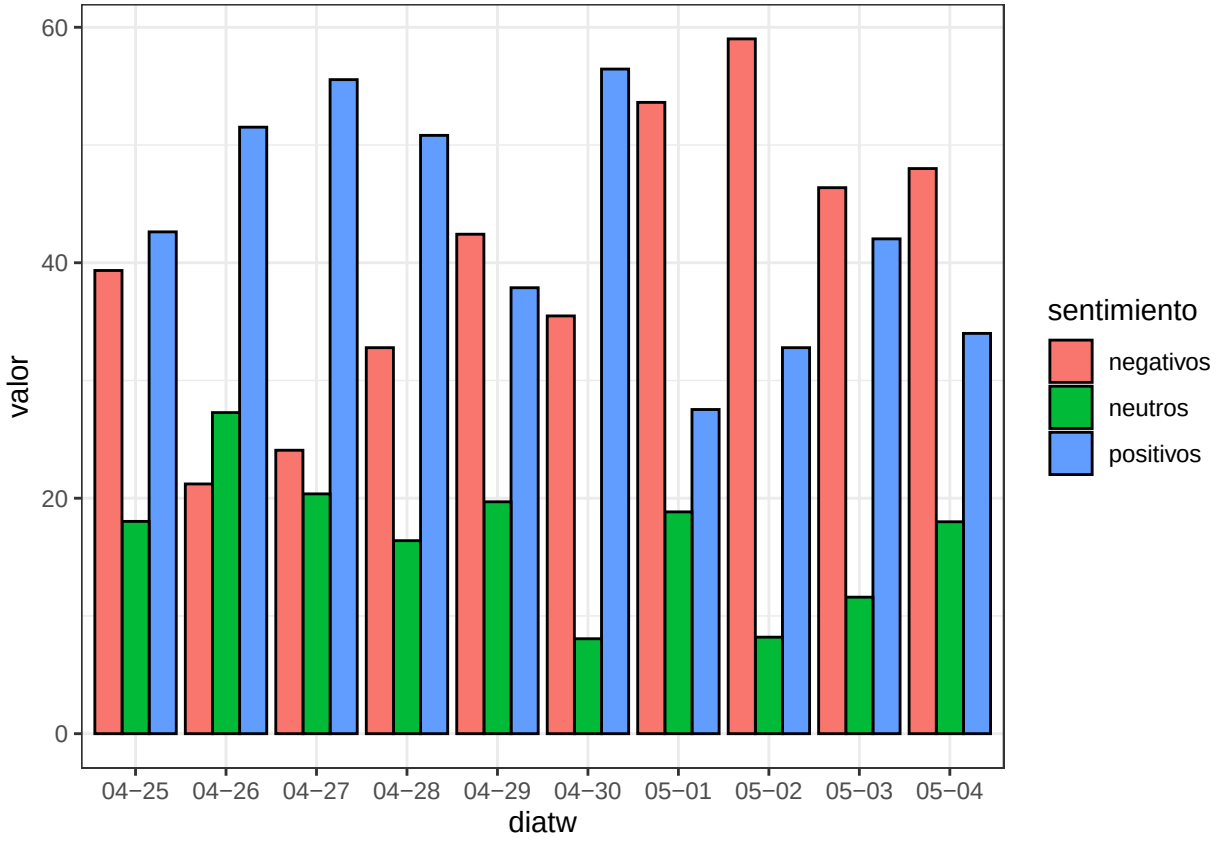
##Análisis de sentimientos

Comenzamos por asignar un sentimiento a cada token haciendo uso del diccionario de R “bing” que clasifica las palabras de forma binaria como positivas o negativas. Para facilitar el cálculo de sentimientos se recodifican los sentimientos de cada Token como +1 para positivo y -1 para negativo.

Esto nos permite realizar el análisis de sentimiento de cada tweet considerando su sentimiento como la suma de los sentimientos de cada una de las palabras que lo forman. Si bien no es la única forma abordar el análisis de sentimientos se consigue un buen equilibrio entre complejidad y resultados.

