

# Grado en Estadística

---

**Título:** Los efectos de los tuits de Donald Trump sobre el mercado bursátil americano

**Autor:** Francesc Salvador Zuriaga

**Director:** Catalina Bolancé Losilla y Salvador Torra Porras

**Departamento:** Departamento de Econometría, Estadística y Estadística Aplicada (Universidad de Barcelona)

**Convocatoria:** Enero 2021



## Resumen

Las redes sociales se están convirtiendo en una de las principales fuentes de datos en la era del *big data*. Muchos analistas financieros usan Twitter para recopilar datos y obtener información valiosa en el análisis financiero. En los últimos años, perfiles políticos reconocidos internacionalmente como Donald Trump han estado usando Twitter para comunicarse con su público. Los tuits del presidente se consideran señales informativas que pueden influir en los consumidores y afectar las decisiones de los inversores en el mercado de valores. El efecto de estas señales se puede medir coincidiendo con los cambios en el mercado de valores. En este proyecto, se aplican análisis textuales y de sentimientos a los tuits de Trump para monitorizar el estado de ánimo presidencial. Luego, se utiliza un análisis de causalidad de Granger y un algoritmo de máquinas de vectores soporte para investigar la hipótesis de que los estados de ánimo del presidente predicen los rendimientos de diferentes mercados americanos. Los resultados indican que existe una relación estadísticamente significativa positiva durante toda la presidencia entre la proporción de tuits negativos escritos por Donald Trump y la tendencia diaria de los rendimientos del índice Dow Jones para el tercer día bursátil siguiente. Sin embargo, este resultado no implica necesariamente que se obtengan mejores resultados predictivos cuando se agrega esta información al modelo de aprendizaje automático. La conclusión del modelo SVM establece que el modelo es neutral a esta nueva información de los indicadores de opinión. Además, se ha descubierto cómo tuits específicos y de carácter financiero de Donald Trump son capaces de generar tendencias anormales en el mercado de valores durante un período de 10 días.

**Keywords**— Redes sociales, efectos de microblogging, tuits presidenciales, minería de texto, análisis textual, análisis sentimental, análisis de causalidad, estudio de eventos, mercados financieros, predicción de mercados, SVM.

## The Effects of Donald Trump's Tweets on American Stock Market

### Abstract

Social media is becoming one of the main sources of data in the *big data* era. Many financial analysts use Twitter to gather data and obtain meaningful insights in financial analytics. Recently, high-profile, internationally known politicians like Donald Trump have been using Twitter to communicate with the public. The president's tweets are considered to be informative signals that may influence consumers and affect investors' decisions in the stock market. The effect of these signals can be measured by coinciding changes in the stock market. In this paper, textual and sentiment analysis are applied to Trump's tweets in order to monitor the presidential mood. A Granger causality analysis and a Support Vector Machine algorithm are then used to investigate the hypothesis that president's mood states predict several American markets' returns. The results indicate that there is a statistically significant positive relation between the ratio of negative tweets written by Donald Trump and daily tendency of Dow Jones Index returns for the third following trading day. However, this result does not necessarily imply the securing of better predictive results when this information is added to the Machine Learning model. The conclusion of the SVM model states that it is neutral to this new information from the opinion indicators. In addition, it has been found how specific financial based tweets from Donald Trump are able to generate non-normal tendencies in the stock market for the timeframe of 10 days.

**Keywords**— Social Media, microblogging effects, president's tweets, text mining, textual analysis, sentimental analysis, causality analysis, event study, financial markets, stock prediction, SVM.

## Clasificación AMS

62P25 Applications to social sciences

68M10 Internet topics

68T50 Natural language processing

62H20 Measures of association

37M10 Time series analysis

62M20 Prediction

## *Agradecimientos*

*M'agradaria donar les gràcies a la meua família per sempre confiar en mi, inclús més que jo mateix i en especial als meus pares per haver-me donat sempre la oportunitat d'aprendre sense posar-hi límits.*

*També m'agradaria agrair als meus tutors, Salvador Torra i Cati Bolancé per la seva ajuda i implicació durant tot el projecte. Donar les gràcies també a en Ramón Alemany per el suport en la fase inicial del treball.*

*Gràcies Alicia per ensenyar-me i demostrar-me com treballa una estadística de veritat. Agrair efusivament als meus companys Josep, Dani, Miquel, Blanca per tots aquests anys de carrera a Barcelona i a Newcastle.*

*Per últim, agrair també a tots aquells que al comentar-los el meu tema van quedar sorpresos positivament i em van animar al respecte.*

*Sense tots vosaltres això no seria possible.*

# Índice

<b>1. Introducción</b>	<b>6</b>
<b>2. ¿Por qué Donald Trump?: Propuesta económica de Donald Trump</b>	<b>9</b>
<b>3. Revisión Literata</b>	<b>10</b>
3.1. Representación Textual . . . . .	10
3.2. Análisis Textual con Twitter . . . . .	10
3.3. Uso de las redes social y su efecto económico-financiero . . . . .	11
3.4. Machine-Learning aplicado a las finanzas: Máquinas de vectores de soporte . . . . .	12
<b>4. Descripción de los datos</b>	<b>14</b>
4.1. Datos financieros . . . . .	14
4.2. Datos de Twitter . . . . .	18
<b>5. Análisis de Sentimientos</b>	<b>23</b>
5.1. Pre-procesamiento de los tuits para el Análisis de Sentimientos . . . . .	23
5.2. Wordclouds . . . . .	24
5.3. Diccionario Léxico . . . . .	26
5.4. Cálculo del sentimiento mediante el uso de un diccionario léxico . . . . .	27
5.5. Conclusiones del Análisis Sentimental . . . . .	30
<b>6. Relación entre variables de opinión y variables financieras: Causalidad de Granger</b>	<b>34</b>
6.1. Relación entre la polaridad de los tuits de Trump y el mercado bursátil americano . . . . .	34
6.2. Granger Causality . . . . .	38
6.3. Modelos con variable respuesta binaria: Logit . . . . .	42
6.4. Relación de las variables de opinión con la volatilidad del mercado . . . . .	46
<b>7. Estudio de eventos aplicado a Twitter</b>	<b>48</b>
<b>8. Predicción y Machine Learning: Support Vector Machines</b>	<b>59</b>
8.1. Support Vector Machines . . . . .	59
8.2. Support Vector Machines sin indicadores sentimentales . . . . .	63
8.3. Support Vector Machines con indicadores sentimentales . . . . .	66
8.4. Comparación de los modelos con y sin indicadores sentimentales . . . . .	69
<b>9. Conclusiones</b>	<b>71</b>
<b>A. Código para la extracción de los tuits de Donald Trump</b>	<b>80</b>
<b>B. Descripción de datos financieros</b>	<b>82</b>
B.1. S&P 500 . . . . .	82
B.2. VIX . . . . .	83
<b>C. Causalidad de Granger</b>	<b>84</b>
C.1. S&P 500 . . . . .	84
C.2. VIX . . . . .	84
C.3. Test por años: 2017, 2018, 2019 y 2020 . . . . .	85
<b>D. Estudio de Eventos aplicado a Twitter</b>	<b>91</b>
D.1. Código para el estudio de eventos . . . . .	91
<b>E. SVM</b>	<b>92</b>
E.1. Funciones utilizadas para la creación de los diferentes modelos SVM . . . . .	92
E.2. Ejemplo visual de SVM con dos variables predictoras . . . . .	93
<b>F. Resto del código en R</b>	<b>94</b>

# 1. Introducción

En la última década, Twitter ha experimentado un tremendo crecimiento en todo el mundo, en línea con el éxito de las redes sociales. No es de extrañar que en el campo de la investigación comenzaran a indagar formas de utilizar esta plataforma de *microblogging* para encontrar nuevas aplicaciones en todo tipo de áreas. El seguimiento de sentimientos en línea se ha vuelto notablemente muy popular en el ámbito social y psicológico, pero el rango de aplicación es más amplio. En particular, se empezó a estudiar la posibilidad de predecir los mercados financieros mediante esta información.

La predicción de los mercados financieros siempre ha sido una área activa de investigación durante los últimos años. La Hipótesis de Mercados Eficientes dice que el precio de los valores bursátiles está en gran parte impulsado por información nueva y que, además, este precio sigue un camino aleatorio. Aunque esta hipótesis esta aceptada comúnmente por la comunidad académica como paradigma central que gobierna los mercados financieros, muchos han intentado extraer patrones o rachas en los comportamientos de los mercados basados en estímulos externos.

Muchos de estos trabajos se han basado en intentar predecir el comportamiento del mercado financiero en base a las premisas de la economía conductual (véase *Bollen, Mao y Zeng, 2010* o *Xue Zhang, Fuehres y Gloor, 2012*). Estas premisas se basan en que las emociones y los estados de ánimo de los individuos afectan en su toma de decisiones. De esta forma, se puede crear una relación entre el sentimiento público que reflejamos en nuestras redes sociales y el sentimiento de mercado. Por ejemplo, el trabajo por *Bollen, Mao y Zeng, 2010* encontró que el estado de ánimo definido como “calmado” contenía información predictiva del mercado bursátil americano. Este trabajo no se basa en la aproximación del estado de ánimo de la sociedad en base a una colección enorme de tuits que puedan reflejar su conducta, sino que se ha escogido alguien con la importancia suficiente para que sus tuits sean capaces de afectar la conducta de sus lectores.

Donald Trump, presidente 45<sup>o</sup> de los Estados Unidos de América, es un usuario ávido de la red social Twitter y a diferencia de sus predecesores en el cargo, utiliza activamente esta red social para expresar su visión de los asuntos que afectan al mundo entero. Uno pensaría que Trump usa Twitter para generar apoyo para su campaña, sin embargo, Trump es conocido por usar Twitter, por ejemplo, como una herramienta estratégica para evitar que las empresas estadounidenses trasladen sus operaciones al extranjero. Con más de 50 millones de seguidores en Twitter, Trump (@realDonaldTrump) se ha convertido en el líder mundial más seguido y el mundo de la inversión actualmente monitoriza sus tuits como indicadores de futuras políticas. Esto le da a Trump poder exclusivo para influir en los mercados financieros con solo 280 caracteres.

Tal es el poder que tiene Donald Trump en su *smarthphone* que recientemente, a causa de los sucesos ocurridos el 6 de Enero de 2021, cuando partidarios del presidente Donald Trump, irrumpieron en la sede del Congreso, violando la seguridad y ocupando partes del edificio durante varias horas, en el ya llamado “Asalto al Capitolio”; los dirigentes de la red social Twitter decidieron suspender permanentemente el perfil de Twitter de Donald Trump. La red social que era hasta el momento el canal preferido de comunicación del presidente de Estados Unidos tomó esta decisión ante el riesgo de “incitación a la violencia” que podrían suponer sus tuits. Este singular acontecimiento demuestra el poder de convicción que tiene y las repercusiones que puede tener un tuit de Donald Trump.

En este trabajo se realizará un estudio exhaustivo de los efectos de los tuits de Donald Trump en los mercados bursátiles americanos. Se quiere comprobar si el estado de “humor” del presidente Donald Trump, reflejado en sus tuits, tiene la capacidad de mover el mercado financiero. Este análisis será completo, en el sentido que el objetivo básico del trabajo es encontrar relación estadística para toda la presidencia Trump, utilizando todos los tuits escritos por él mismo durante su mandato. De tal manera si se encuentran indicios de relación estadística entre los tuits de Donald Trump y los rendimientos del mercado americano se estará probando la ineficacia de la hipótesis de mercados eficientes, ya que la información procedente de un tuit de Donald Trump no será incorporada automáticamente en el mercado y existirá un periodo de tiempo en el que se podrá utilizar esta información de forma lucrativa. Además de buscar esta relación en un periodo de tiempo de 4 años, se intentará aplicar la teoría política de ciclos electorales para ver su implicación en esta relación entre datos de redes sociales y redes financieras. Para eso se agruparán los tuits de Donald Trump en sus años naturales y se buscarán relaciones con los datos bursátiles también estructurados por años. Se considera que este trabajo es exhaustivo porque, por otro lado, aparte, del estudio de todos los tuits de Donald Trump que se consideran una aproximación de su estado de ánimo, también se trabajará con un subconjunto de estos tuits, los tuits considerados de carácter financiero-económico. Al trabajar con este subconjunto de tuits se quiere poner el foco en tuits en concreto que se cree que pueden mover el mercado bursátil americano y se estudiarán sus consecuencias palpables en los mercados. El último objetivo de este trabajo es comprobar si realmente se puede sacar partido de los tuits de Donald Trump, es decir, si incorporar información procedente de estos micro-mensajes puede ayudar a predecir si la bolsa subirá o bajará. Por último, se define este estudio como íntegro porque se

trabajará con los valores bursátiles del índice del Dow Jones, considerado el índice americano más importante y que refleja el comportamiento del precio de la acción de las 30 compañías industriales más importantes y representativas de Estados Unidos y el índice Standard & Poor's que recoge las principales 500 empresas americanas de todos los sectores. Al margen de estos dos mercados también se trabajará con un índice de volatilidad del mercado, el VIX, llevando el estudio a otro nivel, ya que adicionalmente del efecto de los tuits sobre el rendimiento bursátil, también se considerarán las repercusiones de las palabras de Trump sobre el riesgo del mercado, medido por su volatilidad.

Una vez se han definido los objetivos de este trabajo, se puede establecer que el estudio se organizará de la siguiente manera: En la siguiente sección se justificará la elección de Donald Trump para este estudio y se presentarán sus principales ideas que son un reflejo de su ideología. En la sección 3 se discutirán los antecedentes de estudios que han utilizado información procedente de las redes sociales para intentar predecir valores bursátiles. A partir de este momento se inicia el proyecto en si mismo con la presentación y análisis básico de los datos financieros y los datos procedentes de Twitter. En la sección 5 se realiza un análisis textual de los tuits del presidente en forma de nubes de palabras y se desempeña un análisis sentimental para poder clasificar cada tuit con un valor de “positivo”, “neutro” o “negativo” y por lo tanto, también se podrá establecer el “humor” de Donald Trump para cada día en base al sentimiento de sus tuits. En la sección 6, se unen los resultados del análisis de sentimientos en forma de indicadores de opinión a los datos procedentes de los mercados financieros y el índice de volatilidad del mercado. Gracias a la combinación de estos datos se puede realizar una prueba de causalidad entre los tuits de Donald Trump y los rendimientos bursátiles. Una herramienta muy común para esto es el análisis de causalidad de Granger, que permite identificar si una serie temporal tiene información predictiva sobre otra o no. A continuación, nos centraremos en los tuits que se catalogan como tuits financieros escritos por Trump, para poder poner el foco en hechos concretos y ver sus consecuencias palpables en los mercados. También se intentará dar una explicación político-económica a estas relaciones. En la sección 8, se incorporarán las conclusiones procedentes del análisis de causalidad a un algoritmo de Machine Learning para dilucidar si se pueden utilizar la información procedente de los tuits para predecir los comportamientos diarios del mercado del Dow Jones. Finalmente, se presentarán las principales conclusiones, limitaciones y futuras líneas de investigación sugeridas.

Se quiere destacar el carácter progresivo de este estudio, es decir que la información de cada apartado es necesaria para el desarrollo del siguiente. En el sentido de que en el trabajo se verá como las conclusiones obtenidas del análisis textual de los tuits nos permiten definir y conocer la forma con la que se comunica Donald Trump, una forma de comunicación única en él y que el hecho de poderla aprender nos permitirá, en el mismo apartado, la aplicación de un análisis de sentimientos adecuado para la persona que él es. Seguidamente, las conclusiones del análisis de sentimientos y su síntesis en índices de opinión nos permitirá obtener valores de opinión para cada día de su presidencia. Gracias a estos valores diarios, se podrá investigar si estas series temporales poseen información predictiva de los rendimientos bursátiles utilizando el análisis de causalidad. Además, sera debido al análisis de causalidad, que se podrá utilizar sus conclusiones para construir modelos de inteligencia artificial con las variables adecuadas para intentar predecir el comportamiento diario del mercado. Es muy importante entender este hilo conductor a lo largo del trabajo para poder comprender que cada parte del trabajo es importante, pero que es la suma de todas las secciones y su relación entre ellas, lo que permite obtener este análisis completo de los tuits del presidente Donald Trump.

La motivación de este proyecto es la combinación de métodos estadísticos y contexto económico que garantiza este trabajo. La modernidad de los conceptos estadísticos básicos para este trabajo como el procesamiento del lenguaje natural, reflejado en los tuits de un individuo y a la vez poder aplicarlos con cierto sentido en la técnicas de predicción de tendencias en los mercados era una idea que me llamaba mucho la atención y que me ha hecho decidir por este tema. De la misma manera, a nivel personal, me atrae mucho la idea de poder aplicar conceptos estadísticos a ciencias sociales, es decir, que las conclusiones estadísticas se puedan aplicar en la vida real. Este trabajo no va a tener repercusiones directas sobre las acciones de nadie, pero se puede ver una aplicación de la estadística en las ciencias sociales al detalle, en redes sociales, que nos permite sacar conclusiones que afectan al mundo real, es decir a la bolsa. No quiero olvidarme tampoco, de las pequeñas pinceladas de doctrina política, ciencia a la que estaba totalmente ligada la economía hasta hace 100 años, que me ha permitido pintar este proyecto.

Por último, se quiere destacar que uno de los principales rasgos de este proyecto es el alto nivel de conocimientos nuevos que implica. Dado que durante el grado no se han impartido los conocimientos necesarios para afrontar este trabajo, como son el dominio de las técnicas de minería de textos y el procesamiento del lenguaje natural más su implementación, tampoco se han tratado los análisis de causalidad, ni los estudios de eventos. Así mismo, no se han utilizado algoritmos de Machine Learning con datos procedentes de un análisis textual a este nivel. Por eso se ha dedicado mucho tiempo de este trabajo a la búsqueda de recursos, principalmente online para dichos fines. Aunque la aplicación de todos los recursos empleados no aparezca directamente en este proyecto, se encuentran en las referencia al final de la memoria. Paralelamente al aprendizaje de todos estos nuevos conceptos, se han aplicado diferentes métodos aprendidos para la creación de una base de datos con información real procedente de Twitter y datos de

carácter financiero procedentes de los mercados americanos, así como el análisis de estos. Para la realización de este proyecto se ha utilizado el software libre *R* y se ha redactado gracias al procesador de texto de latex, *Overleaf*.

Para aliviar la lectura de esta memoria, a lo largo de la misma solo se presentarán los resultados finales, evitando incluir ningún análisis o procedimiento no definitivo, ya que podrían alejar al lector del seguimiento de este escrito.

## 2. ¿Por qué Donald Trump?: Propuesta económica de Donald Trump

Ante el crecimiento exponencial del uso de las redes sociales, uno se puede preguntar hasta donde pueden llegar a repercutir los mensajes mostrados en las redes sociales y cuales son sus efectos en las acciones de la gente. Además, las redes sociales se pueden utilizar como un medidor del estado de animo de la gente, la cual canaliza sus deseos o sus lamentaciones a través de sus *smartphones*. Tanta importancia se le da hoy en día a las redes sociales, que diferentes estudios se han realizado para ver la relación entre el “humor” de los usuarios y los movimientos del mercado bursátil, relacionando el estado de animo de los usuarios con sus acciones en los mercados. Por ejemplo, si las redes sociales nos indican un gran sentimiento optimista en los mensajes de sus usuarios, es fácilmente entendible que este optimismo mostrado en redes se transforme en un optimismo en el mercado financiero que propicie demanda de productos financieros y, como resultado, un aumento de sus precios. En nuestro caso, es prácticamente imposible realizar un estudio de este tipo, ya que para poder tener una buena aproximación del estado de “humor” de un país a través de sus comentarios en redes sociales, son necesarios millones de tuits y, desafortunadamente, Twitter no dispone de bases de datos de tal envergadura. Por lo que si mi objetivo es encontrar relación entre los tuits y la bolsa, y no tenía disponible una cantidad suficiente de tuits para establecer el estado de ánimo de los usuarios a partir de sus tuits, necesitaba de alguien con la importancia suficiente para que su opinión pueda mover mercados y además que utilice Twitter asiduamente. No me puedo imaginar a alguien que cumpla mejor estas dos premisas que Donald Trump, presidente de los Estados Unidos de América (EE.UU.), usuario recurrente y siempre polémico de dicha red social. Por lo tanto, si de alguien vale la pena hacer un seguimiento de sus tuits y analizar su repercusión en el mercado de valores, ese es Donald Trump. También, se ha de tener en cuenta, el poder de movilización que tiene el presidente de EE.UU., rasgo necesario para que sus mensajes puedan tener repercusiones en la bolsa, mediante el aliento de sus seguidores hacia las dinámicas que el presente en su tuit. Donald Trump, conocido por infundir este poder en sus fanáticos, dispone de este poder de convicción.

Uno de los mayores placeres del presidente Donald Trump, lo ha contado él mismo, es lanzar un tuit y ver cómo crece y marca la agenda de los medios de comunicación como nunca conseguiría el mejor comunicado de prensa. Pero las consecuencias de sus micro-mensajes van mucho más allá. Son cada vez mas frecuentes y en algunos casos pueden mover mercados, en especial si hablan de la guerra comercial con China o la política monetaria de la Reserva Federal. (*Navarro, 2019*)

Donald Trump asumió la presidencia de los Estados Unidos de América el primero de enero de 2017. Su visión de la economía se basa en el nacionalismo, lo que significa un mercantilismo y proteccionismo en las relaciones internacionales, definidos por la protección de las empresas americanas en el territorio nacional para intentar reducir las exportaciones de productos del extranjero, y de esta manera mantener las fábricas y puestos de trabajo en suelo americano. En otros campos el cree firmemente en la limitación de la inmigración, disuadir a las empresas nacionales de su des-localización hacia otro lugar con mano de obra más barata, des-regularizar las actividades económicas para favorecer la competitividad de las empresas americanas en los mercados internacionales y bajar los impuestos para los grandes empresarios. En esta misma línea, la administración Trump ha retirado a los Estados Unidos de algunas organizaciones internacionales con acuerdos de libre comercio. Por otro lado, si que ha iniciado negociaciones para acuerdos comerciales de libre comercio con agentes económicos como la Unión Europea, la NAFTA o otros acuerdos bilaterales con países como China. Un famoso ejemplo de las políticas mercantilistas de Trump son la imposición de un arancel a la importación de acero del 25 % o un arancel a la importación de aluminio del 15 % que Trump estableció en 2018 con México, Canadá y la UE. (*Colonescu, 2018*)

¿Cómo reaccionarán los mercados financieros al mercantilismo y proteccionismo propios de la administración Trump y en particular cuál será el efecto sobre el índice Industrial del Dow Jones? ¿Serán sus tuits, sin filtro, y siempre polémicos capaz de influir en sus lectores hasta el punto de que estos acaben moviendo hacia arriba o hacia bajo los mercados bursátiles de Estados Unidos?

## 3. Revisión Literatura

### 3.1. Representación Textual

Existen tres formas primarias de representar texto en un análisis textual estadístico. Estas son N-grama, modelo de espacio vectorial y flujo de caracteres. La primera de estas técnicas, representaciones N-grama, se ha utilizado durante décadas y plantea el método más simple y directo de representar texto basado en palabras o conteos de secuencias de caracteres. Los modelos de espacio vectorial son una forma más compleja y reciente de desarrollar, estos incluyen la implementación “word2vec”. Finalmente, el flujo de caracteres es la técnica más reciente en ser desarrollada en este campo.