Como resultado, obtenemos los tweets clasificados como negativos, positivos y neutros por día.

diatw	positivos	neutros	negativos
04-25	42.62295	18.032787	39.34426
04-26	51.51515	27.272727	21.21212
04-27	55.55556	20.370370	24.07407
04-28	50.81967	16.393443	32.78689
04-29	37.87879	19.696970	42.42424
04-30	56.45161	8.064516	35.48387
05-01	27.53623	18.840580	53.62319
05-02	32.78689	8.196721	59.01639
05-03	42.02899	11.594203	46.37681
05-04	34.00000	18.000000	48.00000



Finalmente se ponderaron los token de los distintos tweets en base a aquellas métricas de comportamiento que permanecen públicas y pueden verse a simple vista al visitar cualquier perfil. Retuits: número de ocasiones en las que se retuiteó una publicación. Favoritos: número de ocasiones en las que se marcó como favorito una publicación. Esto nos permite ponderar los token de cada tweets en función de la visibilidad de los mismos

que surje de estas métricas. La ponderación de los token se realiza asignandoles un incremento de un 1% por favorito obtenido por el tweet correspondiente y un incremento de un 4% por cada retweet obtenido.

diatw	status_id	favorite_count	retweet_count	token	sentiment	valor	ValorPonderado
04-25	1254097939208663041	1	0	lucrative	positive	1	1.01
04-25	1254097939208663041	1	0	pure	positive	1	1.01
04-25	1253992672416727040	0	0	bad	negative	-1	-1.00
04-25	1253992672416727040	0	0	reasonable	positive	1	1.00
04-25	1253992672416727040	0	0	tumble	negative	-1	-1.00
04-25	1254081106678132737	6	1	excited	positive	1	1.11

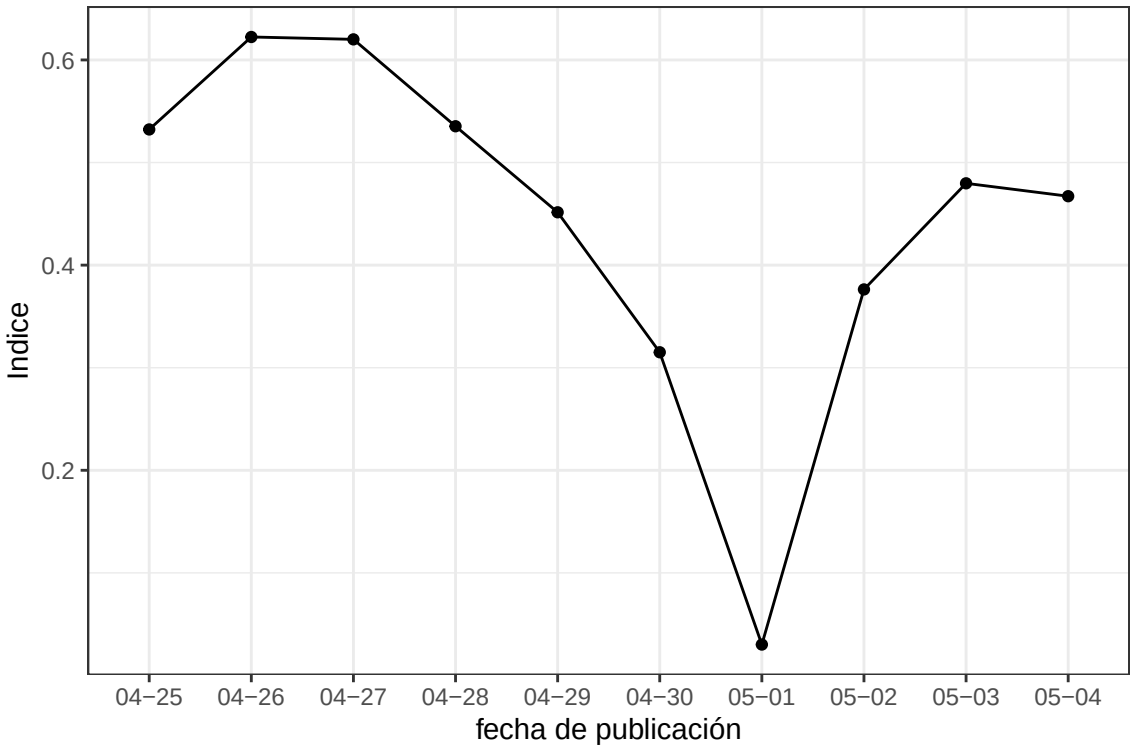
Indice de Sentimiento

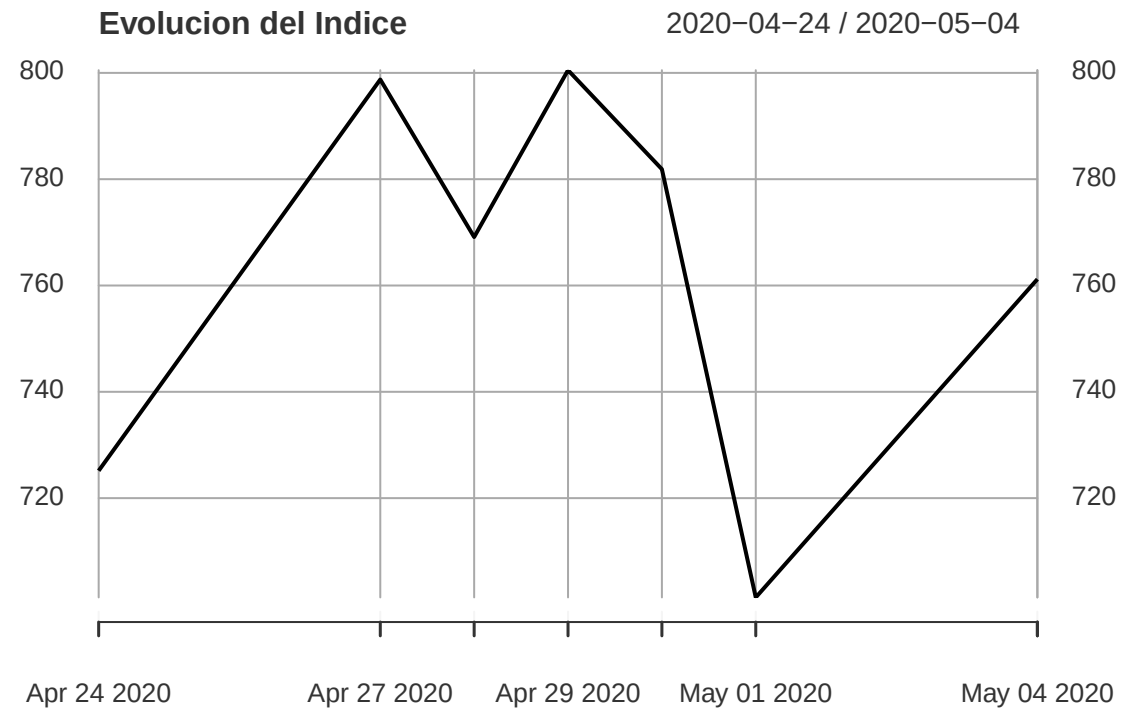
Con los datos ponderados se construye un índice diario calculada a partir de la suma de los sentimentos positivos ponderados sobre los sentimientos totales ponderados por día.

diatw	negative	positive	Indice
04-25	-62.83	71.49	0.5322365
04-26	-50.00	82.41	0.6223850
04-27	-34.27	55.93	0.6200665
04-28	-65.08	74.99	0.5353752
04-29	-93.74	77.16	0.4514921
04-30	-119.75	55.08	0.3150489
05-01	-2191.23	68.08	0.0301331
05-02	-73.56	44.38	0.3762930
05-03	-74.42	68.62	0.4797260
05-04	-71.43	62.63	0.4671789

Conclusión

Habiendo realizado los procedimientos necesarios para la correcta extracción y preparación de los datos, podemos pasar a una etapa de observación y análisis de los mismos. Aunque los resultados financieros de la empresa pueden estar sujetos a distintos factores, se ve una clara relación entre los sentimientos predominantes por día en twitter y la evolución del índice.





El día 1 de mayo se puede observar con mucha claridad cómo hay un pico decreciente tanto la bolsa de valores como en cuanto a los sentimientos (mayor porcentaje de negativos). Esto puede explicarse por la hipotesis del tweet mencionado en el primer párrafo. En conclusión para este caso se observa una correlación interesante entre el análsiis de sentimiento y el desempeño financiero de la empresa estudiada.