Las representaciones N-grama se basan en caracteres simples o conteos de secuencias de palabras. Este tipo de técnicas utilizan un texto completo y se extrae cada carácter o secuencia de palabras de longitud  $n$  para formar un diccionario de palabras, a lo que denominamos grama. La forma en la que extraemos los gramas se tiene que adaptar al ámbito que estamos estudiando y al objetivo que tenemos en mente. Por ejemplo, en el estudio del lenguaje natural podríamos construir los n-gramas sobre la base de distintos tipos de elementos como por ejemplo fonemas, sílabas, letras, palabras. Algunos sistemas procesan las cadenas de texto eliminando los espacios, otros no. En casi todos los casos, los signos de puntuación se eliminan durante el preproceso. Como muestra, el texto “El rápido gato marrón saltó sobre el perezoso perro” tiene las siguientes características en formato de 5-grama: “El rápido gato marrón saltó”, “rápido gato marrón saltó sobre”, “gato marrón saltó sobre el”, “marrón saltó sobre el perezoso” y “saltó sobre el perezoso perro”. Cada frase en el cuerpo del texto se puede transformar como vector basándose en el conteo de las veces que cada frase del diccionario aparece en el texto (véase *Cavnan y Trenkle, 1994*).

Las técnicas de espacio vectorial requieren una parte substancial de texto ya filtrado. Cada palabra del nuevo texto se clasifica en un único vector (“word2vec”), colecciones de palabras se pueden agregar usando la suma y normalización de matrices creadas con estos vectores (mirar *Sutskever y col., 2013*). La relación espacial entre las palabras después se analiza como se describe en *Mikolov y col., 2013*. Por ejemplo, el vector de “reina” se podría calcular a partir de los vectores para “rey”, “hombre” y “mujer” de la siguiente manera: “*rey*” – “*hombre*” + “*mujer*” = “*reina*”. El principal inconveniente de este tipo de representación vectorial es la necesidad de generar un representación global sobre el lenguaje que podría llevar a malinterpretar el uso de la lengua en situaciones de campos muy específicos, donde el significado de las palabras sea ligeramente diferente.

El modelo de flujo de caracteres en minería de texto requiere un diccionario de caracteres validos para aprender de un texto. Esto permite que esta técnica funcione correctamente, cuando se le aporta un set de datos de aprendizaje adecuado, en lenguajes tan dispares como el chino y el inglés. Esto supone un gran avance sobre otros métodos previos que eran muy rígidos respecto al lenguaje para el que estaban diseñados. Hasta el momento, esta técnica muestra un poder de predicción superior, aunque con un coste computacionales superior. (*Xiang Zhang, Zhao y LeCun, 2015*)

Para la realización de este estudio se ha decidido trabajar con uni-gramas y bi-gramas para el análisis textual y con un diccionario léxico para el análisis sentimental como se verá en su redacción.

### 3.2. Análisis Textual con Twitter

Al realizar un análisis sentimental con la ruidosa y sesgada información de Twitter, se ha encontrado que un modelo con nivel de clasificación múltiple puede ofrecer predicciones más robustas y precisas en bases de datos complicadas, *Barbosa y Feng, 2010*. En este modelo los tuits primero se clasifican en base a su objetividad o subjetividad. De esta manera, se puede eliminar parte del ruido de la información que simplemente manifiesta hechos para centrar-se en los tuits que expresan claramente el estado de ánimo del usuario. A partir de aquí, estos se pueden clasificar con mejor precisión, ya que la mayoría de los tuits con carencia de sentimiento que hubiesen causado dificultades han sido eliminados y eficazmente clasificados como neutrales. Este enfoque puede ser particularmente ventajoso para evaluar tuits sobre el mercado bursátil ya que tanto los tuits subjetivos como objetivos pueden aportar información sobre el mercado, pero estos requerirían métodos de procesamiento dispares.

También ha habido varios planteamientos basados en el aprendizaje profundo (Deep learning)<sup>1</sup> para la clasificación de sentimientos en Twitter que se han especializado en explicar los casos donde hay relativamente poca información

---

<sup>1</sup>Los modelos computacionales de Deep Learning imitan las características arquitecturales del sistema nervioso, permitiendo que dentro del sistema global haya redes de unidades de proceso que se especialicen en la detección de determinadas características ocultas en los datos. Este enfoque ha permitido mejores resultados en tareas de percepción computacional, si las comparamos con las redes monolíticas de neuronas artificiales.

disponible en el texto con un máximo de ciento-cuarenta caracteres. Estos estudios, *Tang y col., 2014* y *Barbosa y Feng, 2010*, han encontrado que la combinación de una meta-data especialmente escogida y rasgos textuales específicos con análisis más tradicionales como n-gramas pueden proporcionar un modelo de clasificación más preciso que estos métodos por sí solos. En particular, abordar la existencia de palabras en mayúsculas, emociones y palabras estiradas como: “queeeeee” o “guayyyy” pueden mejorar la clasificación. Cualquier análisis sentimental con redes neuronales profundas se deberían tener en consideración considerando el excelente resultado que han demostrado en los recientes años.

### 3.3. Uso de las redes social y su efecto económico-financiero

Medir el sentimiento del mercado bursátil a partir de la información derivada de las redes sociales y buscar el efecto del análisis sentimental en los mercados financieros es bastante reciente, pero no es un tema nuevo en la literatura de este tema. Diferentes autores han investigado el efecto de las redes sociales como Twitter en los mercados financieros, agrupando datos de no solo un usuario, sino de diferentes usuarios de las redes sociales. Por ejemplo, *Ranco y col., 2015* ya utilizaron esta metodología para encontrar un pequeño pero significativo efecto en el corto plazo de los sentimientos en el valor de las acciones de las 30 compañías del índice industrial del Dow Jones (DJIA)<sup>2</sup>.

En la misma línea, por lo que se refiere a los efectos a corto plazo, *Papaioannou y col., 2013* usaron información de Twitter para modelizar y predecir la fluctuaciones diarias del tipo de cambio Euro/US Dólar. Finalmente observaron que cada vez más y más estudios desafiaban la hipótesis de los mercados eficientes<sup>3</sup> y sugirieron incluir la información procedente de las redes sociales en los modelos de series temporales para predecir los tipos de cambio. A diferencia de otros estudios, este extrajo indicadores específicos procedentes de los tuits, en vez de valorar el sentimiento u opinión dentro de la red social. El estudio concluyó que la información procedente de las redes mejoró la capacidad de ciertos modelos de predecir en el corto plazo, los tipos de cambio intradiarios entre el euro y el US Dólar.

*Bollen, Mao y Zeng, 2010* fueron más generales en su estudio: ¿Puede el sentimiento general de la sociedad afectar a la economía? Estos autores utilizaron información de Twitter usando las plataformas como Opinion-Finder y Google-Profile of Mood States, dos herramientas de minería de textos para determinar el estado de ánimo de la gente a partir de los tuits. El primer instrumento solo facilita una escala de positivo o negativo, mientras que el segundo instrumento mide el estado de ánimo en diferentes dimensiones. Como el primer artículo mencionado, *Bollen, Mao y Zeng, 2010* usó el índice del DJIA para medir los cambios en el mercado, pero solo se basó en los precios de cierre en vez de en los rendimientos diarios. El resultado también mostró que incluir las variables de ánimo en los modelos de series temporales mejoró las predicciones a corto plazo de los precios.

*Xue Zhang, Fuehres y Gloor, 2012* utilizó una muestra de tuits coleccionados en un periodo de 6 meses para el mismo propósito que otros muchos autores como *Bollen, Mao y Zeng, 2010* discutieron de la misma manera que se ha visto antes. Concretamente, para determinar una correlación entre el análisis sentimental procedente de Twitter y algunos indicadores agregados de variables del mercado. A parte del índice del DJIA, *Xue Zhang, Fuehres y Gloor, 2012* también consideró el NASDAQ<sup>4</sup>, S&P 500<sup>5</sup>, como también índices que miden la volatilidad del mercado como el índice VIX<sup>6</sup>. Estos autores encontraron una correlación negativa entre la emoción general (sin importar si esta era negativa o positiva) y los precios de mercado agregados, también encontraron una relación positiva con el índice de volatilidad del mercado. Una carencia de este estudio parece ser el limitado léxico que los autores utilizaron para la calibración del estado de ánimo.

---

<sup>2</sup>El Dow Jones Industrial Average es el índice bursátil de referencia de la bolsa de valores de Nueva York. Este índice refleja la evolución de las 30 empresas industriales con mayor capitalización bursátil que cotizan en este mercado.

<sup>3</sup>La hipótesis del mercado eficiente es una teoría que dice que el precio actual de un activo en el mercado refleja toda la información disponible que existe (histórica, pública y privada), por lo tanto cualquier noticia o evento futuro que pueda afectar a la cotización de un activo, hará que el precio se ajuste tan rápido, que sea imposible obtener un beneficio económico del mismo.

<sup>4</sup>El Nasdaq es el acrónimo de National Association of Securities Dealers Automated Quotation y es la segunda bolsa de valores electrónica automatizada más grande de Estados Unidos. Se caracteriza por comprender las empresas de alta tecnología en electrónica, informática, telecomunicaciones, biotecnología, etc.

<sup>5</sup>El S&P 500 es un índice bursátil que recopila las 500 empresas más grandes de Estados Unidos. De su elaboración se encarga la empresa Standard & Poor's, y de ahí vienen las siglas de “S&P”. La cifra 500 se refiere a las 500 empresas que lo forman.

<sup>6</sup>El índice VIX es el índice de volatilidad (Volatility Index), que mide la volatilidad de las opciones del mercado en el índice S&P500. El valor del precio del índice VIX refleja una medida de volatilidad en términos porcentuales, por lo que su precio se limita a entre 0 y 100.

En la Tabla 1 se presenta un resumen de la literatura más importante que cubre la relación entre la información precedente de Twitter y los rendimientos bursátiles. La columna de horizonte temporal denota el periodo temporal utilizado para el estudio, como también su duración. La columna de datos contiene la frecuencia de los datos, el tamaños del data-set, como el tipo de información financiera. En la columna de metodología el lector puede ver si se ha realizado un análisis sentimental previo y el tipo de relación modelada. Finalmente, se recogen las principales conclusiones de cada estudio.

### 3.4. Machine-Learning aplicado a las finanzas: Máquinas de vectores de soporte

La búsqueda de modelos para predecir los precios de los mercados financieros sigue siendo un tema muy investigado, a pesar de los importantes desafíos relacionados. Los precios de los activos financieros son no lineales, dinámicos y caóticos. Por tanto, son series temporales difíciles de predecir. Entre las últimas técnicas, los modelos de aprendizaje automático son algunas de las más investigados. En concreto, estas técnicas se aplican para la predicción de valores del mercado financiero. Los modelos de predicción más utilizados incluyen, entre otros, máquinas de vectores de soporte, en inglés *Support Vector Machines* (SVM), y redes neuronales. Mientras que las redes neuronales son modelos que intentan minimizar el error de clasificación dentro de los datos de entrenamiento, las SVM pueden cometer errores de clasificación dentro de los datos de entrenamiento para minimizar el error general en los datos de test.

La técnica de Machine-Learning que se utilizará en este trabajo son las SVM. En el trabajo de *Fung, J. X. Yu y Lam, 2002* se utiliza un análisis de regresión para identificar tendencias en los precios mientras que se utiliza un estudio de SVM sobre artículos de medios de comunicación para realizar una clasificación binaria en dos categorías predeterminadas: subidas y bajadas de valores bursátiles. De su investigación, usando 350.000 noticias financieras simularon una estrategia de compra y venta de activos basada en la clasificación de su SVM, los autores mostraron que su técnica de SVM era ligeramente provechosa económicamente.

*Mittermayer, 2004*, también uso las SVM en su investigación para encontrar una técnica de trading óptima. Su estudio se centro en encontrar los limites donde decidir si comprar o vender una acción, basándose en un sistema de clasificación de 3 opciones. Encontró que los beneficios se pueden maximizar mediante la compra o venta de valores, sacando provecho de estos cuando tenían una subida del 1% o una bajada del 3%. Este método mejora al trading aleatorio, ya que aumenta el beneficio respecto a este en un 11% de media.

Más recientemente, *Min e Y.-C. Lee, 2005* explica un modelo de predicción basado en SVM para predecir la tendencia de los mercados de valores. En este trabajo, se muestra que la clasificación por SVM mejora a las redes neuronales en los problemas de predicciones de tendencia. *Dunis y col., 2012*, muestra que es posible predecir algunos periodos del IBEX-35 bajo diferentes supuestos de aversión al riesgo usando clasificadores SVM. Además, se utiliza un algoritmo genérico para optimizar la selección de inputs y los parámetros del modelo SVM. Este metodología se aplico para tratar de averiguar si el mercado subiría o bajaría para el día siguiente usando los índices FTSE 100 y FTSE 200.<sup>7</sup> Los resultados experimentales indicaron que el método utilizado mejoró a técnicas más clásicas como los modelos ARIMA<sup>8</sup> o predictores Bayesianos<sup>9</sup>.

---

<sup>7</sup>El FTSE 100/200 es el índice bursátil de referencia de la Bolsa de Valores de Londres. Está compuesto por las 100/200 compañías de mayor capitalización bursátil del Reino Unido y es indicador del rendimiento financiero de las empresas reguladas por la ley de empresas del Reino Unido.

<sup>8</sup>En estadística y econometría, en particular en series temporales, un modelo autorregresivo integrado de promedio móvil o ARIMA es un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Se trata de un modelo dinámico de series temporales, es decir, las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes.

<sup>9</sup>La inferencia bayesiana es un tipo de inferencia estadística en la que las evidencias u observaciones se emplean para actualizar o inferir la probabilidad de que una hipótesis pueda ser cierta. Hoy en día, uno de los campos de aplicación es en la teoría de la decisión, visión artificial y reconocimiento de patrones.

Autores		Temporalidad	Datos	Metodología	Resultados
Bollen, Mao y Zeng, 2010	10 meses, 2008	Diarios; 9.9M tuits; DJIA	Diccionario Léxico; Regresión Lineal + SOFNN <sup>7</sup>	El estado de animo “calmado” del mercado tiene poder predictivo sobre los rendimientos del DJIA de 1 a 5 días por adelantado.	
Mao, Counts y Bollen, 2011	15 meses, 2010-2011	Diarios y semanales; Twitter, Google Search, Encuestas, Noticias financieras; DJIA , VIX, oro	Sin análisis de sentimiento, Regresión Lineal	Muestra que el sentimiento de los diferentes medios de comunicación esta correlacionado con los rendimientos diarios del DJIA y con el VIX. También encuentra relación significativa del número de tuits optimistas con los rendimientos bursátiles para retardos de 1 y 2 días. Los datos de encuestas y noticias financieras no son significativos.	
Xue Zhang, Fuehres y Gloor, 2012	5 meses, 2009	Diarios; 5.5M tuits; DJIA, S&P 500, NASDAQ	Sin análisis de sentimiento, Correlación	Las palabras con más significación para los rendimientos del día siguiente fueron palabras como “worry”(preocupación), “fear”(miedo), “anxious”(ansioso) y “hope”(esperanza). Los autores encontraron que la combinación de palabras negativas tenia un poder de predicción superior que las positivas en la volatilidad. Existe una relación positiva entre el número de palabras emocionales y la volatilidad del siguiente día, pero negativa con los rendimientos del día siguiente.	
Mittal y Goel, 2012	7 meses, 2009	Diarios; 475M tuits; DJIA	Diccionario Léxico; Regresión Lineal + SOFNN	Los datos de Twitter pueden capturar el estado de animo general y los estados de “happy”(feliz) y “calm”(calmado) retrasados 3-4 días tienen relación con el DJIA . El método SOFNN con información sentimental mejora a otros métodos.	
Ni y col., 2019	No especificado	Diarios; Tuits; rendimientos bursátiles	Diccionario Léxico; Regresión Lineal	El sentimiento de Twitter tiene correlación con los rendimientos de mercado. Se proponen diferentes estrategias de trading incorporando la información sentimental.	
Smailović y col., 2013	9 meses, 2011	Diarios, 0.15M tuits; 8 empresas	Machine Learning, Regresión Lineal	Cambios en los sentimientos positivos tienen capacidad predictiva importante sobre los rendimientos de las empresas, especialmente en grandes variaciones de precios o caídas importantes. Los tuits clasificados como neutrales pueden añadir información cuando se modelan los rendimientos bursátiles.	
Y. Yu, Duan y Cao, 2013	3 meses, 2011	Diarios, 0.05M publicaciones de compañías, fórums y Twitter; 824 empresas	Diccionario Léxico; Regresión de Panel <sup>8</sup>	Los medios de comunicación convencionales y sociales tienen poder predictivo para los rendimientos bursátiles, dicho efecto es superior en las redes sociales. El efecto es positivo sobre el riesgo en blogs y Twitter. El efecto respecto a los foros tiene una correlación negativa con los rendimientos.	
Sprenger y col., 2014	6 meses, 2010	Diarios, 0.25M tuits, empresas.	Machine Learning; Regresión de Panel	Relación significativa entre el optimismo de Twitter y los rendimientos bursátiles. La variabilidad en el estado de animo de la gente se asocia con un incremento de volumen de trading.	
Mao, Counts y Bollen, 2015	36 meses, 2010-2012	Diarios; 0.31M mensajes de Twitter y Google Search; DJIA , S&P 500, Russell 1000 y 2000 <sup>9</sup>	Sin análisis de sentimiento, Regresión Lineal	Twitter mejora a “búsquedas de Google” para estimar el sentimiento del mercado. El optimismo de Twitter puede predecir los rendimientos de índices de Canadá, Reino Unido y los Estados Unidos. Los precios vuelven a sus orígenes después de una semana.	

<sup>7</sup> SOFNN es un tipo de algoritmo de machine-learning basado en una red neuronal auto-organizada.

<sup>8</sup> Un conjunto de datos de panel recoge observaciones sobre múltiples fenómenos a lo largo de determinados períodos. La dimensión temporal enriquece la estructura de los datos y es capaz de aportar información que no aparece en un único corte.

<sup>9</sup> El Russell 1000/2000 es un índice bursátil compuesto por 1000/2000 compañías estadounidenses de baja capitalización.

## 4. Descripción de los datos

En el estudio se investigará la relación entre los valores de mercado y la información de Twitter para el periodo que comprende la administración Trump. El inicio de los datos coincide con el ascenso a la presidencia de Trump el 17 de enero de 2017, semana donde fue nombrado el 45<sup>o</sup> presidente de los Estados Unidos como sucesor de Barack Obama, hasta el 6 de noviembre de 2020, fecha de las elecciones presidenciales del país. Unas elecciones presidenciales perdidas por Donald Trump, en favor del candidato demócrata Joe Biden. El resultado de estas elecciones ha creado mucho revuelo por un supuesto fraude con el que Trump ha intentado invalidar sus resultados. La selección de este periodo temporal favorece evitar que toda la polémica actual sobre la legitimidad de las elecciones provocada por los tuits de Donald Trump afecten a los resultados de este trabajo. Los detalles de los dos tipos de datos se especificarán en el resto de esta sección.

### 4.1. Datos financieros

Este análisis se realizará en base a 3 índices económicos. Primeramente, el índice industrial del Dow Jones, seguidamente el índice Standard & Poor's 500, también conocido como S&P 500 y finalmente, el VIX que es el código del oficialmente llamado Chicago Board Options Exchange Market Volatility Index, que mide la volatilidad del mercado S&P 500. La información financiera se recogerá por un periodo de 4 años, des de el 17 de Enero 2017 al 6 de Noviembre de 2020, gracias al archivo histórico de la plataforma Yahoo Finance.<sup>10</sup>

Hay dos factores que han hecho relacionar los tuits del presidente Trump con el índice DJIA. Por un lado, a diferencia de otros índices, el DJIA está compuesto por 30 de las compañías mas grandes del mercado integrado americano y afectan a negocios en todo el mundo, por lo tanto, estas capturan totalmente las políticas de la administración americana. La mayoría de estas compañías provienen de sectores diferentes como farmacéuticas, bancos, finanzas o software. De este modo, algunas decisiones como la imposición de un arancel sobre el acero puede que no tengan consecuencias significativas directas en el índice DJIA. Sin embargo, a medio plazo los efectos podrían verterse hacia otros sectores si incluidos en el índice DJIA, ya que otros países afectados podrían tomar represalias que afectarán a una gama de productos más amplia. Por otro lado, también ha sido el índice que más ha mencionado Donald Trump en sus tuits.

Por otro lado, el índice S&P 500 recoge la capitalización bursátil de las 500 empresas más grandes de Estados Unidos que cotizan en la bolsa y captura aproximadamente el 80 % de toda la capitalización de mercado en Estados Unidos. Aquí reside la importancia de este índice, ya que en este índice se puede encontrar empresas de sectores muy distintos como sector de consumo no-cíclico, bancos, empresas tecnológicas etc. La ventaja de este índice, también utilizado en la literatura, es que comprende mayor parte del mercado global estadounidense que el DJIA, ya que hay autores que comentan que la utilización solo del DJIA podría conllevar un sesgo hacia el sector industrial, ya que históricamente el DJIA ha sido un índice de empresas industriales. Se cree que el análisis conjunto de estos dos índices es una muestra fiel y completa del mercado integrado americano.

Finalmente, teniendo en cuenta que los dos índices anteriores son índices que miden el comportamiento de las empresas en los mercados, es decir, sus rentabilidades, se ha querido añadir un índice que no mida rentabilidades del mercado, sino la propia volatilidad del mercado, es decir, los momentos de mayor miedo del mercado, donde más vendedores deciden poner su títulos en venta por ejemplo. Esta información nos permitirá añadir una dimensión más al estudio, en forma de ver el papel de los tuits de Donald Trump en el miedo de los inversores, miedo representado en la volatilidad del mercado, gracias al índice VIX.

La información que se obtiene a partir de la fuente Yahoo Finance relacionada con cada índice comprende la siguiente información:

- **Index:** Variable indicadora que muestra la fecha con el siguiente formato: ["yyyy-mm-dd"]
- **Open:** Variable numérica que mide el valor de apertura al inicio del día del índice.
- **High:** Variable numérica que mide el valor máximo que alcanzó el índice durante ese día.
- **Low:** Variable numérica que mide el valor mínimo que alcanzó el índice durante ese día.
- **Close:** Variable numérica que mide el valor de cierre al finalizar el día del índice.<sup>11</sup>

---

<sup>10</sup>Yahoo! Finance es un servicio de Yahoo! que proporciona información financiera, incluyendo cotizaciones de bolsa, índices bursátiles, comunicados de prensa corporativos y financieros, y foros de discusión para discutir las perspectivas de empresas y la valoración de las mismas.

<sup>11</sup>El valor del cierre del día  $t$  ( $P_t$ ) no tiene porque corresponder con el valor de apertura del día  $t + 1$  ( $P_{t+1}$ ) debido a las subastas de apertura y de cierre que se dan antes de que el mercado este abierto al intercambio de acciones.

- **Volume:** Variable numérica que mide el número de transacciones realizadas durante ese día.<sup>12</sup>
- **Adjusted:** Variable numérica que mide el valor de cierre al finalizar el día del índice ajustado por factores que pueden afectar al precio después del cierre de mercado.<sup>13</sup>

Con la información anterior se calculará una nueva variable para cada índice, conocida como **rendimiento**:

- **Rendimiento:** Variable numérica que mide la variación porcentual entre el precio de cierre del día actual y el precio de cierre del día anterior.

Para los valores bursátiles se calcula el rendimiento diario  $R_t$  de forma simple:

$$R_t = \frac{(P_t - P_{t-1})}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1$$

El rendimiento del día  $t$  ( $R_t$ ) se obtiene como el cociente del valor de cierre del día  $P_t$  sustrayendo el valor de cierre del día  $P_{t-1}$  y el propio valor del día  $P_{t-1}$ , lo que es lo mismo, el cociente de valores de cierre menos 1.

Para calcular estos rendimientos se podría haber hecho también de forma compuesta, es decir aplicando el logaritmo natural a los precios de cierre. No se ha considerado necesario utilizar la fórmula de rentabilidad compuesta ya que el objetivo de nuestro estudio no es maximizar una rentabilidad utilizando todas las herramientas posibles, más capitalizaciones de dinero en el caso de la rentabilidad compuesta. Sino que nuestro objetivo es localizar para cada día de la presidencia Trump, de forma discreta un valor de la serie financiera conjuntamente con un valor procedente de las redes sociales, por lo que no es necesario la utilización del logaritmo natural. Además, con la fórmula simple de rentabilidad es mucho más fácil interpretar los resultados, ya que son incremento porcentuales, esto nos facilitará la comprensión de los efectos inmediatos de los tuits del presidente en el mercado bursátil americano. Adicionalmente, la diferencia entre estos dos métodos para rentabilidades bajas como son las diarias es prácticamente residual. Esta metodología es habitual en trabajos de este tipo como podemos ver en *Ranco y col., 2015*.

A continuación, se procede a hacer un análisis descriptivo individual para cada una de las variables. En primer lugar, se empiezan mostrando, en el siguiente Cuadro 2, los estadísticos básicos para cada una de ellas.

Variable	Min	1r Quartil	Mediana	Media	3r Quartil	Max
DJI.Open	19028	23443	25258	24882	26526	29440
DJI.High	19121	23531	25396	25017	26638	29669
DJI.Low	18214	23340	25079	24737	26415	29407
DJI.Close	18592	23440	25283	24883	26527	29551
DJI.Volume	86150000	265450000	314780000	342339188	384560000	2190810000
DJI.Adjusted	18592	23440	25283	24883	26527	29551
Rendimiento	-0.1292	-0.0033	0.0009	0.0005	0.0054	0.1137

Cuadro 2: Estadísticos descriptivos univariantes de las diferentes variables del DJIA

Una vez realizado el análisis descriptivo inicial de los datos financieros que forman parte del estudio, se procede a realizar un análisis gráfico, mostrando los gráficos más adecuados para cada variable y para qué se está midiendo en cada caso.

En este trabajo nos centraremos en el *close price* y, con su cálculo, más especialmente en los rendimientos. El precio de cierre es muy útil para los inversores para valorar los cambios de *stock* en el tiempo. El precio de cierre puede ser comparado con el precio de cierre del día previo, de un mes antes o incluso de un año antes. Por otro lado, poder utilizar los rendimientos permite una serie de beneficios:

<sup>12</sup>Para el índice VIX al ser un índice artificial que depende de los valores del índice S&P 500 y no un mercado en sí mismo, no existe valor de volumen ya que no existen transacciones.

<sup>13</sup>El precio de una acción generalmente se ve afectado por la oferta y la demanda de los participantes del mercado. Sin embargo, algunas acciones corporativas, como la división de acciones, los dividendos y las ofertas de derechos, afectan el precio de una acción. Los ajustes permiten a los inversores obtener un registro preciso del rendimiento de las acciones.

- La variable que refleja los precios está en una métrica comparable, permitiendo que se evalúe su relación con otras variables. Esta característica será de gran importancia cuando se usen las técnicas de Machine learning.
- La sustracción del precio anterior en el uso del rendimiento permite pasar de una serie seguramente no estacionaria a una serie estacionaria a través de esta primera diferencia. La diferenciación elimina la tendencia.

Por estos motivos, solo se mostrarán los gráficos relacionados con las variables con las que trabajaremos, es decir el *close price* y el *rendimiento*.

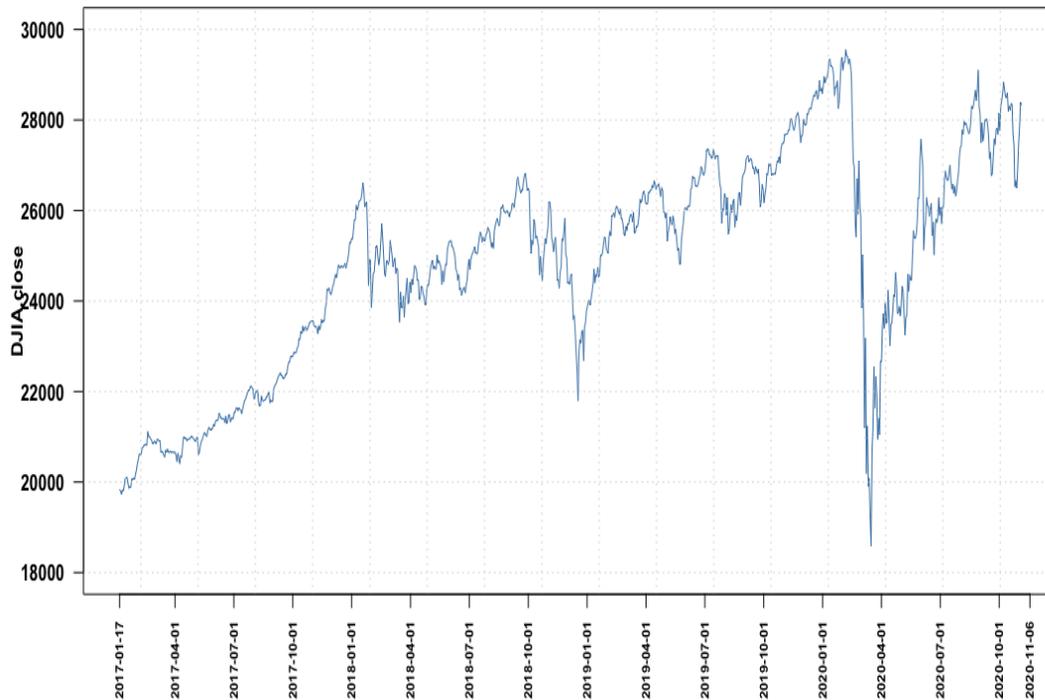


Figura 1: Valores de cierre para el índice DJIA del 17-01-2017 al 06-11-2020

Se puede ver en la Figura 1 la evolución del índice industrial Dow Jones durante la presidencia de Donald Trump, desde su posesión del cargo el 20 de enero de 2017 hasta las elecciones presidenciales de 2020. La evolución de este índice bursátil refleja un aumento sostenido de la valoración de las empresas que lo forman desde el inicio de su mandato hasta el inicio de la crisis del COVID-19<sup>14</sup>. En marzo de 2020, a causa de la crisis sanitaria, económica y social por la propagación del virus COVID-19 se observa un desplome de la bolsa americana que cae de los casi 30.000 puntos a bajar de los 20.000 puntos en un periodo muy corto de tiempo. También se puede destacar una ligera bajada de este índice a finales de 2018, a causa del miedo generalizado a una ralentización económica, las tensiones comerciales con China y las dudas sobre la política monetaria de la Reserva Federal. El mismo Donald Trump llegó a culpar a la Reserva Federal y su máximo dirigente Jerome Powell de “no tener tacto con los mercados” y “no entender las disputas comerciales que sufría el país”. Finalmente, se puede ver que después del desplome acontecido por la crisis del COVID-19, el mercado americano vuelve a conseguir valores parecidos a los de antes de la crisis en julio del mismo año, a pesar de que el virus COVID-19 seguía activo.

<sup>14</sup>La COVID-19 es una pandemia ocasionada por el virus SARS-CoV-2. Se identificó por primera vez en diciembre de 2019 en la ciudad de Wuhan, en la República Popular China. Esta llevó a el colapso del mercado de valores de 2020, un colapso global del mercado que comenzó en febrero de 2020 durante la pandemia de COVID-19 de 2020. El promedio industrial Dow Jones, el índice S&P 500 y el NASDAQ-100 cayeron abruptamente el 27 de febrero durante una de las peores semanas de negociación desde la crisis financiera de 2007-08

A partir de este punto nos centraremos en el análisis de los rendimientos del índice DJIA, anteriormente se ha detallado las ventajas de utilizar esta información en vez de los precios de cierre, y es que de esta forma podemos aislar los rendimientos diarios de la situación general del mercado. En la Figura 2 se puede observar los rendimientos del índice DJIA para el periodo ya comentado.

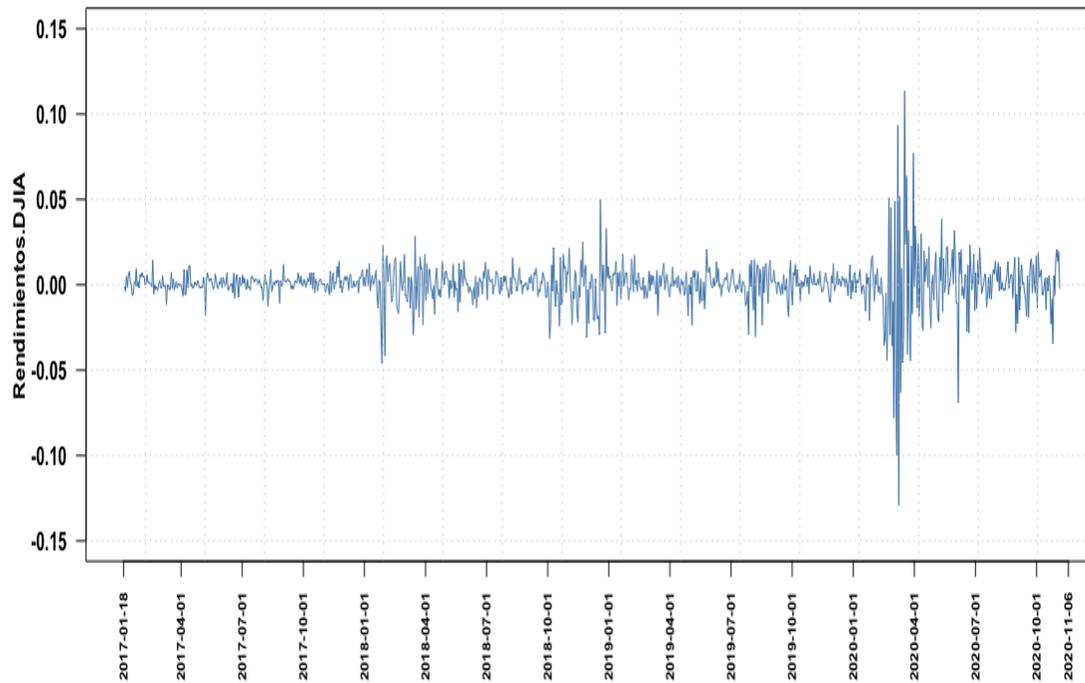


Figura 2: Rendimientos del índice DJIA del 18-01-2017 al 06-11-2020

En la Figura 2 se puede apreciar la volatilidad de este índice americano. Al visualizar los rendimientos diarios en escala porcentual se puede ver cuando los inversores han sufrido más riesgo en sus operaciones. Es muy común asociar el riesgo a la volatilidad de un activo, ya que cuando más volátil es este activo, más posibilidad hay de ganar mucho dinero si escoges buenos momentos para invertir, pero también es más fácil perder mucho dinero si escoges los momentos menos adecuados. A partir de esta información se puede establecer periodos de volatilidad baja entre enero de 2017 y enero de 2018, también entre febrero de 2019 y marzo de 2020 donde los rendimientos están entre el -2% y 2%. Por otro lado periodos de volatilidad media, rendimientos entre el -5% y 5%, que se dieron a cabo entre febrero del 2018 y enero del 2019 y, también a partir de julio del 2020. Para acabar existe un periodo de alta volatilidad, el rendimiento de los activos diarios alcanza cifras superiores al 5% tanto en positivo, como en negativo. Este periodo coincide con el inicio de la crisis del COVID-19 en marzo de 2020 hasta junio del mismo año, con un máximo de volatilidad, una bajada del 13% el 16 de Marzo de 2020, el primer lunes con el estado de emergencia declarado por el presidente Trump. Los mismos gráficos se pueden ver para el mercado S&P 500 y para el índice de volatilidad VIX en el Anexo B.

En este trabajo se quiere estudiar la relación de los datos financieros con los datos procedentes de los tuits de Donald Trump. El presidente tuitea prácticamente todo los días del año, por lo que, como se verá en el próximo apartado de esta sección, se tiene información procedente de Twitter para cada día del año, en cambio, información financiera sólo disponemos para los días bursátiles, días laborables, como se necesita emparejar para cada día la información de las dos fuentes, se ha decidido realizar una interpolación para los fines de semana y otros festivos. Esta interpolación, para poder asignar a cada día información financiera y de la red social nos permite maximizar el uso de la información disponible para este estudio. Con este procedimiento, no se pierde la información sentimental de los fines de semana y festivos.

Esta interpolación se ha realizado gracias a la función *na.approx* del paquete *zoo* de R que permite reemplazar los valores faltantes de los días no bursátiles por una aproximación lineal entre los valores disponible cercanos, es decir, para interpolar los valores del sábado y del domingo se realizaría una interpolación lineal teniendo como referencia los valores del viernes y del lunes y interpolando para los puntos de los días no bursátiles.

A continuación se estudiará el boxplot y el histograma de estos rendimientos, una vez aplicada la interpolación, en la Figura 3.

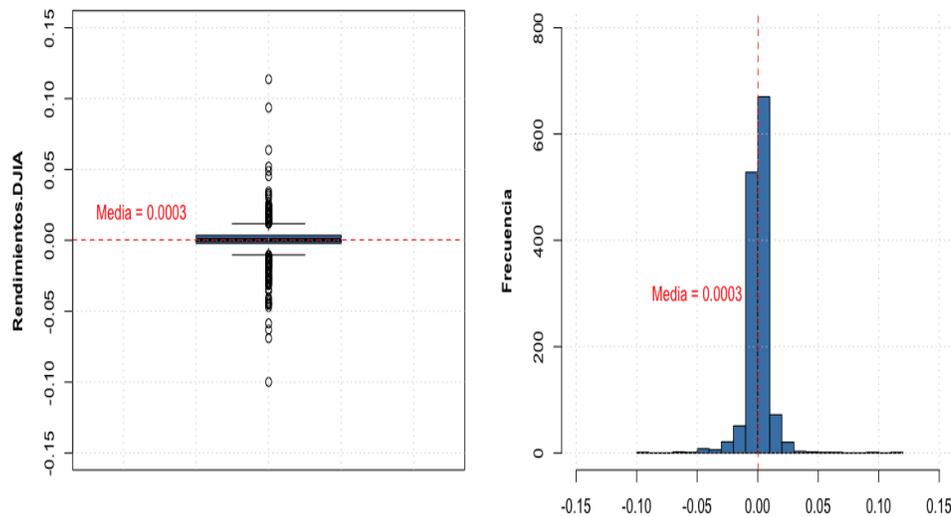


Figura 3: Boxplot e histograma de los rendimientos del índice DJIA del 18-01-2017 al 06-11-2020

Observando este boxplot y el histograma se puede ver como la distribución de los rendimientos recuerda a una distribución normal pero mucho más estrecha, por lo tanto leptocúrtica; es decir, con un coeficiente de curtosis bastante más alto que el valor de 3, típico de la distribución normal. También destacan la presencia importante de valores muy anómalos o extremos que nos hacen intuir que la regla de que el 99.70 % de los valores se encuentran en tres desviaciones típicas no se cumple en este caso. Esto nos puede llevar a pensar que esta distribución es mucho más probable observar valores extremos de lo que se tendría que esperar.

Variable	Min	1r Quartil	Mediana	Media	3r Quartil	Max	Sd.	Curtosis
Rendimientos (interpolado)	-0.0999	-0.0020	0.0003	0.0003	0.0034	0.1137	0.0102	28.56

Cuadro 3: Estadísticos descriptivos univariantes para la variable rendimientos

En el Cuadro 3 se pueden ver los estadísticos descriptivos básicos para la variable rendimientos del índice DJIA. El rendimiento diario más bajo, una vez interpolado es del casi 10 %, por lo que se puede ver como el valor mínimo antes de interpolar se correspondía con el valor de un rendimiento de viernes a lunes, ya que ahora ha desaparecido para un rendimiento mínimo superior. La media de los rendimientos es casi 0 (rendimiento medio del 0.03 %), por lo que se cumple la hipótesis de mercados eficientes. Además, su máximo rendimiento es del 11.37 %. La desviación típica es del 1.02 % y la curtosis de los datos es de 28, muy superior al valor de 3 de la distribución normal.

## 4.2. Datos de Twitter

La segunda fuente de datos proviene de Twitter y consiste en los tuits del presidente Donald Trump desde el inicio de su mandato, la semana el 17 de enero de 2017, hasta las elecciones presidenciales americanas de 2020. Los datos

fueron recolectados mediante la Twitter Search API, donde se realiza una búsqueda para todos los tuits del presidente Donald Trump des de su cuenta de Twitter (@realDonaldTrump). Para este periodo existen un total de 15.483 tuits realizados por el presidente Trump. Los tuits para su análisis han sido proporcionados por el archivo de Twitter de Trump (<http://www.trumptwitterarchive.com/>)

Para su recolección, gracias al archivo de Twitter de Trump, ha sido necesario el uso de diferentes funciones de R. Estas funciones han tenido que obtener los tuits de Donald Trump des de el año 2008 hasta la actualidad, buscando las fechas de los tuits que coincidiesen con dichos años a través de la librería *rtweet*. Los tuits han sido descargados primero en formato JSON<sup>15</sup> y posteriormente convertidos a *data frame* mediante la librería *jsonlite*. Esta metodología esta basada en el trabajo de Michael Kearney que se puede ver en su archivo de GitHub<sup>16</sup> llamado *mkearney*, este código modificado para el uso de este trabajo se puede ver en el Anexo A.

Primero se seleccionarán y se renombrarán las variables más importantes. Se descartan dos variables, el ID de la persona a la cual responde el presidente con su tuit, si el tuit no es una respuesta, esta variable toma el valor NA y una variable lógica que indica “TRUE” si el tuit es un retuit. Como se ha dicho anteriormente, para este estudio solo utilizaremos tuits y no retuits por tanto esta variable es “FALSE” para todos los registros.

Por lo tanto solo se utilizarán las siguientes variables: (una vez renombradas)

- **Fuente:** Variable factor que indica desde que plataforma se hizo el tuit. Niveles: [“Media Studio”, “Twitter Ads”, “Twitter for Android”, “Twitter for iPad”, “Twitter for iPhone”, “Twitter Media Studio”, “Twitter Web App”, “Twitter Web Client”]
- **Fecha:** Variable de fecha que indica en que momento se realizó el tuit, con el siguiente formato: [“yyyy-mm-dd”, “hh-mm-ss”]
- **Id:** Variable identificadora del tuit.
- **Texto:** Variable carácter con el tuit escrito por el presidente.
- **rt:** Número de retuits del tuit del presidente.
- **fav:** Número de favoritos del tuit del presidente.

Se empieza analizando la variable **fuente**, esta es una variable categórica y su representación gráfica se realizará mediante un diagrama de barras en la Figura 4.

Se puede observar que prácticamente todos los tuits del presidentes provienen desde la aplicación de Twitter para iPhone (con un 95.71 % de los tuits). De forma más marginal, se puede ver como las otras fuentes para sus tuits serian Twitter for Android y Twitter Media Studio.<sup>17</sup> Estos últimos serian tuits procedentes de su equipo de campaña y por tanto al no estar escritos por Donald Trump se eliminarán de la muestra. De esta forma después de eliminar los tuits que no escribe personalmente Donald Trump, se obtiene un conjunto de 14.074 tuits.

La siguiente variable a estudio, es la variable **fecha**, la cual muestra en que momento se realizo cada tuit. Para poder representar gráficamente esta variable se han agrupado los tuits que hace el presidente por cada mes del año y se gráfica el número de tuits por mes, durante sus casi 4 años de mandato.

Cuando se representa la variable **fecha** en la Figura 5, se puede ver como el volumen de tuits del presidente ha ido aumentando en función que avanzaba su mandato, se podría decir que le ha ido cogiendo el gusto a esto de tuitear durante su paso por la Casa Blanca. Empieza su mandato con unos 125 tuits al mes y va aumentando progresivamente hasta casi los 250 de Noviembre de 2017. A partir de este punto su actividad social se reduce hasta unos 160 tuits por mes. En el inicio de verano de 2018 vuelve a aumentar su actividad a ritmo de unos 300 tuits por mes. En Mayo

---

<sup>15</sup>JSON (acrónimo de JavaScript Object Notation, “notación de objeto de JavaScript”) es un formato de texto sencillo para el intercambio de datos.

<sup>16</sup>GitHub es una plataforma de desarrollo colaborativo de software para alojar proyectos utilizando el sistema de control de versiones Git. El código se almacena de forma pública, aunque también se puede hacer de forma privada, creando una cuenta de pago.

<sup>17</sup>Twitter Media Studio es un panel de control accesible desde la página de Twitter desde donde se pueden gestionar todos los elementos multimedia de tu cuenta de Twitter (imágenes, GIF, videos, etc.) tanto los que han sido tuiteados como los que no. También permite programar tuits, autorizar a otras cuentas a administrar tus contenidos, compartir imágenes, videos y animaciones entre cuentas de manera privada y otras funciones.

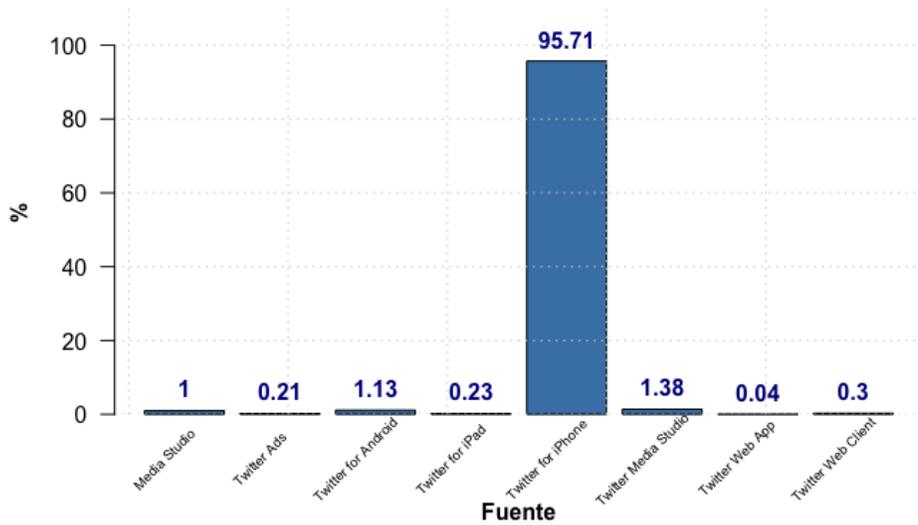


Figura 4: Fuente de procedencia de los tuits del presidente Trump

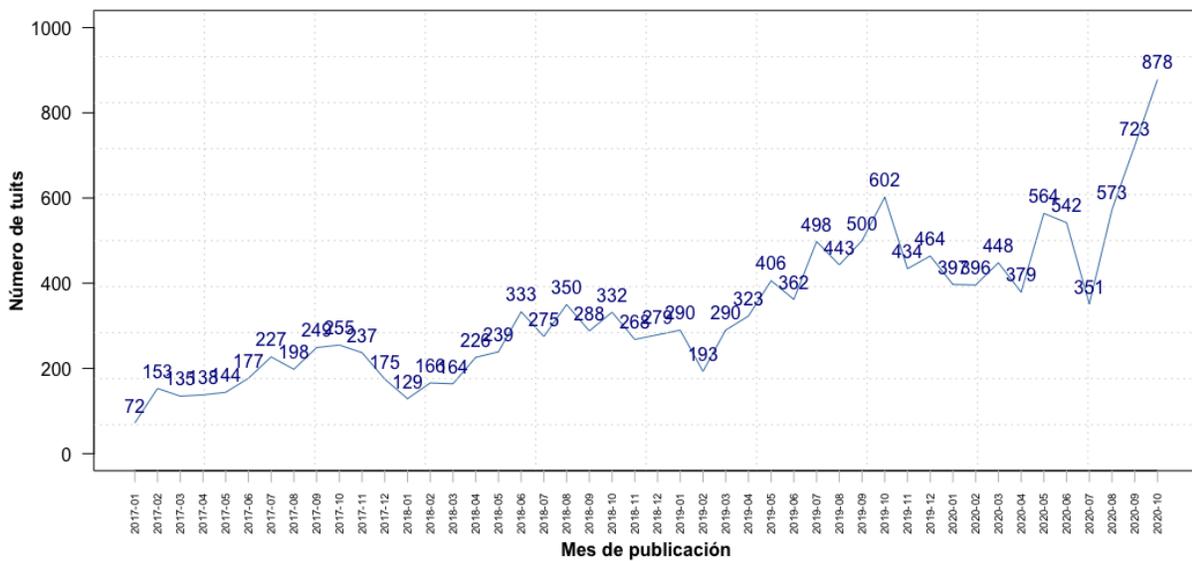


Figura 5: Evolución de los tuits del presidente Trump por mes

de 2019 realiza unos 400 tuits, siendo el mes con más tuits durante su mandato, hasta el momento. No se trata de un caso aislado, sino de una tendencia, ya que a partir de este momento no bajará de la cifra de 400 tuits al mes, prácticamente 13 al día (sin contar retuits), sino que al contrario, habrán meses como Octubre de 2019 y Mayo y Julio de 2020 que superará los 550 tuits por mes. A partir de verano de 2020 y seguramente a causa de la cercanía de las siguiente elecciones presidenciales americanas su actividad en Twitter se dispara hasta 723 y 878 tuits por mes

en los dos últimos meses donde se han contabilizado datos (Septiembre y Octubre de 2020). Gracias a esta variable, se puede ver como el presidente Donald Trump ha ido aumentando su actividad en la red social Twitter durante su mandato, llegando casi a multiplicar por 7 su actividad desde el inicio al final de su mandato.

Variable	Min	1r Cuartil	Mediana	Media	3r Cuartil	Max	Sd.
Retuit(RT)	0	13.564	18.978	21.762	26.345	414.644	14.421,45
Favorito (FAV)	0	59.287	82.330	95.526	123.510	1.884.649	66.422,90

Cuadro 4: Estadísticos descriptivos univariantes para las variables rt y fav

Si nos fijamos en el Cuadro 4, se puede ver los descriptivos básicos para las variables que recogen el número de RT y de FAV que recibe cada tuit del presidente. Se observa que el mínimo para los dos variables es de 0, esto se debe a que han sido tuits borrados por Twitter o borrados por el mismo presidente a posteriori, y por lo tanto, no queda recogida la información sobre retuits y favoritos. De media cada tuit del presidente recibe unos 21.762 retuits y unos 95.526 favoritos o *likes*. Si nos centramos en la mediana, se puede ver que el 50% de los tuits del presidente reciben más de 18.978 retuits y más de 82.330 favoritos o *likes*. El tuit con más de RT ha tenido 414.644 retuits y el tuit con más favoritos o *likes* pasa del millón (1.884.649) de *likes*. Por último la desviación típica de la variable RT es de 14.421, significando que el 67% de sus tuits reciben entre 7341 y 36.183 retuits [ $21.762 \pm 14.421$ ]. La desviación típica de la variable FAV es de 66.423, significando que el 67% de sus tuits reciben entre 29.103 y 161.949 favoritos o *likes* [ $95.526 \pm 66.423$ ]. Esta información nos muestra la influencia que tiene el presidente Trump en Twitter y como sus mensajes son capaces de llegar a millones de usuarios.

A continuación se va a realizar la representación gráfica de estas variables. Al tratarse de unas variables cuantitativas que se pueden considerar continuas se ha decidido que la mejor manera de visualizar sus distribuciones sería mediante un boxplot. Estos se pueden ver en la Figura 6.

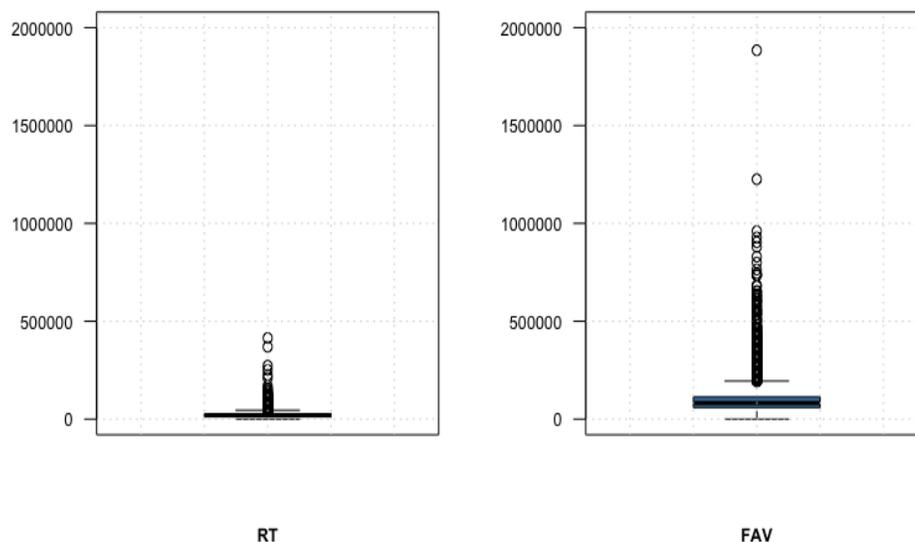


Figura 6: Boxplot del número RT y FAV por tuit del presidente Trump

Mediante la observación de los boxplots se puede concluir como estas dos variables tienen distribuciones con largas colas a la derecha, con los valores de los estadísticos anteriormente comentados. Parece ser que la distribución de las dos variables es muy parecida pero con un factor de escala multiplicativo para pasar de la variable RT a FAV. A continuación se estudia la relación entre las dos variables con la Figura bivalente 7.

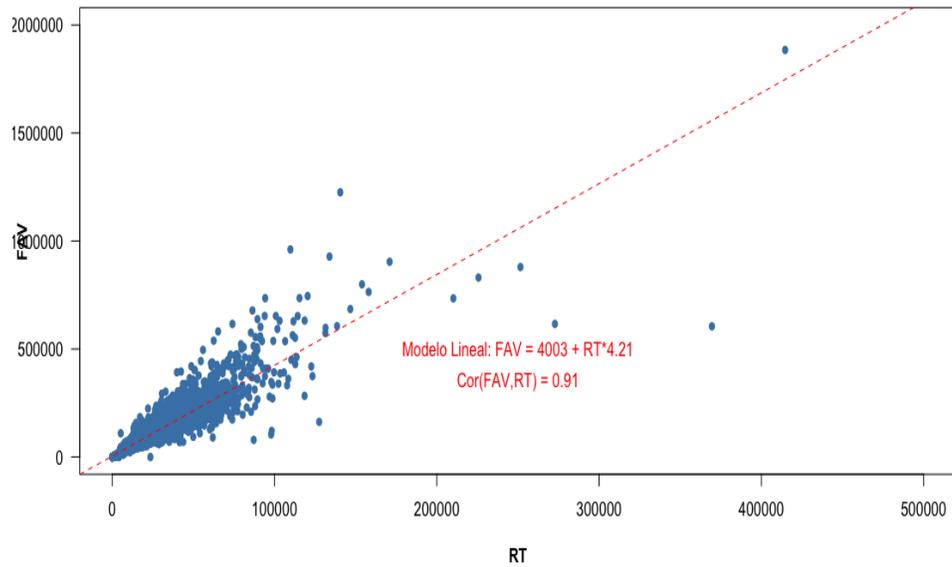


Figura 7: Relación entre el número RT y de FAV por tuit del presidente Trump

En esta Figura 7, donde se visualizan los RT (eje X) y FAV (eje y) por cada tuit del presidente Donald Trump, se puede observar una relación positiva entre las dos variables, con una correlación muy alta (0.91). La relación lineal entre las dos variables es positiva, a más RT que tiene un tuit más FAV o *likes* tendrá. Esta relación queda explicada por el siguiente modelo lineal.  $FAV = 4003 + 4.21 * RT$ . A pesar de esta clara relación, parece ser que existe una frontera en el número máximo de FAV que tendrá un tuit, a partir de los 200.000 RT parece ser que el número de FAV no aumenta en la proporción esperada.

## 5. Análisis de Sentimientos

Dentro de todas las áreas que abarca la minería de textos, este trabajo se centrará en el análisis de sentimientos. El análisis de sentimientos tiene como objetivo detectar la actitud del escritor clasificando los textos u oraciones según su polaridad: positiva, negativa o neutra. Se trata de un área reciente dentro del ámbito de la minería de datos, en la que destacan notablemente los trabajos de *Pang y L. Lee, 2008* y de *Pak y Paroubek, 2010*.

Este trabajo se centrará en la realización de un análisis de sentimientos sobre los tuits del Presidente de los Estados Unidos, Donald Trump. Los tuits a evaluar serán aquellos que emitan una opinión o evaluación. Este tipo de estudios se resuelven categorizando la polaridad de una opinión, determinando si es positiva o negativa respecto al tema estudiado.

Para detectar la polaridad de un tuit destacan dos técnicas: el uso de métodos de aprendizaje computacional (*machine learning*) y el empleo de *diccionarios léxicos*.

El *machine learning* o aprendizaje computacional tiene como objetivo analizar la información de manera automática. Para detectar la polaridad de un tuit es recomendable usar técnicas de aprendizaje automático supervisado que se basan en sets de entrenamiento, los cuales son usados para catalogar el resto de opiniones encontradas en la web, realizando pruebas y luego validándolas (*Campeny, 2019*). Para hacer uso de dicha metodología en este trabajo, se debería crear un modelo a partir de datos ya polarizados y entrenarlo. Sin embargo, si bien existen conjuntos de tuits centrados en diversos ámbitos políticos ya polarizados, actualmente no se dispone de un conjunto de ellos en un contexto similar al que se analiza en este trabajo. Por lo tanto, se ha considerado que no sería adecuado utilizar dichos sets de entrenamiento en este entorno.

En vista de la no viabilidad de la primera opción, se optará por trabajar con la segunda de las comentadas anteriormente, el uso de un diccionario léxico para el análisis de sentimientos. Este método se basa en una lista de palabras con una puntuación determinada y/o categoría emocional. Cabe destacar, que este método es uno de los que ofrece mayor porcentaje de efectividad. Muchos de estos diccionarios en inglés están disponibles de forma gratuita y con un acceso relativamente fácil en Internet.

### 5.1. Pre-procesamiento de los tuits para el Análisis de Sentimientos

Antes de realizar cualquier método de extracción de sentimiento, se suele practicar un pre-procesamiento de la información, de esta forma se podrá mejorar la calidad de la clasificación de sentimiento y se reducirá la complejidad computacional. Los pasos de pre-procesamiento que se tiene la intención de realizar son los siguientes:

- **Stemming or lemmatisation.** Stemming (Derivar) es un procedimiento por el cual se remplazan las palabras por su lexema o raíz. De esta forma palabras como “escribir”, “escritor” o “escribiendo” se asignan como la única palabra “escrib”. Sin embargo, se ha de ir con cuidado cuando se realiza el *stemming*, ya que puede llevar a error. Por ejemplo, un error de *stemming* sería unificar las palabras “experimento” y “experiencia” en el único lexema “exper”. Sobre-derivar podría llevar a problemas de baja precisión, mientras que sub-derivar nos llevaría a obtener sesgo de memoria. El impacto global de la derivación depende de la base de datos y del algoritmo de derivación. El algoritmo más popular de derivación es el de Porter (*Willett, 2006*). Finalmente no se a realizado una *lemmatisation* de las palabras porque el diccionario léxico utilizado no incorpora palabras sin derivar, sino palabras derivadas. Además, se corría el riesgo de cometer el posible error comentado de *stemming*.
- **Eliminación de las stop-words.** Las stop-words son palabras que no aportan información, como conectores, preposiciones y artículos. Estas palabras se eliminan de los tuits cuando se realiza el análisis para que no aparezcan en el momento de seleccionar los términos más usados.
- **Negaciones.** Las negaciones son un factor clave cuando se realiza un análisis de sentimientos a través de tuits. La presencia de negaciones puede invertir el sentido de la oración de manera que, si la oración tenía una polaridad positiva, pase a tener una polaridad negativa. Por ejemplo: “Estos resultados electorales no son buenos”. No obstante, la negación no siempre invierte el sentido de la oración, como, por ejemplo: “No dudo que los intereses sean los más altos del mercado”. Por eso es necesario encontrar patrones gramaticales que identifiquen el sentimiento adecuado en presencia de negaciones.
- **Intensificadores.** Existen términos que reducen o aumentan el sentimiento total del mensaje a analizar. Estos son los intensificadores y las variaciones. Palabras como “muy”, “más” o “bastante”, aumentan la intensidad del mensaje. Por el contrario, palabras como “poco”, “menos” o “casi”, disminuyen dicha intensidad.



(noticia), *fake* (falso), *democrats* (demócratas)<sup>18</sup>. Esta nube de palabras dando mayor peso a la gente, al país, a si mismo (presidente) y a la oposición (demócratas) ya nos podría dar información sobre las prioridades del presidente Donald Trump a la hora de hablar de cara a su país mediante el uso de la red social Twitter.

A partir de esta primera *wordcloud* se puede decir del presidente Trump, que donde él quiere poner más énfasis con sus tuits es en la propia gente, en su propio pueblo. Como cualquier político sabe que él es presidente de los Estados Unidos por que la gente así lo ha votado, por lo tanto, es normal que su palabra más utilizada sea la de *people*. Lo mismo pasaría con la palabra *country*, de el siempre se ha dicho su intención de poner a Estados Unidos por delante de cualquier cosa, por eso que la palabra “país” sea de las más utilizadas. También se puede observar como el hecho de que la palabra “presidente” sea una de las que más escribe (también vemos que destaca su apellido “Trump”), muestra como el mismo se considera una figura muy importante y nunca se olvida de remarcar su rol como “presidente” de los Estados Unidos. Asimismo, el hecho que una de las palabras que aparezcan en este gráfico sea como el llama a la oposición *democrats* (demócratas) ya nos informa que a pesar de estar él como presidente no se olvida en sus tuits de los que forman la oposición. Por último, aparecen las palabras *fake* (falso) y *news* (noticias) que en el siguiente gráfico se explicaran ya que son el bigrama (unión de dos palabras) más utilizado por el presidente Donald Trump.



Figura 9: Wordcloud de los Bigramas de los tuits de Donald Trump

En la Figura 9 se pueden ver la bigramas más usados por el presidente Trump en sus tuits. Como se comentaba antes, la expresión *fake news*<sup>19</sup> (bulos) és la expresión más común de Donald Trump, seguido de: *witch hunt* (caza de brujas), *white house* (casa blanca), *Joe Biden*, *total endorsement* (apoyo total o respaldo total) o *radical left* (izquierda radical).

<sup>18</sup>Aquí el término demócrata no hace referencia a una persona partidaria de la democracia sino a una persona seguidora del partido demócrata de Estados Unidos, partido más progresista y opuesto al republicano, al cual pertenece Donald Trump.

<sup>19</sup>Las noticias falsas, conocidas también con el anglicismo *fake news*, son un tipo de bulo que consiste en un contenido pseudo-periodístico difundido a través de portales de noticias, prensa escrita, radio, televisión y redes sociales y cuyo objetivo es la desinformación.

Un lector con conocimiento previo del tema se podría preguntar porque no aparece uno de los lemas más conocidos y con más exposición mediática, incluso en campos de marketing. El termino MAGA (Make America Great Again) no aparece en ninguna de estas wordclouds porque cuando es utilizado por Donald Trump, siempre lo hace en forma de hashtag (#MAGA) y los hashtag no forman parte del estudio. Aun así si realizáramos el gráfico en forma de nube de palabras de gramas de 4 términos si que aparecería este lema.

Como conclusión ha este apartado y de mucha importancia para el siguiente apartado, es que se han podido identificar una serie de palabras, que son muy usadas por Donald Trump en sus tuits, que a priori y en la mayoría de diccionarios léxicos carecen de valoración, pero que al ser escritas por el presidente de los Estados Unidos y más en concreto por Donald Trump si que se pueden considerar positivas o negativas. Estas palabras son:

- *Border* (Frontera), *China*, *Russia*, *Which Hunt* (Caza de Brujas), *Korea*, *México*, *Wall* (Muro).

Todas estas palabras se han añadido al diccionario léxico, que se explicará en el siguiente apartado, como se puede ver en el código. Se han añadido al diccionario con una connotación negativa equivalente a -1, ya que se ha visto que el presidente Trump utiliza estos términos de forma negativa. Estos términos se ven reflejados en el Cuadro 5 con su puntuación.

Palabra	Puntuación
witch hunt (caza de brujas)	-1.00
wall (muro)	-1.00
border (frontera)	-1.00
China	-1.00
Russia	-1.00
Korea	-1.00
Mexico	-1.00

Cuadro 5: Ejemplos de algunas palabras del diccionario léxico añadidas

A posteriori, se ha visto que añadir estas palabras con esta connotación al diccionario léxico ha permitido clasificar una gran cantidad de tuits, que anteriormente habían sido considerados como neutrales, como negativos. Como resultado se puede decir que el análisis sentimental a realizar después ganará en carácter y expresividad, ya que este se centra en el número de tuits positivos o negativos.

### 5.3. Diccionario Léxico

Como se ha comentado en el apartado anterior, para detectar la polaridad de los tuits se hará uso de diccionarios léxicos.

Los métodos basados en diccionarios (o lexicones, del inglés lexicon) hacen uso de listados de palabras y frases previamente etiquetadas con la polaridad de sentimiento que expresan y, en ocasiones, además con su intensidad o la fuerza de dicho sentimiento. Los textos por clasificar se dividen en unidades más pequeñas, como palabras o frases, y se buscan en los diccionarios de sentimiento. Así, el sentimiento global del texto vendrá dado por algún tipo de función matemática que tenga en cuenta el sentimiento individual de las unidades de trabajo y en base a lo indicado para ellas en el diccionario.

Como uno puede esperar, el papel del diccionario de palabras en las conclusiones del análisis sentimental es vital. La realización del mismo procedimiento de análisis de sentimientos llevará a resultados diferentes según la utilización de que diccionario léxico, incluso se ha visto como para el mismo diccionario léxico los resultados han variado mucho solo por el hecho de añadir sentimiento a 7 palabras, gracias a los resultados de las nubes de palabras. La causa de estos diferentes resultados entre diccionarios léxicos se basa en que muchas veces hay divergencias de opinión a la hora de catalogar palabras como positivas o como negativas. Por ello es muy importante la correcta selección del diccionario.

Existen muchos diccionarios léxicos en inglés, se ha decidido usar la función *sentiment* del paquete de R *sentimentr*. Esta paquete ya incluye diferentes diccionarios y en este trabajo se ha decido trabajar con el diccionario *Combined Jockers & Rinker Polarity Lookup Table*. Este diccionario es la combinación aumentada de dos diccionarios ya asentados como son el de *Jockers & Rinker's (2017)* y *Hu & Liu (2004)* proporcionado por *Rinker y Spina, 2016*. Este diccionario contiene 11710 palabras y se trata de un conjunto de palabras puntuadas entre -1 a 1 según si son

palabras percibidas de manera más negativa o más positiva, este diccionario también incluye 13 expresiones negativas que tienen un valor de entre -1 y -2. Este diccionario ha sido el escogido porque se trata de un diccionario de palabras americano muy completo, más de 11000 palabras, que además también contiene un diccionario de *stop-words* y enfatizadores. Se ha visto en diferentes estudios que este diccionario de R, es adecuado para tratar con textos procedentes de redes sociales (Naldi, 2019). Además, esta función de R, donde se encuentra este diccionario permite la actualización de este, característica muy adecuada para poder adecuar el diccionario a las connotaciones de Donald Trump. En el Cuadro 6 se muestra un ejemplo de palabras puntuadas según el diccionario léxico utilizado.

Palabra	Puntuación
unduly (indebidamente)	-2.00
aberration (aberración)	-0.80
accomplishment (cumplido)	0.50
accomplishments (cumplidos)	1.00
admirably (admirable)	1.00

Cuadro 6: Ejemplos de algunas palabras del diccionario léxico utilizado

Este paquete y más en concreto, esta función de R, también permite añadir un diccionario de 140 enfatizadores. Estas palabras reducen o aumentan el sentimiento total del mensaje a analizar. Algunos de estos intensificadores se pueden ver en el Cuadro 7 con su clasificación en negaciones (1), amplificadores (2), atenuadores (3) o conjunciones adversativas (4).

Los negaciones varían el signo de la polaridad del resto de la frase hasta que se encuentren con un punto en el texto. Realmente lo que hace es elevar a la -1 dicha polaridad, de esta manera si en la frase hay dos negativos la polaridad queda positiva, ya que se considera que dos negaciones en la misma frase, se contrarrestan. Para los amplificadores se incrementa la polaridad de la frase en la que se encuentran por 1 + el valor de peso asignado (entre 0 y 1). Los atenuadores trabajan bajando el signo de la polarización de la palabra. Finalmente, los conjunciones adversativas hacen que baje el grado de polaridad de la frase inmediatamente anterior y aumentan el grado de polaridad de la frase inmediatamente posterior. Más adelante se ve la función utilizada en R explicada para cada uno de sus parámetros.

Palabra	Tipo de enfatizador
absolutely (absolutamente)	2
almost (casi)	3
don't (negador verbal)	1
but (pero)	4
didn't (negador verbal)	1
huge (enorme)	2

Cuadro 7: Ejemplos de algunos enfatizadores del diccionario léxico utilizado

## 5.4. Cálculo del sentimiento mediante el uso de un diccionario léxico

El uso de un diccionario léxico permite calcular el sentimiento para un texto dado gracias a la polaridad de sus palabras en el texto (Turney, 2002). Para este método es necesario un diccionario de palabras con su respectiva polaridad asignada. Hay múltiples ejemplo de léxicos existentes como: Opinion Lexicon, SentiWordNet, AFINN Lexicon, Loughran-McDonald Lexicon. En este trabajo se utilizara el léxico de *Combined Jockers & Rinker Polarity Lookup Table*, ya se han comentado anteriormente sus ventajas.

La metodología para el cálculo del sentimiento se describe en los siguientes pasos:

- **Pre-processing.** El texto (tuit) se somete al proceso de pre-procesamiento descrito en la sección anterior: Eliminación de las stop-words, manejo de las negaciones y tokenización de los tuits.
- **Chequear cada token con su polaridad en el léxico.** Cada palabra de la bolsa de palabras se compara con el diccionario. Si la palabra se encuentra en el diccionario la polaridad de esta palabra se añade a la puntuación sentimental de la frase. Si la palabra no se encuentra en el léxico la polaridad considerada es de cero. Se tienen en cuenta las negaciones, intensificadores, atenuadoras y conjunciones adversativas. Finalmente se calcula la

polaridad de cada tuit en función de la polaridad de las frases que este contiene. Para obtener la polaridad del tuit se suman las polaridades de cada frase que contiene el tuit, se considera que dentro del mismo tuit si que puede haber compensación de polaridades ya que la longitud de un tuit suele ser corta.

- **Calcular el sentimiento final del tuit.** Después de asignar puntuaciones de polaridad a todas las frases del tuit, se asigna un sentimiento final al tuit siendo este positivo si la suma de polaridades es mayor que 0, neutro si esta es 0 o negativo si la suma de puntuaciones es inferior a 0.

Como se ha comentado en la subsección anterior, se ha modificado el diccionario léxico con la incursión de algunos términos. Esto ha sido posible gracias a la función `update_key` del paquete `sentimentr` de R. Esta función permite actualizar y ampliar nuestro léxico inicial, `lexicon::hash_sentiment_jockers_rinker`, con una serie de palabras y su connotación, adaptándolo así al uso de Trump de estas palabras.

```
# Ampliamos el diccionario con los terminos que ha usado mucho
# Trump y tienen una interpretacion diferente por ser el

mykey <- lexicon::hash_sentiment_jockers_rinker
mykey <- update_key(mykey, x = data.frame(x = c("border", "china", "wall", "russia",
      "which_hunt", "korea", "mexico"), y = rep(-1,7),
stringsAsFactors = FALSE))
```

Para calcular el *score* de polaridad de cada tuit se ha utilizado la función `sentiment()` del paquete `sentimentr` de R. Esta función rescata cada frase dentro de cada tuit y da una puntuación de sentimiento a cada frase. En el parámetro `polarity_dt` te permite escoger que diccionario quieres trabajar. Adicionalmente, el parámetro `valence_shifters_dt` permite añadir un diccionario de cambiadores de valencia como son las negaciones “*don't, didn't...*”, intensificadores de opinión “*absolutely*”, atenuadores de opinión “*vastly*”, conjunciones adversativas “*but, however...*”. Esta función ya contiene por defecto un diccionario completo con estos enfatizadores, lo que permite que la posterior clasificación de sentimiento por frase sea más rica y completa. El parámetro `hyphen` hace referencia al funcionamiento de los guiones en esta función. Por defecto, la opción activada en nuestro caso, considera como si el guión no estuviera, es decir la palabra “*sugar-free*” la buscaría como “*sugarfree*”. También se le puede dar un peso para los intensificadores/atenuantes de opinión `amplifier.weight`, este peso entre 0 y 1 hará que el termino polarizado quede multiplicado por 1 más el peso seleccionado (por defecto este valor es de 0.8). También se puede asignar el número de palabras antes `n.before` y después `n.after` de un cambiador de valencia, que serán las que se les aplicará dicho cambiador de valencia. El peso de las conjunciones adversativas sobre la frase en las que se encuentran también puede ser modificado por el parámetro `adversative.weight`. Finalmente, hay un criterio `neutral.nonverb.like` para decidir si asignar o no polaridad a verbos como: “*is*”, “*am*”, “*had*”, “*been*”. Esto es bastante útil en el caso de “*like*” ya que de por si obtendría una polaridad positiva cuando muchas veces actúa como verbo sin polaridad definida. Se ha decidido no considerar palabras como “*like*” con polaridad. En *Rinker y Spinu, 2016* se pueden ver las ecuaciones con las que se ha definido este *score* al detalle.

En nuestro caso se ha utilizado esta función de R con los siguientes parámetros:

```
#sentiment analysis
auto_sent <- sentiment(tweets$text, polarity_dt = mykey,
  valence_shifters_dt = lexicon::hash_valence_shifters,
  hyphen = "",
  amplifier.weight = 0.8, n.before = 5, n.after = 2,
  adversative.weight = 0.25,
  neutral.nonverb.like = FALSE)
```

Se prosigue con la metodología para el cálculo del sentimiento:

- **Calcular el sentimiento final para cada día.** Como nuestro objetivo es comparar el análisis sentimental de los tuits del presidente con los valores diarios en bolsa del Dow Jones, es necesario tener una valoración diaria de la polaridad de los tuits de Trump, para ese día. De esta forma se ha decidido crear 4 indicadores diarios, los cuáles, serán utilizados para relacionarlos con las series bursátiles.

ÍNDICE DE POLARIDAD DIARIO 1: Aporta información sobre la relación relativa entre tuits positivos y tuits negativos para el mismo día. Si este es mayor que 0, en general los tuits del presidente ese día habrán sido positivos, mientras que si el índice es inferior a 0, los tuits del presidente ese día serán mayormente negativos.

$$\text{Índice de polaridad diario} = \frac{\# \text{ Tuits Positivos} - \# \text{ Tuits Negativos}}{\# \text{ Tuits Positivos} + \# \text{ Tuits Negativos}}$$

ÍNDICE DE POLARIDAD DIARIO 2: Aporta información sobre la relación absoluta entre tuits positivos y tuits negativos para el mismo día. Si este es mayor que 0, ese día habrán más tuits positivos que negativos, en caso contrario, habrán más tuits negativos que positivos y si este es 0 habrán los mismos tuits positivos que negativos. Este índice permite valorar diferente un día con sólo 1 tuits positivo, que un día con 15 tuits todos positivos, ya que el Índice de Polaridad 1 los consideraría iguales.

$$\text{Índice de polaridad diario} = \# \text{ Tuits positivos} - \# \text{ Tuits Negativos}$$

RATIO DE TUIITS NEGATIVOS DIARIO: Aporta información sobre el número de tuits negativos sobre el total de tuits con polaridad para el mismo día. Si este es cercano a 1, los tuits del presidente ese día habrán sido la mayoría negativos, mientras que si el índice es cercano a 0, no habrán tuits negativos casi para ese día.

$$\text{Ratio de negatividad diario} = \frac{\# \text{ Tuits Negativos}}{\# \text{ Tuits Positivos} + \# \text{ Tuits Negativos}}$$

RATIO DE TUIITS POSITIVOS DIARIO: Aporta información sobre el número de tuits positivos sobre el total de tuits con polaridad para el mismo día. Si este es cercano a 1, los tuits del presidente ese día habrán sido la mayoría positivos, mientras que si el índice es cercano a 0, no habrán tuits positivos casi para ese día.

$$\text{Ratio de positividad diario} = \frac{\# \text{ Tuits Positivos}}{\# \text{ Tuits Positivos} + \# \text{ Tuits Negativos}}$$

Se considera que la utilización de estos cuatro indicadores, puede mostrar una imagen muy completa del sentimiento del presidente Trump al escribir los tuits. Por un lado se trabaja con un indicador de polaridad diario mostrando si en general el día ha sido sentimentalmente positivo o negativo. Esta sería se puede relacionar con mucho sentido con una serie bursátil que muestre por cada día si la bolsa ha subido o ha bajado respecto al día anterior. Por otro lado los dos ratios de negatividad y positividad se pueden utilizar para estudiar si son los tuits negativos o positivos los que afectan más al mercado de valores americanos.

- **Normalización de los datos.** Excepto para los movimientos bursátiles, todos los índices o ratios sentimentales se normalizaran para poder hacer comparaciones en futuros análisis. Se calculan los *z-scores* mediante la sustracción de la media local y la división entre la desviación estandard. Para la variable X, el *z-score* de  $x_i$ ,  $Z(x_i)$  se calcula de la siguiente manera.

$$Z(x_i) = \frac{x_i - \mu(X)}{\sigma(X)}$$

Una vez finalizado el procedimiento de catalogación de los tuits, se verá una tabla con ejemplos de tuits clasificados según dicho proceso.

Texto	Polaridad
Totally Negative China Virus Reports. Hit it early and hard. Fake News is devastated. They are very bad (and sick!) people!	-3.68
Four people in Nevada viciously robbed and killed by an illegal immigrant who should not have been in our Country. 26 people killed on the Border in a drug and gang related fight. Two large Caravans from Honduras broke into Mexico and are headed our way. We need a powerful Wall!	-2.22
On my way to the #ArmyNavyGame in Philadelphia, Pennsylvania!	0.00
Almost One Million people request tickets for the Saturday Night Rally in Tulsa, Oklahoma!	0.00
Tariffs will bring in FAR MORE wealth to our Country than even a phenomenal deal of the traditional kind. Also, much easier & quicker to do. Our Farmers will do better, faster, and starving nations can now be helped. Waivers on some products will be granted, or go to new source!	2.29
Stock market hits new high with longest winning streak in decades. Great level of confidence and optimism - even before tax plan rollout!	1.56

Cuadro 8: Muestra de tuits clasificados por diccionario léxico

En el Cuadro 8 se pueden ver 6 tuits de ejemplo, con las polaridad asignadas por la función de R. Se puede ver como tuits con palabras como “China”, “Virus”, “Fake News”, “devastated” han sido clasificados como muy negativos. Seguidamente tuits que aparentemente no muestran ninguna opinión como lo es un tuit que habla de que Trump se encuentra de camino a Philadelphia han sido catalogados con una polaridad de 0. Finalmente tuits con palabras como “high”, “winning” o “optimism” han sido valoradas con una polaridad positiva.

Se ha decidido la creación de estos indicadores, en vez de utilizar una serie temporal con la media o otra medida de agregación del sentimiento de los tuits para cada día. Esta decisión se ha tomado para paliar los efectos de la selección del diccionario en los resultados finales de este trabajo, ya que una serie temporal con el sentimiento promedio de los tuits para cada día dependería demasiado de la elección del diccionario léxico. De esta forma, se considera que se está dando más consistencia a los resultados del análisis sentimental. Además se recuerda que el objetivo del trabajo es considerar los efectos de los tuits de Donald Trump en la economía americana para toda la presidencia Trump, para este objetivo no tenemos la necesidad de contabilizar el sentimiento específico de cada tuit, sino su clasificación en positivo o negativo. Como resultado, será la estimación del estado anímico de Donald Trump a partir de sus tuits lo que veremos reflejado en los indicadores escogidos. A pesar de esto, si que se utilizara una variable temporal basada en la puntuaciones de los tuits para cada día en un sub-apartado específico de este trabajo.

## 5.5. Conclusiones del Análisis Sentimental

En este apartado se ha realizado un análisis textual exhaustivo de todos los tuits escritos por Donald Trump durante su presidencia. Después de la realización de un análisis textual preliminar en forma de *Wordclouds* se han podido identificar los términos más usados por el presidente de los Estados Unidos en sus tuits, esto nos ha permitido tener una primera aproximación a qué dice y cómo habla Donald Trump en la plataforma de Twitter. Gracias a estas nubes de palabras se ha podido tunear, esto es modificar, el diccionario léxico a utilizar durante el análisis de sentimientos para adaptar dicho diccionario a la forma de hablar del presidente. Seguidamente, se ha realizado un análisis de sentimiento del presidente Trump, donde se ha cotejado cada palabra de sus tuits en un completo diccionario de palabras americano que contenía una puntuación para cada palabra según su connotación positiva (+), negativa (-) o neutra (0). Por causa de este procedimiento, se ha podido asignar una puntuación a cada tuit del presidente, para finalmente obtener para cada día el número de tuits positivos, neutros y negativos. Finalmente, se han escogido cuatro indicadores sentimentales y se han construido gracias a los datos sobre tuits positivos y negativos para cada día. Debido a este exhaustivo proceso se pueden sacar las siguientes conclusiones:

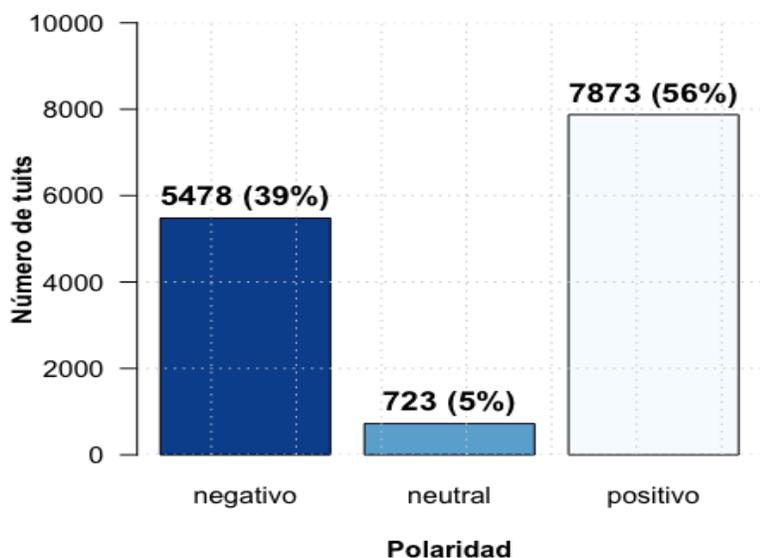


Figura 10: Distribución de los tuits según polaridad

En el Cuadro 10 se puede ver la distribución de los 14.074 tuits del presidente según si han sido clasificados como negativos, neutrales o positivos según las pautas y el algoritmo anteriormente mencionado, es decir, si un tuit tiene una puntuación de 0 es un tuit neutro, si tiene una puntuación inferior a 0 es un tuit negativo y si la puntuación sentimental superior a 0 está etiquetado como positivo. Del total de 14.074 tuits del presidente, 5478 de estos tuits, un 39%, han sido etiquetados como negativos, 723 tuits, un 5%, han sido clasificados como neutros y finalmente un 7873 positivos, es decir, un 56%. Se puede ver como, al contrario de lo podría parecer, la mayoría de los tuits del

presidente Trump tienen una connotación positiva. Esta idea de que los tuits de Donald Trump son mayoritariamente negativos seguramente provenga del sesgo que provocan los medios de comunicación en nuestro pensamiento, ya que estos seguramente solo resalten los tuits negativos del presidente.

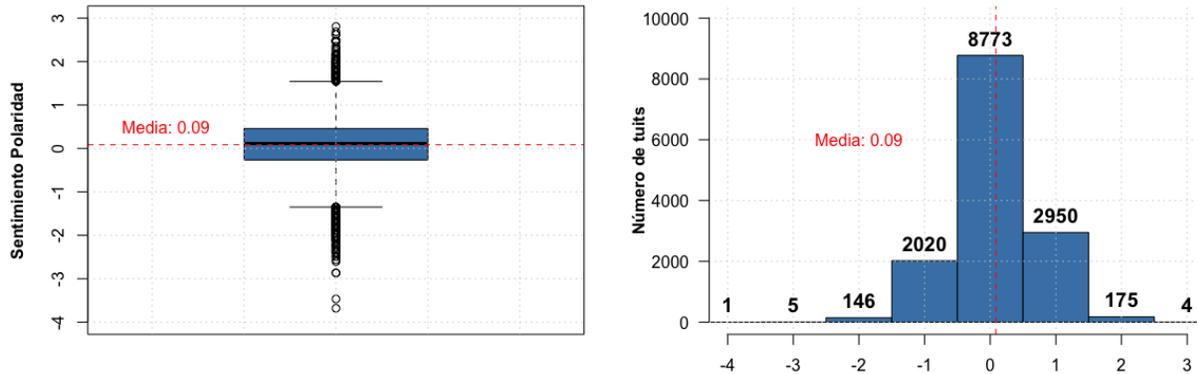


Figura 11: Boxplot e histograma de los tuits según polaridad

Esta información sobre la distribución de polaridad de los tuits, se puede complementar con lo que se ve en la Figura 11, donde se observa el boxplot y el histograma del sentimiento de polaridad de los tuits. Se puede ver como la media de puntuación del sentimiento de los tuits es de 0.09, con más tuits clasificados con un sentimiento mayor que 1 que tuits clasificados con un sentimiento menor a -1. No obstante, los tuits clasificados con un valor absoluto mayor son tuits de connotación negativa ya que en el boxplot vemos como los puntos más extremos en negativo se encuentran un puntuaciones cercanas al -4, en cambio en el extremo opuesto las puntuaciones positivas llegan al +3.

A continuación se detallaran las series temporales referidas a los índices de opinión creados.

Primeramente, se verán los resultados obtenidos por los dos índices de polaridad explicados.

El primer índice de polaridad, que mide el número de tuits positivos menos los negativos para cada día teniendo en cuenta el número de tuits con sentimiento, nos muestra en la Figura 12 como este índice se encuentra casi siempre por encima del 0, por lo tanto, los tuits positivos son más abundantes que los tuits negativos. Sólo hay un periodo donde los tuits negativos son mas abundantes ligeramente que los positivos que es durante el primer trimestre de 2019. Los picos con mayor diferencia, a favor de los positivos, entre positivos y negativos de forma relativa al total de tuits se encuentran en verano de 2017 y Marzo de 2020, con valores cercanos a 0.5, es decir que el número de tuits positivos es un 50 % superior al número de tuits negativos.

El segundo índice de polaridad, que mide la diferencia entre número de tuits positivos y tuits negativos para cada día, se observa en la Figura 13. Se puede ver los valores por meses de este índice, como este se encuentra casi siempre por encima del 0, por lo tanto, los tuits positivos son más abundantes que los tuits negativos. Sólo hay un periodo donde los tuits negativos son mas abundantes ligeramente que los positivos que es durante el primer trimestre de 2019, de la misma manera que se veía en el gráfico anterior. Los picos con mayor diferencia, a favor de los positivos, entre positivos y negativos se encuentran en verano 2017, Septiembre-Octubre de 2018, Marzo de 2020 y Septiembre-Octubre de 2020, sobretodo estos dos últimos periodos. Seguramente a causa del aumento del número total de tuits en los últimos meses que haga que la diferencia relativa entre tuits positivos y negativos sea mayor.

Por último, se verá la evolución de los ratios de negatividad y de positividad a lo largo de estos casi 4 años.

En la Figura 14, se observa como los ratios de negatividad y positividad son muy simétricos entre sí, no al 100 % a causa de los días en que todo son tuits neutros, ya que la construcción de estos así lo propicia. Se puede ver lo que se decía anteriormente como los tuits positivos son mayores que los tuits negativos ya que el ratio de negatividad se

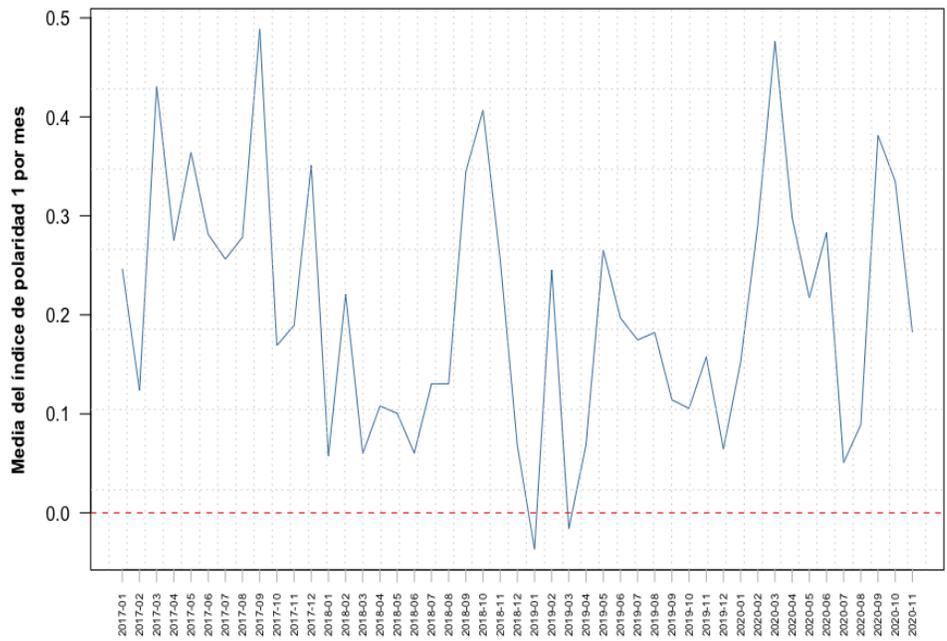


Figura 12: Índice Polaridad 1 por mes

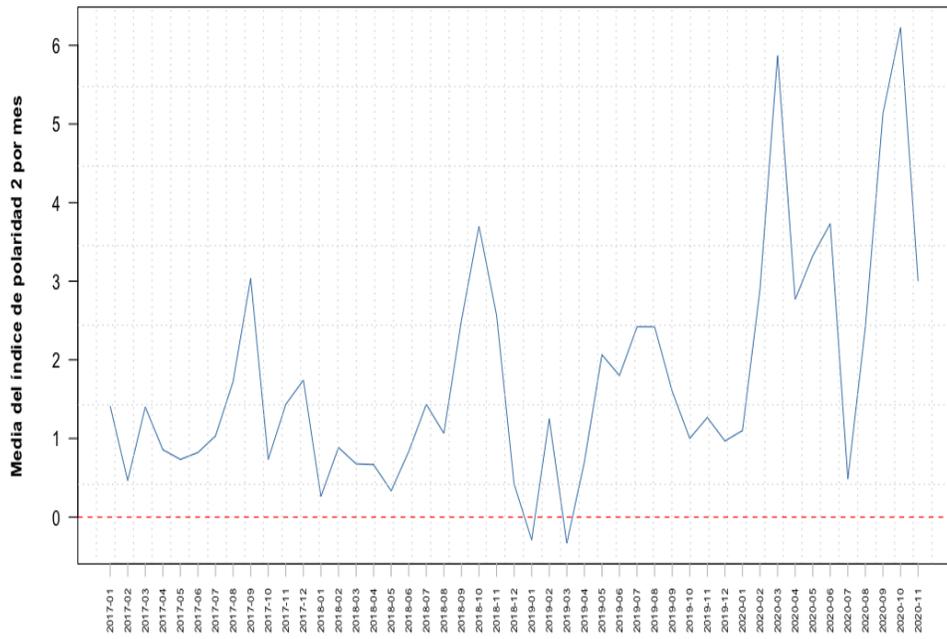


Figura 13: Índice Polaridad 2 por mes

mueve entre 0.25 y 0.50, en cambio, el ratio de positividad se mueve entre 0.50 y 0.75. Los puntos álgidos, es decir, donde mayor diferencia o menor hay entre los ratios se dan para Septiembre de 2017, con un porcentaje muy alto de tuits positivos y muy bajo de negativos. También pasa lo mismo en Octubre de 2018 y Marzo de 2020. Por otro lado, en el primer trimestre de 2019 y sólo en este periodo, es donde parece que los números de tuits negativos y positivos son muy parecidos entre sí.

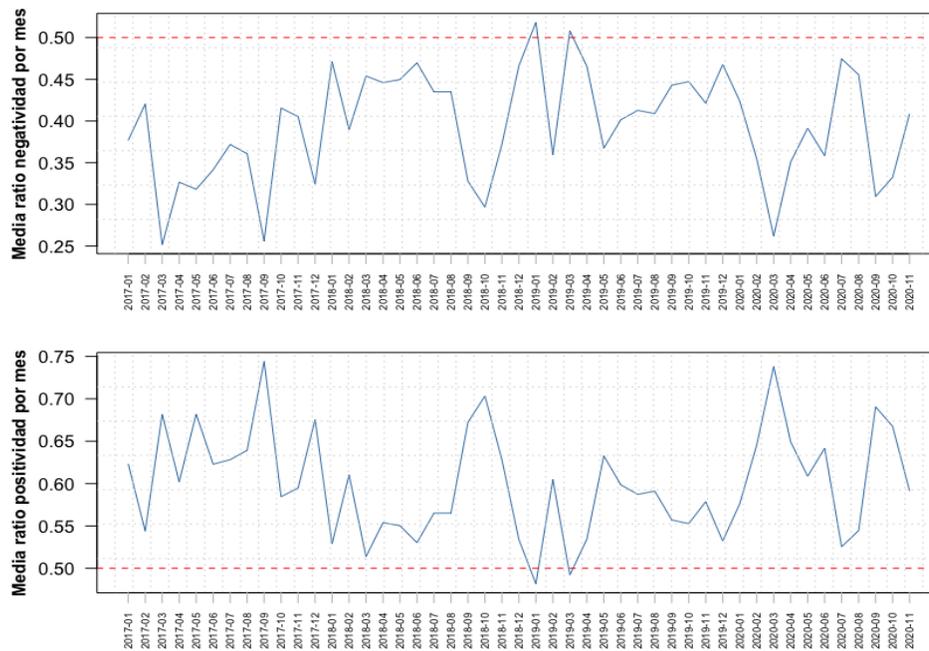


Figura 14: Ratio de Negatividad y de Positividad por mes

En líneas generales para todas las variables sentimentales se pueden ver ideas parecidas. El “humor” de Donald Trump reflejado en sus tuits y resumido en función de estas variables se podría sintetizar como que empieza su mandato con optimismo y buenas intenciones, ya que en sus primeros meses de mandato las variable sentimentales reflejan un gran número de tuits positivos respecto a los negativos. También destaca el alto número de tuits positivos para otoño de 2018, justo antes de un cambio de estado de animo hacia negativo hasta finales de 2019, periodo marcado por las tensiones internas con la creación del muro y las tensiones externas con sus políticas exteriores comerciales hacia la China. A finales de 2019, vuelve a ver una temporada con un estado de ánimo positivo hasta el inicio de la crisis del COVID-19, a partir de cuando su estado de ánimo se vuelve muy volátil.

## 6. Relación entre variables de opinión y variables financieras: Causalidad de Granger

En esta sección del trabajo, se realizara un estudio conjunto de las variables financieras procedentes de los mercados y los indicadores sintéticos extraídos gracias al análisis sentimental de los tuits del presidente Trump. Primeramente, se mostrarán diferentes gráficos con cada indicador creado y los rendimientos bursátiles del DJIA para el mismo periodo y con las dos series temporales escaladas para facilitar su comprensión. Una vez se haya podido observar gráficamente dichas series temporales, se buscará encontrar una relación entre los indicadores sentimentales retrasados hasta 10 días y los rendimientos bursátiles (rendimientos como valores numéricos o como variable binaria que mide si estos han sido positivos o negativos), de esta manera se podrá ver si de alguna forma las subidas y bajadas bursátiles se pueden ver con anterioridad en las variables sentimentales. Así, se podrá intentar dar respuesta a una de las preguntas básicas de este apartado: ¿Tienen los tuits del presidente Trump alguna influencia en el mercado bursátil estadounidense? Los resultados de este análisis se sintetizan en los modelos más adecuados para ser entendidos, como puede ser un modelo Logit para explicar la relación con variables respuesta binarias. Por último, para tener una inspección más completa de los efectos de los tuits del presidente, también se indagará en el efecto que tiene la volatilidad en sus tuits, es decir si hay relación entre los tuits con carácter muy diferente y extremos en valoración para el mismo día y la volatilidad del mercado, basada en el índice VIX de volatilidad del mercado S&P 500.

Para las diferentes tablas que contienen p-valores <sup>20</sup> que aparecerán a partir de este punto se establece que se marcarán los p-valores inferiores a 0.10 con (\*), inferiores a 0.05 con (\*\*), y iguales a 0 con al menos 4 decimales iguales a 0 con (\*\*\*)). El nivel de confianza para los diferentes test será del 90% a no ser que se justifique lo contrario. Se cree que este trabajo es muy ambicioso en sus proyecciones, ya que es muy complicado encontrar significación estadística sostenida del efecto de los tuits de Donald Trump en la economía para un periodo tan largo de tiempo como son 4 años, además de los diferentes factores que han hecho que estos 4 años, no sean un periodo tradicional y normal. Por lo tanto se ha decidido que los resultados de este trabajo se muestren con un nivel de confianza del 90%.

### 6.1. Relación entre la polaridad de los tuits de Trump y el mercado bursátil americano

Una vez se ha realizado un análisis de sentimientos con la muestra de 14.074 tuits del presidente Donald Trump, se intentará entrelazar la información obtenida de dicho análisis con los valores bursátiles americanos. Para poder incorporar la información de opinión a los datos de rendimientos del Dow Jones, se utilizarán los indicadores y ratios anteriormente comentados. El objetivo de esta primera fase es representar gráficamente las diferentes relaciones con las variables de opinión para introducir lo que será luego el análisis estadístico de causalidad entre las variables financieras y las de opinión. Para ello se gráfica cada variable de opinión conjuntamente con los rendimientos del Dow Jones para el mismo periodo, ambas series escaladas, para facilitar la comparación. Para facilitar la observación visual también se han representado no las series de datos como tal, sino la media móvil de 30 días de la serie.<sup>21</sup> Este método de agregación nos permitirá reducir el ruido diario de la series temporales y suavizar las fluctuaciones de plazos cortos para centrarnos las tendencias y los ciclos de plazos largos. Se recuerda que se han interpolado algunos de los rendimientos del Dow Jones para poder asignar a cada día la información de Twitter con los datos financieros, esto pasa principalmente los fines de semana.

Se empieza comparando el índice de polaridad 1, que mide la relación relativa, al número total de tuits con sentimiento, entre tuits positivos y tuits negativos. Por lo tanto este índice antes de escalar, se puede mover entre -1 (todos los tuits son negativos) y 1 (todos los tuits del día son positivos).

Se puede ver en la Figura 15 que aparentemente no hay ninguna relación clara entre las dos series escaladas. Si se mira con más detenimiento parece ser que en momentos como principios de 2017, Octubre del 2018, principios de 2019 y principios de 2020 son los momentos donde el índice de polaridad 1 es más grande, la diferencia entre tuits positivos y negativos es mayor en favor a los tuits positivos, son momentos donde el mercado tiende a bajar o al menos

---

<sup>20</sup>En estadística general y contrastes de hipótesis, el p-valor se define como la probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta. En términos simples, el p-valor ayuda a diferenciar resultados que son producto del azar del muestreo, de resultados que son estadísticamente significativos.

<sup>21</sup>una media móvil es un cálculo utilizado para analizar un conjunto de datos en modo de puntos para crear series de promedios. Así las medias móviles son una lista de números en la cual cada uno es el promedio de un subconjunto de los datos originales. Por ejemplo, si se tiene un conjunto de 100 datos y queremos una media móvil de 25 días, el primer valor de la serie de medias móviles podría ser el promedio de los primeros 25 términos, luego el promedio de los términos 2 al 26, el tercer elemento de los términos 3 al 27 y así, hasta por último el promedio de los últimos 25 números del 76 al 100.

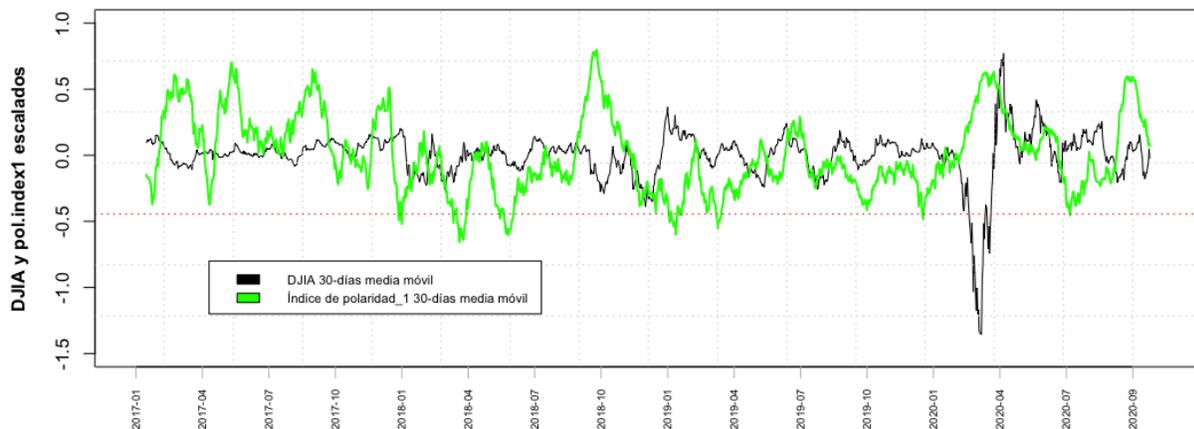


Figura 15: Serie temporal del índice de polaridad 1 y el índice DJIA

a tener un comportamiento en forma de valle. En cambio, existen otros momentos donde parece que la relación es a la inversa, es decir, cuando el índice de polaridad 1 sube, también sube la bolsa o si baja el índice hay una tendencia bajista en la bolsa como puede pasar en Junio de 2018, Septiembre-Octubre de 2019 o Septiembre de 2020. Parece ser que esta relación positiva se da más en periodos más concretos, que la anterior tendencia negativa. Se ha añadido, para facilitar la comprensión de las tendencias, una línea roja con el valor escalado del 0 para saber cuando los tuits para cada día (en media móvil) son positivos (por encima de la línea) o negativos (por debajo de la línea).

Para el estudio del índice de polaridad 2, se recuerda que este mide la diferencia absoluta entre tuit positivos y negativos para cada día.

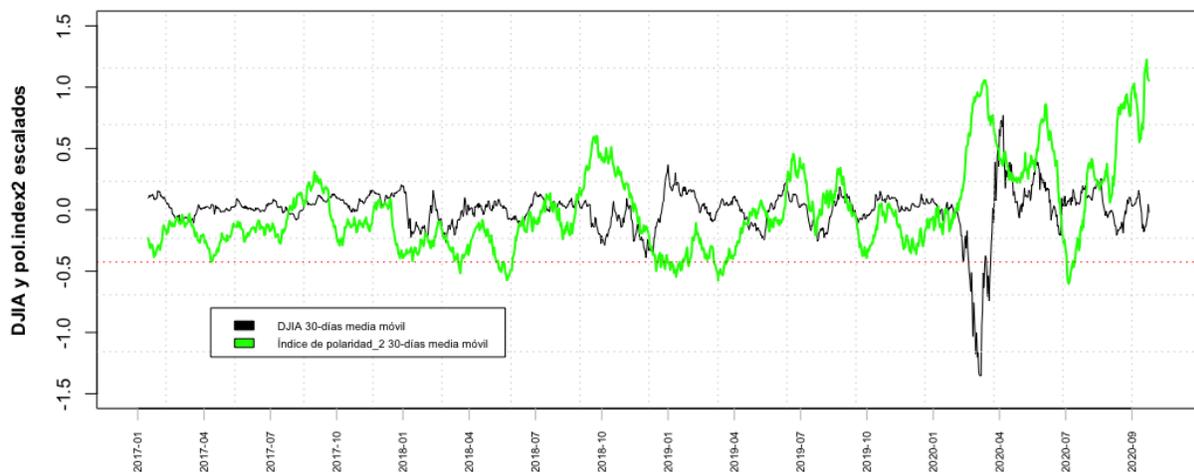


Figura 16: Serie temporal del índice de polaridad 2 y el índice DJIA

En la Figura 16 se puede ver como el inicio de la presidencia de Donald Trump se caracteriza por unos rendimientos poco volátiles y sostenidamente positivos, y un comportamiento de esta variable de opinión bastante estable en el tiempo. A partir de Octubre de 2018 se inicia un periodo con una relación inversa entre estas variables, cuando una sube la otra baja y viceversa que finaliza en mediados de 2019, donde parece que se inicie un breve periodo donde la relación pasa a ser positiva, cuando una sube la otra también sube y viceversa. Esto se da hasta el estallido de la crisis del COVID-19 donde los rendimientos caen sistemáticamente y el índice de polaridad 2 sube a valores muy elevados, la relación entre tuits positivos y negativos pasa a ser muy favorable de los positivos. Esta tendencia se rompe en verano de 2020 donde se registra la mayor diferencia entre tuits positivos y negativos en favor de los negativos del momento, a partir de este momento el índice de polaridad refleja valores muy extremos de tuit positivos y la bolsa parece independiente a dicho comportamiento. Al final de este gráfico, en referencia a la variable de opinión, se puede ver, lo que ya se había visto anteriormente y es el efecto del aumento drástico de tuits realizados por Donald Trump durante el final del estudio. Al ser este indicador una diferencia absoluta, el hecho de que el número de tuits aumente notablemente, hace que esta diferencia también tienda a aumentar en los dos extremos. Se ha añadido un línea roja con el valor escalado del 0 para saber cuando los tuits para cada día (en media móvil) son positivos (por encima de la línea) o negativos (por debajo de la línea).

A continuación, se realizará el mismo análisis para el ratio de tuits positivos y negativos respecto al total de tuits de cada día.

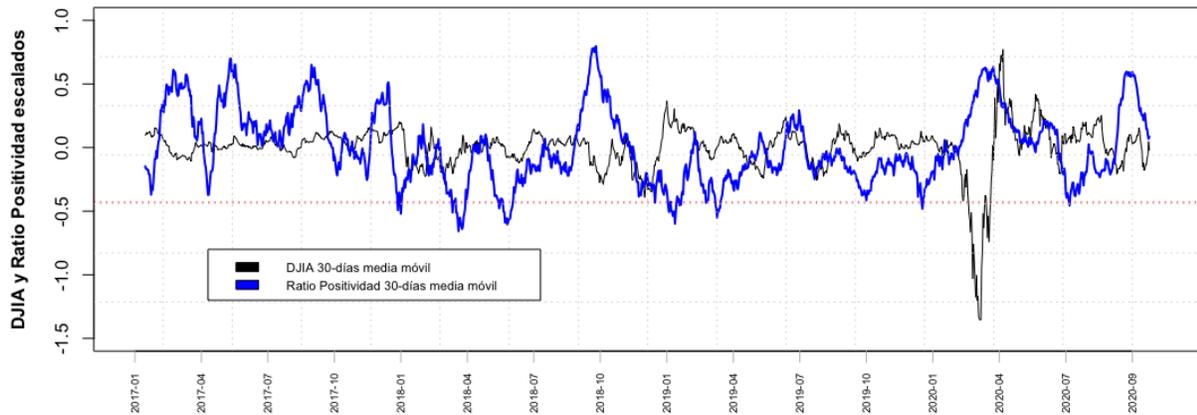


Figura 17: Serie temporal del ratio de positividad y el índice DJIA

Se puede observar en la Figura 17 el ratio de tuits positivos y los rendimientos del DJIA para la presidencia Trump, ambos valores escalados. Al inicio de esta gráfico vemos unos rendimientos con poca volatilidad y un ratio de tuits positivos bastante grande. A partir de 2018 se ve como este ratio disminuye visualmente hasta Octubre de 2018, para este periodo parece que los rendimientos también disminuyen ligeramente. En el pico de Octubre 2018 se puede ver claramente como la variable sentimental realiza un pico y la variable financiera un valle, ejemplificando este comportamiento inversamente proporcional que también vemos en la crisis del COVID-19 y en algún otro momento puntual. Poca información relevante se puede sacar del año 2019 a simple vista. Finalmente se puede ver como en Septiembre de 2020 hay un pico del ratio de positividad con un pico de la variable financiera, lo que demuestra que en algunos casos la relación entre las variables escaladas es positiva. Se ha añadido un línea roja con el valor escalado del 0.5 para saber cuando los tuits para cada día (en media móvil) son positivos (por encima de la línea) o negativos (por debajo de la línea).

En la Figura 18 se puede ver el ratio de tuits negativos y los rendimientos del DJIA para la presidencia Trump ambos valores escalados. Al inicio de esta gráfico vemos unos rendimientos con poca volatilidad y un ratio de tuits negativos inferior a su media. A partir de 2018 se ve como este ratio aumenta visualmente hasta Octubre de 2018, para este periodo parece que los rendimientos disminuyen ligeramente. En Octubre de 2018 se puede ver claramente



Figura 18: Serie temporal del ratio de negatividad y el índice DJIA

como la variable sentimental realiza un valle y la variable financiera también, ejemplificando este comportamiento proporcional que también vemos en la crisis del COVID-19 y en algún otro momento puntual. Poca información relevante se puede sacar del año 2019 a simple vista. Finalmente se puede ver como en Septiembre de 2020 hay un valle del ratio de negatividad con un pico de la variable financiera, lo que demuestra que en algunos casos la relación entre las variables escaladas es negativa. Este gráfico es muy parecido al anterior, y las conclusiones que se pueden sacar son parecidas, teniendo en cuenta que la relación entre los dos ratios está definida por construcción y están fuertemente correlacionados. Se ha añadido, para facilitar la comprensión, un línea roja con el valor escalado del 0.5 para saber cuando los tuits para cada día (en media móvil) son positivos (por debajo de la línea) o negativos (por encima de la línea).

Después de ver la relación gráfica entre las variables extraídas gracias al análisis sentimental de los tuits del presidente y los rendimientos bursátiles del Dow Jones, se considera que podría enriquecer el estudio hablar de la teoría de ciclos electorales y más en concreto, aplicándola al caso americano. El ciclo político es la política económica que siguen los gobiernos ante un proceso electoral. Cuando se acercan las elecciones, se suelen poner en marcha medidas de política económica que permitan conseguir el mayor número de votos para el partido gobernante, y alcanzar así su objetivo más importante, permanecer en el poder el mayor tiempo posible. En este sentido, la política económica está determinada por el ciclo político, es decir, por la proximidad o lejanía de unas elecciones.

Si nos centramos en el ciclo presidencial estadounidense, descrito inicialmente por Yale Hirsch en 2003 (Y. Hirsch y J. Hirsch, 2003), la tendencia es la siguiente: Los dos primeros años de presidencia acostumbran a tener un comportamiento del mercado bajista y recesiones, para dejar paso a una expansión del mercado y tiempos más prósperos para la última mitad del ciclo presidencial, tal como se ha explicado en el párrafo anterior. Históricamente, el Dow Jones ha subido un 10% el primer año de presidencia y un 8% el segundo. Observando los datos también se puede ver como el tercer año de presidencia es el más fuerte con una subida media de 13 puntos porcentuales, el ritmo baja notablemente hasta unos rendimientos del 5% para los años electorales. Para el caso de Donald Trump, se puede ver como su ciclo presidencial basado en los rendimientos del Dow Jones no sigue para nada la tendencia clásica descrita por los modelos de ciclos presidenciales. Los rendimientos del Dow Jones para el primer año de mandato del presidente Trump supusieron una subida del 32%, seguida por una caída del 5% el segundo año y un rebote del 19% durante los 365 días del tercer año. Para el último año de mandato y contabilizando hasta Octubre de 2020, el parqué americano del Dow Jones había perdido el 2.5% de su cotización, notablemente influido por la crisis económico-sanitaria del COVID-19. Se puede relacionar esta información con las conclusiones del análisis sentimental donde se decía que los primeros años de presidencia el presidente parecía estar “de buen humor”, siempre basándonos en su reflejo en Twitter. Seguidamente, teníamos una época marcada por un aumento del sentimiento negativo en los tuits de Donald Trump para tener un mejor “humor” a finales de 2019 y, finalmente, un “humor” muy volátil en 2020. Esto se puede ligar ligeramente con los rendimientos para cada año presentados anteriormente.

Referido, a la crisis sanitaria del COVID-19 se ha podido observar como esta causó una fuerte crisis financiera con una caída muy fuerte del DJIA, la más fuerte en todo el periodo de estudio. También, se ve como el mercado es capaz de recuperarse de esta fuerte caída y 5 meses más tarde ya volvía a sus valores anteriores a la crisis. Por lo que respecta, a como el presidente encara esta crisis, comportamiento reflejado en sus tuits, se puede ver como justo cuanto más afectaba esta crisis a la economía, más optimistas eran sus tuits, tal vez a posteriori, se puede decir que relativizo demasiado la magnitud de esta crisis. A medida, que esta crisis se convierte en persistente y los indicadores sanitarios empeoran, se ve como los tuits de Trump bajan su nivel de optimismo considerablemente, aunque estos vuelven a subir a partir de finales de este último verano. Esto, seguramente, sea más consecuencia de la proximidad de la elecciones y efecto de la teoría electoral que de la evolución de la pandemia.

En este apartado, se ha podido ver una buena introducción a la relación entre variables financieras y variables sentimentales. Algunas de las cosas más importantes que se han visto y que nos servirán para el posterior análisis de esta relación son: Primeramente, y tal vez, el hecho más destacado, es esta dicotomía entre la relación entre indicadores sentimentales y rendimiento económico que se alterna con el tiempo, se han encontrado periodos temporales donde el hecho de que los tuits del presidente sean más positivos se asocia con un buen rendimiento de los valores bursátiles, mientras que también se han encontrado otros periodos donde son los tuits de carácter más negativo, los que se asocian con buenos rendimientos bursátiles. Por último, también se ha comparado los rendimientos bursátiles de la presidencia Trump con la hipótesis de ciclos políticos y se ha visto que la presidencia Trump no parece seguir lo que dice la teoría de ciclos presidenciales. Tampoco siguen la teoría electoral su comportamiento en Twitter, ya que si nos basamos en la teoría su mejor “humor” se debería dar en sus dos últimos años electorales. Lo que si que es verdad, es que al final del periodo de seguimiento de sus tuits se ve una subida importante de los tuits positivos, reflejo de su humor, y esta subida coincide con la proximidad de los comicios electorales en los Estados Unidos.

A raíz de estas ideas, se ha decidido tratar el data-set de forma compacta para toda la presidencia Trump, pero también se realizará el análisis diferenciado por cada año, de esta forma también se podrá ver el papel del ciclo presidencial en los diferentes años y sobretodo, se entenderá mejor la relación dual entre indicadores sentimentales y variables financieras, ya que se ha visto que a veces la relación podía ser positiva y otras negativa.

## 6.2. Granger Causality

En este apartado nos centramos en saber si los indicadores de opinión o indicadores sentimentales expresados anteriormente en este trabajo se correlacionan con cambios en el mercado bursátil, en particular con los rendimientos del Dow Jones. También se mirarán correlaciones con el mercado de S&P 500 y su índice de volatilidad el VIX.

Uno de los métodos más populares para buscar correlación entre diferentes variables es la técnica econométrica del análisis de causalidad de Granger, el cual está basado en la regresión lineal. Este método se basa en que si la variable X causa la variable Y, entonces los cambios aparecerán antes en X que en Y. De esta manera, los *lagged values*<sup>22</sup> de X deberían estar estadísticamente correlacionados con la variable Y. Sin embargo, a pesar de su nombre, la causalidad de Granger solo muestra correlación, lo que no implica causalidad. Igualmente, sigue siendo una buena herramienta para indicar si una serie tiene información predictiva de otra serie.

En la Figura 19 se puede ver el funcionamiento de este test. Cuando la serie de tiempo X Granger-causa la serie de tiempo Y, los patrones en X se repiten aproximadamente en Y después de un lapso de tiempo (dos ejemplos se indican con flechas). Por lo tanto, los valores pasados de X pueden usarse para la predicción de valores futuros de Y.

Cuando se realiza este test, se obtiene un p-valor, si este valor es más bajo que el nivel de significación, esto significa que la serie temporal independiente *Granger-causa* a la serie temporal dependiente.

Se intenta mejorar nuestro modelo lineal incorporando los valores retrasados (*lagged values*) del índice bursátil. Matemáticamente y estadísticamente, la causalidad de Granger no existe si y solo si ningún valor retrasado de X es utilizado en la auto-regresión univariante de Y. Los valores de X se utilizan si son individualmente significativos según un test de t-Student y si colectivamente añaden poder explicatorio según un test de Fisher.

Nuestra serie temporal bursátil, denotado como  $D_t$ , refleja movimientos diarios en el mercado bursátil americano del Dow Jones, como se ha explicado anteriormente,  $D_t$  denota la diferencia relativa de los valores de  $t$  y  $t - 1$ . Para testear si nuestra serie sentimental basada en el índice de polaridad predice cambios en los valores de mercado, se compara la varianza explicada por dos modelos lineales mostrados en las ecuaciones 1 y 2. El primero modelo ( $L_1$ ) solo contiene  $n$  valores retrasados de la variable económica  $D_{t-i}$ , mientras que el segundo modelo ( $L_2$ ) contiene  $n$  valores retrasados de la variable económica  $D_{t-i}$  y añade  $n$  valores retrasados del índice de polaridad  $X_{t-i}$  que recoge el sentimiento de los tuits del presidente Donald Trump.

<sup>22</sup>Valores anteriores en el tiempo de la misma variable

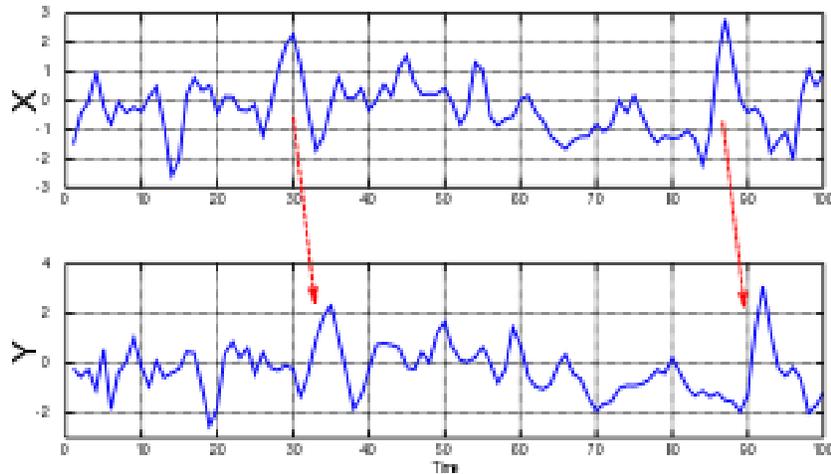


Figura 19: Ejemplo del funcionamiento del análisis de Granger  
Fuente: Granger causality 2021

Se realiza el análisis de causalidad de Granger según los modelos  $L_1$  y  $L_2$  detallados en las ecuaciones 1 y 2 para el periodo temporal de la presidencia Trump entre el 20 de Enero de 2017 y el 4 de Noviembre de 2020.

Este test se puede implementar en R mediante la función *grangertest* del paquete *lmtest*. Esta función necesita un vector  $x$  con la serie temporal independiente, un vector  $y$  con la serie temporal dependiente y un valor para *order* indicando que orden de valores retrasados queremos añadir al modelo ampliado.

$$L_1 : D_t = \alpha + \sum_{i=1}^n B_i D_{t-i} + \epsilon_t \quad (1)$$

$$L_2 : D_t = \alpha + \sum_{i=1}^n B_i D_{t-i} + \sum_{i=1}^n \gamma_i X_{t-i} + \epsilon_t \quad (2)$$

Que se traducirían en las siguientes hipótesis para cada retardo  $i$ :

$$H_0 : \gamma_i = 0 \text{ para cada valor de } i \text{ dentro de } [i, n]$$

$$H_1 : \gamma_i \neq 0 \text{ para algún valor de } i \text{ dentro de } [i, n]$$

Estas hipótesis del conjunto de análisis de Granger, permiten entender que si existe algún retardo significativo, se tendrán indicios estadísticos de que existe causalidad de Granger entre la serie independiente y la serie dependiente.

Listing 1: R output Ejemplo Test de Granger

```
Granger causality test

Model 1: df.DJIA$return ~ Lags(df.DJIA$return, 1:10) + Lags(df.DJIA$index.pol, 1:10)
Model 2: df.DJIA$return ~ Lags(df.DJIA$return, 1:10)
Res.Df  Df      F Pr(>F)
1    1143
2    1153 -10 0.317 0.977
```

En el Listado 1 se puede ver la salida en  $R$  de un ejemplo de este test de causalidad de Granger para la adición de los valores retardados hasta el orden 10 del índice de Polaridad 1. Se puede ver como el modelo 1 del test se correspondería con el modelo  $L_2$  que hemos definido nosotros que contiene las variables sentimentales y el modelo 1 se correspondería con el modelo solo con información financiera, que se ha llamado en este trabajo  $L_1$ . También se puede ver como el estadístico  $F$  de Fisher que mide la suma de cuadrados de los dos modelos, el restringido ( $L_1$ ) y completo ( $L_2$ ), conjuntamente con la probabilidad ( $Pr(> F)$ ) de que los datos se rijan bajo la hipótesis nula ( $H_0$ ). También se ven los grados de libertad para cada modelo según el número de variables que utiliza.

Este estadístico  $F$  se calcula de la siguiente manera y se distribuye según una distribución  $F$  de Snedecor con  $p_2 - p_1$  y  $T - p_2$  grados de libertad, donde  $p_1$  es el número de parámetros del modelo restringido  $L_2$ ,  $p_2$  es el número de parámetros del modelo sin restringir  $L_1$  y  $T$  el número de observaciones en los modelos.

$$F = \frac{\frac{RSS_{L_2} - RSS_{L_1}}{p_2 - p_1}}{\frac{RSS_{L_1}}{T - p_2}}$$

Donde  $RSS_{L_1}$  es la suma de cuadrados de los residuos del modelo sin restringir y  $RSS_{L_2}$  la suma de cuadrados de los residuos del modelo restringido, es decir con las variables sentimentales.

En el siguiente Cuadro 9 se pueden ver los p-valores de los diferentes Test de Granger para los 4 indicadores sentimentales observados: Indicador de Polaridad 1 (*Ind.Pol.1*), Indicador de Polaridad 2 (*Ind.Pol.2*), Ratio de Negatividad (*Ratio Neg.*) y Ratio de Positividad (*Ratio Pos.*). Para cada uno de estos indicadores se muestra el Test de Granger con variables retrasadas de hasta 1 día a hasta 10 días. No se piensa que los efectos de los indicadores se puedan ver vistos en las variables económicas con más de 10 días de retraso, se han hecho algunas pruebas para comprobarlo si así fuera pero se ha visto que los p-valores a partir del *lag 10* aumentaban.

DJIA : Lag	Ind. Pol.1	Ind.Pol.2	Ratio Neg.	Ratio Pos.
1 día	0.72	0.49	0.73	0.73
2 días	0.90	0.83	0.88	0.93
3 días	0.95	0.82	0.94	0.96
4 días	0.99	0.92	0.99	0.99
5 días	0.99	0.82	0.99	0.99
6 días	0.90	0.44	0.90	0.91
7 días	0.93	0.51	0.92	0.93
8 días	0.86	0.28	0.85	0.86
9 días	0.97	0.62	0.97	0.97
10 días	0.98	0.67	0.98	0.98

Cuadro 9: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el DJIA

En este Cuadro 9 se puede observar la significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el DJIA para diferentes *lags*. Se puede concluir no tenemos indicios para rechazar la hipótesis nula que dice que las series sentimentales no predicen los valores del DJIA dado un nivel de confianza del 90%. Por lo tanto, se puede afirmar que estos 4 indicadores no tienen un efecto significativo causal sobre las variaciones del Dow Jones.

En el Anexo C se puede encontrar más información de este mismo test pero realizado con los rendimientos del mercado S&P 500 y del índice artificial de volatilidad VIX, en vez de sobre el mercado Dow Jones. Las conclusiones en estos casos y para este mismo periodo temporal son las mismas y es que no se ha encontrado ningún indicio de que estos indicadores sentimentales puedan tener relación con los valores del mercado S&P 500 o del índice VIX con un nivel de confianza del 90%. Para el caso del VIX se estudiará más al detalle en uno de los subapartados posteriores.

A continuación se realizará un test parecido pero en vez de centrarnos en buscar la relación con los rendimientos de los mercados, pondremos el foco en establecer si hay causalidad entre las variables sentimentales y el comportamiento de los mercados, entendiendo este comportamiento en si el mercado ha subido o ha bajado respecto al día anterior.

Para ello, crearemos una nueva variable llamada  $Binary_t$  que aporta información sobre si ese día el mercado ha subido o bajado respecto al día anterior para los dos mercados bursátiles del DJIA y del S&P 500.

$$Binary_t = \begin{cases} 1 & \text{si } ret_t(R_t) > 0 \\ 0 & \text{si } ret_t(R_t) < 0 \end{cases} \quad (1)$$

A partir de esta variable y sus valores retrasados podemos escribir las ecuaciones para poder realizar el Test de Granger y ver si las variables sentimentales *Granger causan* las subidas y bajadas diarias del mercado.

$$L_3 : Binary_t = \alpha + \sum_{i=1}^n S_i Binary_{t-i} + \epsilon_t \quad (3)$$

$$L_4 : Binary_t = \alpha + \sum_{i=1}^n S_i Binary_{t-i} + \sum_{i=1}^n \gamma_i X_{t-i} + \epsilon_t \quad (4)$$

Donde ahora se está intentando predecir si el mercado subirá o bajará para el día  $t$  en función de si han subido o bajado  $Binary_t$  los días anteriores ( $L_3$ ) o en función de si ha subido o bajado días anterior y también en función de los valores de las variables sentimentales  $X_{t-i}$  durante los días anteriores ( $L_4$ ).

DJIA .binary : Lag	Ind. Pol.1	Ind.Pol.2	Ratio Neg.	Ratio Pos.
1 día	0.65	0.86	0.73	0.58
2 días	0.35	0.94	0.35	0.36
3 días	0.10	0.91	<b>0.09*</b>	0.11
4 días	0.22	0.88	0.21	0.24
5 días	0.38	0.76	0.37	0.40
6 días	0.28	0.67	0.27	0.29
7 días	0.33	0.76	0.32	0.35
8 días	0.41	0.84	0.39	0.43
9 días	0.59	0.88	0.57	0.61
10 días	0.56	0.91	0.56	0.57

Cuadro 10: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el DJIA .binary

En este Cuadro 10 se puede observar la significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el DJIA.binary para diferentes *lags*. En esta tabla resalta que para una confianza del 90% el p-valor de 0.09 (p-valor inferior a 0.10) procedente del indicador llamado Ratio de Negatividad es significativo en su tercer valor retrasado. Este resultado, nos da indicios para poder rechazar la hipótesis nula, de que el ratio de negatividad retrasado un máximo de 3 días, no puede predecir si el mercado del Dow Jones subirá o bajará. Por lo tanto, se puede decir que los valores de la variable sentimental tienen poder predictivo significativo sobre si el mercado subirá o bajará durante 3 días. Para el resto de variables y de retraso se puede concluir no tenemos indicios para rechazar la hipótesis nula que dice que las series sentimentales no predicen los valores del DJIA dado un nivel de confianza del 90%.

En el siguiente subapartado, se indagará más en profundidad sobre esta relación significativa encontrada, al ser la variable respuesta una variable dicotómica que solo puede valer 1 o 0, se ha decidido utilizar un modelo logit para poder estudiar esta relación más al detalle.

Para el resto de variables económicas utilizadas en este trabajo (S&P 500 y VIX) tampoco se ha encontrado ninguna relación entre las variables binarias indicadoras de rendimientos positivos o negativos (construidas de la misma forma que para la variable *binary.DJIA*) y los valores retardados de las variables sentimentales utilizadas para un nivel de confianza del 90%. Estas conclusiones también se pueden encontrar en el Anexo C.

Como se ha comentado anteriormente, el ciclo político-económico tiene un papel importante sobre el mercado bursátil americano. Además, tal y como se ha visto anteriormente parece que la relación entre variables financieras y sentimentales es ambigua, a veces un rol más positivo desde el Twitter de Trump conlleva a rendimientos bursátiles positivos y otras veces se ha visto que se asocia con rendimientos bursátiles negativos. Por lo tanto, se ha decidido dividir la presidencia Trump en cuatro periodos, referentes a los cuatro años que está en mandato, y volver a realizar los Test de causalidad de Granger explicados anteriormente para cada año en concreto. Así se podrá ver si en periodos de tiempo más cortos se pueden estudiar las relaciones de manera más clara y sobretodo de forma más aislada.

Por lo tanto, se ha dividido nuestro datos en cuatro periodos: 2017, 2018, 2019 y 2020 según los días de cada año y se han repetido los test anteriores para intentar predecir por un lado los valores bursátiles del Dow Jones  $ret.DJIA_t$  y por otro el hecho de si estos rendimientos serán positivos o negativos  $binary.DJIA_t$  en función de los indicadores sentimentales (de misma manera con el mercado S&P 500).

Los resultados significativos de estos test se pueden encontrar en el Anexo C: Test por años: 2017, 2018, 2019 y 2020. Las conclusiones de dichos test para un nivel de confianza del 90 % son que:

- El modelo con retardos de hasta 9 días del Ratio de Negatividad tienen información predictiva sobre los valores de los rendimientos del Dow Jones para el año 2017 en concreto. Esta relación es positiva y estadísticamente significativa solo para el retardo del día 8, por lo tanto, para el año 2017 cuando aumenta el ratio de negatividad, es decir, el número relativo de tuits clasificados como negativos, también aumentan los rendimientos bursátiles al cabo de 8 días. Más concretamente, aumentar el ratio de negatividad en 1 punto, aumenta los rendimientos esperados en 0.0174, por lo tanto en un 1.74 %. (test significativo con un nivel de confianza del 95 %, ver Anexo C.)
- El modelo con retardos de hasta 10 días del Ratio de Positividad, del Ratio de Negatividad y del índice de Polaridad 1 tienen información predictiva sobre la variable binaria del S&P 500 para el año 2018. Se ha visto como aumentar los índices de Polaridad 1 o el Ratio de Positividad hace más probable que al cabo de 6 y 10 días el rendimiento bursátil del S&P 500 sea negativo. Al contrario pasa con el 5 día, donde vemos que la relación es positiva para las variables del Ratio de positividad y del índice de polaridad 1. Aumentar la proporción de tuits positivos hace aumentar la probabilidad de tener un rendimiento positivo al cabo de 5 días. Al revés pasa para el Ratio de Negatividad, por construcción, aumentar el número relativo de tuits negativos hace disminuir la probabilidad de obtener rendimientos positivos al cabo de 5 días y hace aumentar la posibilidad de obtener rendimientos positivos al cabo de 6 y 10 días. Por lo tanto, se puede ver como incluso en periodos temporales cortos existe esta dualidad de relación entre las variables sentimentales y financieras. (test significativo con un nivel de confianza del 90 % para los días 5-6 y significativo con nivel de confianza del 95 % para el retardo del día 10, ver Anexo C.)
- No se han encontrado indicios para establecer ninguna otra relación entre las variables financieras en forma de rendimiento o en forma binaria y los retrasos de las variables sentimentales.

En este apartado donde se ha empezado a ver la relación entre variables financieras y indicadores de opinión de Donald Trump, procedentes del análisis sentimental de sus tuits se ha encontrado relación significativa de su ratio de tuits negativos para toda su presidencia para el mercado del Dow Jones, en su variable respuesta binaria. Es decir, que su número de tuits negativos respecto al total de tuits con sentimiento, tiene información predictiva a 3 días de los rendimientos de dicho mercado. También, se ha visto, como al separar la muestra por años, existe relación positiva en el año 2017 entre el ratio de negatividad y los valores numéricos de los rendimientos para 8 días vista, para el mercado del Dow Jones, tendencia parecida en efecto a la vista en toda la presidencia. Adicionalmente, para el mercado del S&P 500, se ha encontrado relación entre las variables de opinión (excepto el índice de polaridad 2) y los rendimientos en su formato binario, siendo esta relación ambigua entre positiva o negativa. Finalmente, se puede decir que no hay indicios de ninguna otra relación, a parte de las ya mencionadas, para el periodo completo, ni para los años 2019 y 2020. Tampoco se ha encontrado ninguna relación entre los tuits del presidente y el índice de volatilidad del mercado VIX. En el siguiente apartado se estudiará más a fondo la relación para el periodo completo de la presidencia Trump con el modelo que mejor se ajuste a los datos.

Para finalizar este apartado se quiere subrayar el hecho de que se han realizado muchos test de causalidad de Granger en esta sección, por lo que sería correcto realizar una corrección de los p-valores mostrados en los cuadros de este apartado en función del número de test e hipótesis totales hechas. La prueba de hipótesis estadística se basa en rechazar la hipótesis nula si la probabilidad de los datos observados en las hipótesis nulas es baja. Si se prueban múltiples hipótesis, aumenta la probabilidad de un evento raro y, por lo tanto, aumenta la probabilidad de rechazar incorrectamente una hipótesis nula error de tipo 1<sup>23</sup>. El objetivo de esta corrección sería compensar este aumento del error de tipo I al realizar muchos test.

### 6.3. Modelos con variable respuesta binaria: Logit

La estimación e interpretación de los modelos probabilísticos lineales plantea una serie de problemas que han llevado a la búsqueda de otros modelos alternativos que permitan estimaciones más fiables de las variables dicotómicas. Para

---

<sup>23</sup>El error de tipo I, también denominado error de tipo alfa o falso positivo, es el error que se comete cuando el investigador rechaza la hipótesis nula  $H_0$  siendo esta verdadera en la población. Es equivalente a encontrar un resultado falso positivo, porque el investigador llega a la conclusión de que existe una diferencia entre las hipótesis cuando en realidad no existe.

evitar que la variable endógena estimada pueda encontrarse fuera del rango (0, 1), las alternativas disponibles son utilizar modelos de probabilidad no lineales, donde la función de especificación utilizada garantice un resultado en la estimación comprendido en el rango 0-1. Una de las funciones de distribución que cumple este requisito es la regresión logística, ya que utiliza una función continua que toma valores comprendidos entre 0 y 1.

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelizar la probabilidad de un evento ocurriendo como función de otros factores. El análisis de regresión logística se enmarca en el conjunto de Modelos Lineales Generalizados que usa como función de enlace la función logit.

El modelo logit se especifica la probabilidad de un suceso de nuestra variable respuesta como:

$$P(Y = 1) = \frac{\exp(X\beta + \epsilon)}{1 + \exp(X\beta + \epsilon)} = p$$

Donde la variable respuesta (variable aleatoria  $Y$ ) es dicotómica o binaria (toma dos valores: 0 y 1), en nuestro caso si el rendimiento bursátil ha sido positivo o negativo para un día concreto, como consecuencia,  $P(Y = 0) = 1 - p$ . Las variables predictivas (vector aleatorio  $X$ ) son continuas,  $\beta$  es un vector de parámetros a estimar y  $\epsilon$  un error aleatorio. El modelo logit es un caso particular de los llamados modelos lineales generalizados. Para los detalles del modelo véase *Ranganathan, Nakai y Schonbach, 2018*.

El objetivo de este apartado será estudiar la relación entre las variables significativas del apartado anterior con la variable respuesta sobre las que han resultado significativas. En este caso se quiere ver cual es el efecto del los tres retardos del ratio de negatividad sobre la variable respuesta que mide si un día la bolsa ha subido o bajado. Recordando la visión global del trabajo se quiere ver si los tuits de Donald Trump, de los cuales se han sintetizado 4 indicadores después de realizar un análisis de sentimientos tienen una relación significativa con los movimientos del mercado a posteriori.

En software *R*, los modelos lineales generalizados se ajustan con la función *glm*. La principal diferencia con la función *lm* para ajustar modelos lineales es que le tenemos que proporcionar la familia de la distribución. En nuestro caso, como es una variable dicotómica, la familia es la binomial y la función de de enlace que escogemos es *logit*.

Listing 2: R output Logit Modelo con el periodo completo

```

Call:
glm(formula = df.DJIA$binary ~ lagpad(df.DJIA$binary, 1) + lagpad(df.DJIA$binary,
  2) + lagpad(df.DJIA$binary, 3) + lagpad(df.DJIA$ratio_neg,
  1) + lagpad(df.DJIA$ratio_neg, 2) + lagpad(df.DJIA$ratio_neg,
  3), family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8429  -1.0362   0.7273   0.9382   1.7351

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.7832    0.1912  -4.096  4.2e-05 ***
lagpad(df.DJIA$binary, 1)  1.1733    0.1230   9.538 < 2e-16 ***
lagpad(df.DJIA$binary, 2)  0.5471    0.1276   4.288  1.8e-05 ***
lagpad(df.DJIA$binary, 3) -0.4078    0.1269  -3.214  0.00131 **
lagpad(df.DJIA$ratio_neg, 1) -0.1450    0.2576  -0.563  0.57350
lagpad(df.DJIA$ratio_neg, 2)  0.3333    0.2597   1.284  0.19929
lagpad(df.DJIA$ratio_neg, 3)  0.5529    0.2564   2.157  0.03101 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1813.7  on 1319  degrees of freedom
Residual deviance: 1664.8  on 1313  degrees of freedom
(69 observations deleted due to missingness)
AIC: 1678.8

Number of Fisher Scoring iterations: 4

```

La interpretación de los p-valores del *listing 2* es similar a la del modelo lineal. Podemos ver que dos de las variables referidas a los indicadores sentimentales no son significativas en el modelo (p-valor mucho mayor de 0.05), mientras que el retardo del ratio de negatividad del tercer día y las tres variables financieras son significativas (p-valor menor a 0.05). Según la literatura financiera es normal que los precios tengan correlación estadística con los precios de los días anteriores, la intensidad de la relación se pierde a medida que nos alejamos del día  $t$ , por eso los p-valores resultantes de los test de significación aumentan a medida que miramos retardos más lejanos. Destaca la significación del retraso sentimental del tercer día en el modelo.

En cuanto a la interpretación de los parámetros estimados en un modelo Logit, el signo de los mismos indica la dirección en que se mueve la probabilidad cuando aumenta la variable explicativa correspondiente, sin embargo, la cuantía del parámetro no coincide con la magnitud de la variación en la probabilidad. En el caso de los modelos Logit, al suponer una relación no lineal entre las variables explicativas y la probabilidad de ocurrencia del acontecimiento, cuando aumenta en una unidad la variable explicativa los incrementos en la probabilidad no son siempre iguales ya que dependen del nivel original de la misma. Esto se debe a que el modelo Logit no ajusta la variable respuesta sino una función de enlace. En el caso del modelo logit esta función es:

$$\eta = \log\left(\frac{p}{1-p}\right)$$

Como habíamos definido anteriormente,  $p = \frac{\exp(X\beta+\epsilon)}{1+\exp(X\beta+\epsilon)}$ , es la probabilidad de que el día bursátil sea positivo en la variable dicotómica. Al cociente  $\frac{p}{1-p}$  se le conoce como *odds ratio*. Su interpretación es la “ventaja” o preferencia de la opción 1 frente a la 0, es decir, el número de veces que es más probable que ocurra el fenómeno frente a que no ocurra. El *odds ratio*, tal y como está construido (cociente entre probabilidades), siempre será mayor o igual que 0. El campo de variación del ratio va desde 0 hasta  $+\infty$ , y su interpretación se realiza en función de que el valor sea igual, menor o superior a la unidad: si toma el valor 1 significa que la probabilidad de que ocurra la alternativa

1 es la misma que la de que no ocurra; si el ratio es menor que 1 indica que la ocurrencia de la alternativa 1 tiene menor probabilidad que la ocurrencia de la alternativa 0; mientras que si es mayor que la unidad la opción 1 es más probable que la 0.

Por lo tanto, los coeficientes del modelo logit se interpretan como el logaritmo del *odds ratio*. Si nos fijamos en el coeficiente de la variable **lagpad(ratio\_neg,3)** (0.5529), nos está indicando que el logaritmo del *odds ratio* de que el día bursátil sea positivo aumenta 0.5529 unidades por cada unidad que aumenta la puntuación del ratio de negatividad.

Una forma de facilitar la interpretación de los coeficientes es exponenciando:

Variable:	Intercept	<i>binary.DJIA<sub>t-1</sub></i>	<i>binary.DJIA<sub>t-2</sub></i>	<i>binary.DJIA<sub>t-3</sub></i>	<i>ratio_neg<sub>t-1</sub></i>	<i>ratio_neg<sub>t-2</sub></i>	<i>ratio_neg<sub>t-3</sub></i>
Coefficientes Exponenciados	0.4570	3.2327	1.7282	0.6651	0.8650	1.3355	1.7383

Cuadro 11: Coeficiente Exponenciados del modelo logarítmico

Que se corresponde con el modelo:

$$odds = e^{\beta} * e^{\beta_1 * x_1} * e^{\beta_2 * x_2} * e^{\beta_3 * x_3} * e^{\beta_4 * x_4} * e^{\beta_5 * x_5} * e^{\beta_6 * x_6}$$

La conclusión de esta transformación es que ahora si que podemos interpretar los coeficientes. Centrándonos como objetivo de nuestro trabajo en la repercusión de las variables sentimentales significativas sobre los rendimientos, aumentar el ratio de negatividad en 1 punto tres días antes del día *t* hace aumentar en un 73% las posibilidades de que el rendimiento del día *t* sea positivo (*ratio\_neg<sub>t-3</sub>*). Mirando al resto de coeficientes significativos, se puede ver como los rendimientos del día anterior influyen en el rendimiento del día *t*, siendo este más probable de ser positivo si los dos días anteriores los rendimientos son positivos (coeficiente de *binary.DJIA<sub>t-1</sub>* y *binary.DJIA<sub>t-2</sub>* > 1) y si el rendimiento del tercer día anterior es negativo (coeficiente *binary.DJIA<sub>t-3</sub>* < 1).

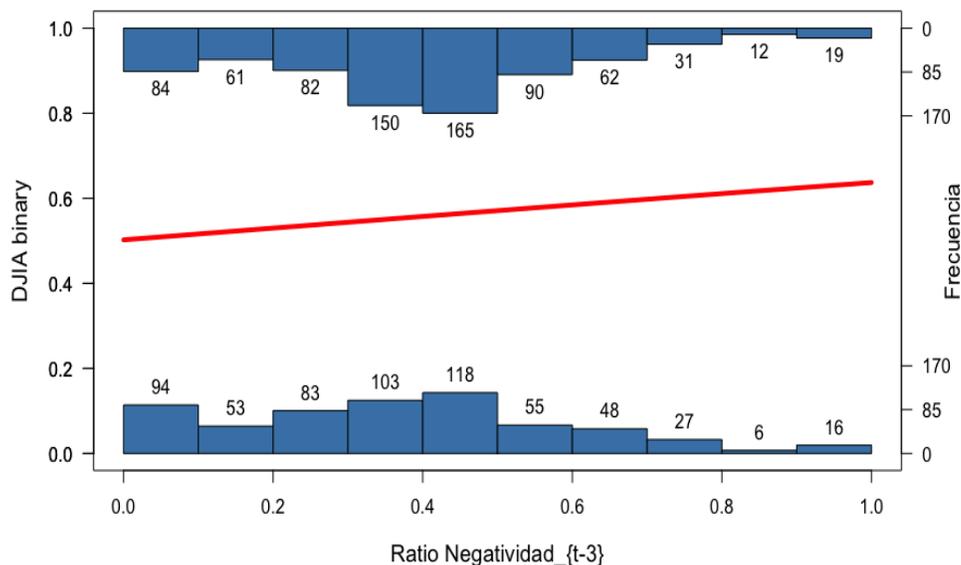


Figura 20: Efecto del tercer retardo de Negatividad sobre DJIA.Binary en modelo logit

En la Figura 20 se puede ver la relación entre el Ratio de Negatividad retrasado 3 días y la variable binaria que mide si el Dow Jones ha subido o bajado para un día *t*. También se puede ver la frecuencia de casos donde el mercado bursátil sube o baja para diferentes intervalos de la variable sentimental. Esta figura señala una ligera tendencia por

la que aumentan las frecuencia de días donde la bolsa sube en días donde el valor del ratio de negatividad retrasado es superior. Este gráfico visualiza los resultados mostrados anteriormente y nos permite entender la relación entre esta variable sentimental y los rendimientos financieros.

Este resultado, será vital para la evaluación de si los tuits de Trump tienen efectos persistentes en el mercado, con este resultado podemos decir que una variable sentimental, en este caso, una variable que mide el número de tuits negativos respecto al total de tuits con opinión, tiene influencia sobre el mercado bursátil del Dow Jones en un periodo de tiempo de 4 años. Por lo que en este caso, se ve que hay una relación positiva entre el número de tuits catalogados como negativos de Trump en el día  $t - 3$  y la categorización de los rendimientos entre positivos y negativos.

## 6.4. Relación de las variables de opinión con la volatilidad del mercado

De momento, se ha observado la relación entre las variables sentimentales y los diferentes mercados bursátiles como el Dow Jones y el S&P 500, y también con el índice de volatilidad del mercado S&P 500 cómo es el VIX. Estas variables sentimentales se han creado a partir del sentimiento promedio de cada día para los tuits de Donald Trump. A lo largo del estudio, se ha considerado que para el índice VIX sería interesante relacionarlo no con el sentimiento promedio para cada día sino con la desviación estándar de este sentimiento promedio. De esta forma, llevamos el análisis a un nivel superior, ya que por un lado analizamos si el hecho de que los tuits de Donald Trump sean catalogados como positivos o negativos influye en la volatilidad del mercado y adicionalmente en este apartado si en los días que los tuits del presidente son más “volátiles”, es decir, tiene tuits muy negativos y muy positivos, influye en la propia volatilidad del mercado, medida por el índice VIX.

Por lo tanto, el objetivo de este subapartado es estudiar si el hecho de que el presidente Trump tenga tuits muy dispares, muy diversos para un día, afecta a la volatilidad de ese día o de los días siguientes. Es una forma de comparar “volatilidades” de las dos formas. Para realizar esto se asignará como volatilidad procedente del análisis sentimental para un día  $t$ , la desviación típica entre las puntuaciones de sentimiento de los diferentes tuits realizados por el presidente en ese día  $t$ .

En el siguiente Cuadro 12 se rescata una variable preliminar del análisis de sentimiento de los tuits de Donald Trump. Esta variable **sentimiento promedio** mide para cada tuit su puntuación sentimental, basada en su contenido y en el diccionario léxico. Se recuerda que esta puntuación sera positiva si el tuit contiene palabras definidas con connotación positiva en el diccionario y negativa si el tuit contiene palabras definidas con connotación negativa en el diccionario. Para este análisis se recogen dos valores para esta variable **sentimiento promedio**, por un lado la desviación típica de tuits escritos el mismo día, y la media de este sentimiento promedio para los tuits escritos el mismo día. De esta manera se obtiene una serie temporal con estos dos valores para cada día. Una vez obtenidas estas dos series temporales: **sd(sentimiento\_promedio)** y **mean(sentimiento\_promedio)** se puede realizar un estudio de Granger para ver si estas variables tienen información predicativa sobre los valores del índice de volatilidad del mercado.

VIX: Lag	sd(sentimiento_promedio)	mean(sentimiento_promedio)
1 día	0.17	0.97
2 días	0.41	0.83
3 días	0.58	0.74
4 días	0.79	0.92
5 días	0.74	0.95
6 días	0.76	0.15
7 días	0.82	0.16
8 días	0.84	0.17
9 días	0.84	0.24
10 días	0.88	0.25

Cuadro 12: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para la desviación típica y la media del sentimiento promedio de cada tuit y el VIX

En este Cuadro 12 se puede observar los p-valores de la correlación bivalente de causalidad de Granger para la desviación típica y la media del sentimiento promedio de cada tuit y el VIX. No se encuentra ningún valor significativo, por lo que parece que ni los valores de la media ni desviación estándar por cada día y por sus diferentes retrasos contienen capacidad predictiva sobre los valores del índice de volatilidad VIX.

En este apartado se considera, a diferencia del apartado anterior, que no es necesaria una variable binaria que mida si el índice VIX ha subido o bajado respecto al día anterior, ya que a diferencia de los rendimientos bursátiles donde es importante y trascendente si un día es positivo o negativo aunque la diferencia en términos numéricos sea baja, en el índice VIX no es importante ni trascendente el caso de que el VIX de un día suba un 1 % o baje un 1 % ya que el nivel de volatilidad sigue siendo prácticamente el mismo.

La conclusión de este subapartado es que no hay ninguna relación estadística significativa entre los tuits del presidente y la volatilidad del mercado durante los 4 años de presidencia. Esto no implica que no haya ninguna relación, sino que para el computo global de estos 4 años y para la variable concreta del VIX no se ha encontrado ninguna relación suficientemente fuerte como para hablar de relación estadística entre tuits y volatilidad del mercado bursátil.

Se recuerda, el análisis doble realizado sobre la volatilidad, por una lado se ha mirado si la puntuación de los tuits en forma de positivos o negativos tiene efecto sobre la volatilidad, es decir, si por ejemplo en días donde los tuits son más negativos la incertidumbre del mercado sube. Otra faceta del estudio, ha sido mirar si la desviación en la puntuación de estos tuits influye en la volatilidad, dicho de otra manera, si los días donde hay más diferencia entre la puntuación de los tuits del presidente son los días con más volatilidad en el mercado. De esta manera, se considera que se ha llegado a un nivel de análisis mucho más rico y profundo.

## 7. Estudio de eventos aplicado a Twitter

Un estudio de eventos es un método estadístico para evaluar el impacto de un evento en el valor de una empresa, en este caso en el valor de un índice financiero. La idea básica es encontrar el rendimiento anormal atribuible al evento que se está estudiando ajustando el rendimiento que deriva de la fluctuación de precios del mercado en su conjunto.

El objetivo de este apartado es realizar un estudio más específico, poniendo el foco no el sentimiento global de los tuits de Donald Trump, sino en sus escritos de carácter puramente económico o financiero, ya que uno de los objetivos de este trabajo son los efectos que se producen en los mercados, por lo tanto, es lógico centrarse en los tuits que más relación podrían tener con el mercado. Este apartado, también permitirá tener una visión más completa de los tuits de Donald Trump, ya que de momento se ha trabajado con una muestra de todos sus tuits, ahora sólo se trabajará con los que cumplan una serie de condiciones y por lo tanto será más fácil visualizar la muestra. Para la realización, de esta parte del estudio se seguirá la metodología de estudio de eventos financieros desarrollada por *Brown y Warner, 1980* en su artículo sobre como medir la seguridad de las acciones. En este apartado, se seguirá el procedimiento descrito por los autores anteriormente nombrados para poder identificar cuales han sido los tuits de Donald Trump que han tenido mayor efecto en el mercado, para medir dicho efecto se tendrá en cuenta la media de los rendimientos del mercado durante la ventana del evento, para antes y para después del evento y también se tendrá en cuenta que la media de los rendimientos para la ventana del evento sea significativamente diferente a la media de la ventana de estimación. Una vez encontremos los tuits que cumplan estas condiciones se intentará dar una explicación de carácter general y económico a porque estos tuits en específico han tenido estas repercusiones en el mercado.

El primer paso para realizar un estudio de eventos es definir la ventana del evento. La ventana del evento es el intervalo de tiempo donde ocurre el evento. Esta ventana puede incluir un cierto número de días antes y después del anuncio.

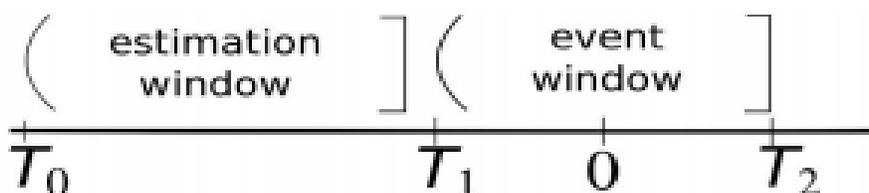


Figura 21: Timeline del Event Study

Fuente: *Rayarel, 2018*

Esta cronología que se puede ver en la Figura 21 define los periodos clave de un estudio de eventos. En la cronología,  $t = 0$  se corresponde con el día en el que el tuit ocurre. La ventana de evento va de  $T_1$  a  $T_2$  donde  $T_1$  es el primer día en la ventana del evento y  $T_2$  el último día. No hay un consenso claro en cual ha de ser la longitud correcta de la ventana del evento y dependerá siempre del tipo de evento. En la literatura existente, la ventana del evento varía de 1 a 40 días. *Born, Myers y Clark, 2017*, quienes también trabajaron con el efecto de los tuits de Donald Trump en el mercado financiero, incorporaron un rango de ventana de 10 días, desde -5 (5 días antes del evento) a 5 (5 días después del anuncio en Twitter). En este estudio se utilizará una ventana diferente que ira de 10 días antes del evento a 10 días después del evento, ya que en apartados anteriores se ha visto que los tuits de Donald Trump pueden tener efectos sobre el mercado financiero hasta prácticamente 10 días después de su tuit. Además esta ventana se considera adecuada para detectar que no hay eventos anteriores que afecten a los rendimientos y también permite estimar el efecto temporal de los tuits.

El segundo paso de un estudio de eventos es definir la ventana de estimación. La ventana de estimación es el intervalo de tiempo anterior a la ventana del evento que se usa para estimar los rendimientos esperados (o rendimientos normales). Para seleccionar la ventana de estimación, es un método común, según *Peterson, 1989*, escoger una ventana de estimación de una longitud de 100 a 300 días. Cuantos más días se usan, más precisa sera la estimación de los rendimientos normales del mercado. En este estudio la ventana de estimación es de 250 días. Se ha añadido también un espacio de 10 días entre la ventana de estimación y la ventana del evento para asegurarnos de prevenir que la información del tuit no afecte a los rendimientos esperados normales, es decir que no hayan posible repercusiones antes del tuit, procedentes de la información del tuit, que afecten a la ventana de estimación.

El siguiente paso para realizar un estudio de eventos es definir la muestra de tuits. Para esto se buscará identificar todos los tuits de Donald Trump de carácter económico o financiero. Para esto, se decide incluir en la muestra todos los tuits de Donald Trump que contengan alguno de los siguientes términos:

- Lista de términos económico-financieros: “DJIA”, “dow”, “Dow Jones”, “bearish”, “bear market”<sup>24</sup>, “best stock”, “bullish”, “bull market”<sup>25</sup>, “finance”, “finance news”, “financial news”, “fiencial market”, “long stock”, “SP500”, “stock”, “stock market”, “stock decline”, “stock fall”, “market crash”, “stock news”, “market today”, “stock price”, “to buy”, “wall street”, “news today”, “unemployment”, “economy”, “trade”, “tariff”, “jobs”, “SP”, “deal”, “money”, “industry”, “companies”, “Fed”<sup>26</sup>, “dollar”, “inflation”, “tax”, “taxes”, “oil”, “economic”, “economy”, “business”, “subsidies”, “inflation”, “tariffs”.

Estos términos han sido escogidos basándose en estudios parecidos realizados sobre magnitudes financieras para diferentes empresas. En este caso se han adaptado al vocabulario de Donald Trump con la revisión de las nubes de palabras procedentes del análisis textual.

En global, la muestra entera de tuits económico-financieros es de 117 tuits que contienen alguno de los términos anteriormente listados.

Ahora que el periodo de estimación se ha definido y la muestra ha sido seleccionada, los rendimientos normales (o esperados) se calculan para el índice DJIA. Hay dos modelos usados comúnmente para calcular los rendimientos normales. El modelo de rendimientos de media constante (*Constant Mean Return Model o CRM*) es un modelo estadístico que postula que los rendimientos esperados son simplemente la media de los rendimientos en la ventana de estimación. Otro enfoque diferente y centrado en el rendimiento de compañías y no de mercados, es el modelo de mercado que calcula el rendimiento esperado para una compañía teniendo en cuenta los rendimientos pasados del mercado y el grado de variabilidad de la rentabilidad de una acción de la compañía respecto a la rentabilidad promedio del mercado en que se negocia (comúnmente a este grado se le conoce como coeficiente beta  $\beta$ ). En nuestro caso, como se quiere investigar la relación de los tuits del presidente sobre el mercado y no su efecto sobre compañías o sectores concretos se decide optar por el primer método, el *Constant Mean Return Model*.

El modelo CRM requiere la siguiente estimación:

Para obtener los rendimientos anormales para el estudio de evento del tuit, se restará al rendimiento de cada día  $R_{i,t}$  de la ventana de evento (10 días antes y días después) el valor esperado del rendimiento del mercado  $E(R_{i,t})$ .

$$AR_{i,t} = R_{i,t} - E(R_{i,t}) \quad (5)$$

Donde se utiliza como estimación del rendimiento esperado del mercado la media de los rendimientos durante la ventana de estimación que va del día  $-271$  al día  $-21$  antes del evento. En nuestro caso la  $N$  será igual a 250.

$$E(R_{i,t}) = \frac{1}{N} \sum_{j=t-271}^{t-21} R_j \quad (6)$$

Para calcular los rendimientos anormales acumulados *Cumulative Abnormal Returns (CAR)* se sumaran los  $AR_t$  de los periodos correspondientes, de la siguiente manera. Los rendimientos anormales acumulados servirán para ver con que rapidez se ajusta el mercado a la información de los tuits de Trump.

$$CAR_{(T_1, T_2)} = \sum_{t=T_1}^{T_2} AR_t \quad (7)$$

*Brown y Warner (1980)* establecieron un test para comprobar la significación de los valores de *Abnormal Returns (AR)* y de *Cumulative Abnormal Returns (CAR)*. La ventaja de este test es que compensa la dependencia de los rendimientos a través de eventos estimando la desviación típica de los  $AR$  usando la ventana de estimación.

<sup>24</sup>Los inversores que creen que el precio de las acciones se le conoce como bajista o “bearish”. El oso representa el bear market, caracterizado por la bajada de los precios y las actitudes pesimistas de los inversores.

<sup>25</sup>Cuando la bolsa de valores cierre siempre en alza es propio de acciones “toro”. Con el bullish, no existen actitudes ni acciones negativas sino todo lo contrario. Este término se utiliza para describir situaciones financieras en las que el valor de un activo, índice bursátil o materia prima, cotiza en alza.

<sup>26</sup>El Sistema de Reserva Federal también conocido como FED (acrónimo de Federal Reserve System), es el modelo de sistema bancario existente en Estados Unidos que se ocupa de controlar la política monetaria del país y de ejercer el poder de supervisar al resto de instituciones bancarias con el objetivo de alcanzar estabilidad.

El t-test para el evento “0” examina si el valor de medida de rendimiento en el evento “0” estandarizado es diferente de 0. El t-test para este método tiene en cuenta la dependencia de los rendimientos en el mercado DJIA. Para este procedimiento, la desviación estándar del evento “0” se estima gracias a los valores de la ventana de estimación. Si los rendimientos para cada evento son normales, independientes y idénticamente distribuidos, entonces bajo la hipótesis nula de no significación del evento (explicada posteriormente), el rendimiento del evento “0” estandarizado se distribuye como una t de Student con  $248 N - 2$  grados de libertad.

El t-estadístico se calcula de la siguiente manera, para nuestro caso:

$$t_{AR} = \frac{AR_t}{\sigma_{\hat{AR}}} \quad (8)$$

donde:

$$\sigma_{\hat{AR}} = \sqrt{\frac{\sum_{t=-271}^{-21} (AR_t - \bar{AR})^2}{250}} \quad (9)$$

donde  $\sigma_{\hat{AR}}$  es la desviación estándar considerada para los *abnormal returns*,  $t = -271$  a  $t = -21$  es la ventana de estimación y 250 es el número de días de la ventana de estimación.  $\bar{AR}$  es la media de los rendimientos anormales medios sobre la ventana de estimación.

$$\bar{AR} = \frac{\sum_{t=-271}^{-21} AR_t}{250} \quad (10)$$

El valor crítico para el modelo CRM que se ha usado se calcula a través un una t de Student con  $N - 2$  grados de libertad, donde  $d$  son los días de la ventana de estimación, en nuestro caso 250 días.

El estadístico de test para los CAR se obtienen a través de:

$$t_{CAR} = \frac{CAR_{(T_1, T_2)}}{(T_2 - T_1 + 1)^{(1/2)} * \sigma_{\hat{AR}}} \quad (11)$$

Donde  $T_2 - T_1$  es la longitud de la ventana de evento correspondiente.

Para mirar si los tuits de Trump tienen efecto significativo sobre el mercado financiero.

$$\begin{aligned} H_0 : AR &= 0 \\ H_1 : AR &\neq 0 \end{aligned}$$

Si se rechaza la hipótesis nula, se puede decir que Trump si que tienen un impacto en los mercados financieros.

Los *CAR* se analizaran para ver la duración del efecto.

$$\begin{aligned} H_0 : CAR &= 0 \\ H_1 : CAR &\neq 0 \end{aligned}$$

Si se rechaza la hipótesis nula, la eficiencia del mercado vendra marcada por ver cuantos días duran los efectos de los *CAR*. Generalmente, si se encuentra que los *CAR* son superiores a 1 día, se puede hablar de ineficiencia del mercado.

(ecuaciones obtenidas de *Peterson, 1989* y *Rayarel, 2018* )

Una vez se han derivado todos los valores necesarios para el análisis del estudio de eventos se han escogido dos tuits significativos de la muestra de 117 tuits económico-financieros de Donald Trump para realizar este estudio.

Estos dos tuits se han seleccionado de la siguiente manera y gracias a la función que uno puede encontrar en el Anexo D.

Esta función de R recibe un tuit (con su número) (*num\_tweet*), una muestra de todos los tuits (*eco\_df*), una base de datos del índice económico; en nuestro caso el DJIA (*df*), los días anteriores del evento para la ventana de evento (*days\_before*), los días posteriores del evento para la ventana de evento (*days\_after*) y el total de días para la ventana de estimación (*estimation\_window*).

A partir de estos inputs, lo primero que se hace es establecer las fechas de la ventana del evento anteriores al evento y las fechas de la ventana del evento posteriores al evento. Se ha de tener muy en cuenta si el evento (tuit de Donald Trump) se produce antes o después del cierre de mercado, ya que si este se produce dentro del horario de mercado, el rendimiento del día 0 ya podría incorporar el efecto de su tuit, en cambio si el tuit se escribe después del cierre de los parqués, su efecto se vería en el día después del evento, esta función tienen en cuenta esta casuística. Una vez se ha hecho la correcta división de la ventana del evento se calcula los valores anormales, los valores acumulados anormales y sus correspondiente t-estadísticos. Finalmente esta función devuelve el resultado del test de medias entre los días antes del evento y después del evento (en ambos casos dentro de la ventana del evento), los valores de los rendimientos anormales  $AR_i$  para la ventana del evento, los valores acumulados de rendimientos anormales  $CAR_i$  para ventana de evento y los t-estadísticos de estas series.

Se ha pasado la base de datos de 117 tuits a esta función y esta ha seleccionado los tuits que:

- Para la ventana de evento escogida (10 días antes y 10 días después del evento) se han escogido aquellos tuits que al hacer un test de medias de la media 10 días antes y la media de los 10 días después, estas sean significativamente diferentes, ya que esto indicará que el tuit ha cambiado la tendencia del mercado.
- Al calcular los estadísticos t de los rendimientos anormales para los 20 días de la ventana de evento alguno sea significativamente diferente de 0. Lo que quiere decir, que el tuit ha fomentado que el rendimiento de algún día dentro de su ventana de evento sea considerado diferente al rendimiento normal de mercado.

De los 117 tuits de la muestra, sólo 4 de estos tuits cumplen con las condiciones anteriormente descritas. Se ha decidido coger dos de estos tuits y realizar un estudio de evento a a partir de estos tuits.

El primer de los tuits es el siguiente:



Figura 22: Tuit 1 para el estudio de eventos

En este tuit de la Figura 22 se ve un escrito del 24 de Diciembre de 2018 hablando sobre “el Muro” que el presidente Trump quería construir como frontera entre Estados Unidos y México. El tema del muro ha sido muy polémico durante toda la presidencia Trump a causa de el matiz anti-inmigración que considera. También tuvo relación en el ámbito económico, ya que la construcción de dicho muro cuesta muchos dinero y de eso es lo que habla en este tuit, en como se pagará el muro. Se puede ver que este tuit fue escrito por la tarde, más tarde del cierre del Dow Jones, por lo tanto sus efectos se verán el día siguiente al evento.

El otro tuit a analizar será:

En este segundo tuit, Figura 23, se trata de un hilo de tuits escritos el 18 de Febrero de 2020 en referencia a la guerra comercial con China. Trump en este tuit habla sobre el papel de las empresas americanas en el comercio internacional, dice que quiere que las empresas americanas vendan sus productos a China y a todo el mundo. Lo que Trump quiere aclarar en este hilo es que el quiere facilitar y no complicar el comercio internacional de las empresas americanas. En este hilo quiere quitar de la cabeza a sus ciudadanos que su política comercial va a llevar a las empresas a huir de Estados Unidos para ser competitivas internacionalmente.



Figura 23: Tuit 2 para el estudio de eventos

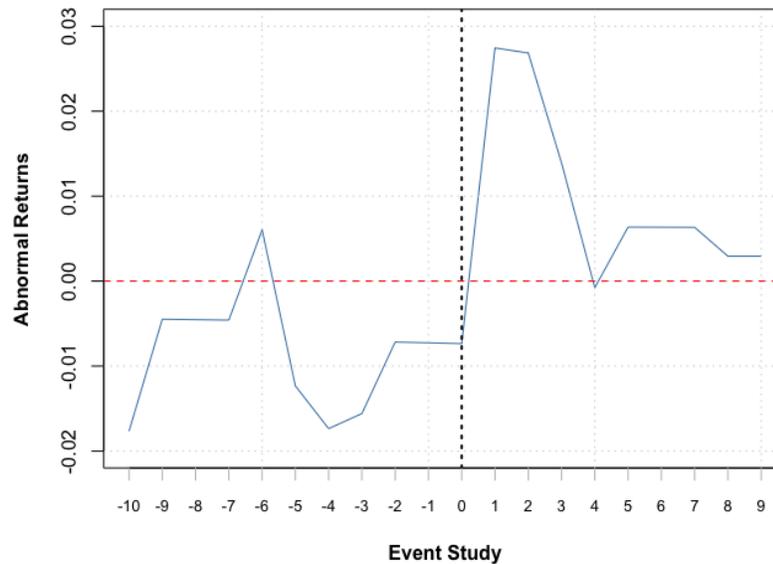


Figura 24: Abnormal Returns DJIA tuit 1

En la Figura 24 se puede ver los rendimientos anormales, es decir los rendimientos obtenidos una vez sustraída la media de la ventana de estimación para los 20 días de la ventana de evento. Se puede destacar como estos rendimientos son en general negativos para los días anteriores al evento, y positivos para los días posteriores. Aún más importante, el rendimiento anormal del día justo después del anuncio, por lo tanto el día que mas recoge el impacto del tuit, ya que el tuit se escribió fuera del horario de mercado, es el rendimiento más alto de toda la ventana de estimación, y

pasa de una bajada del -0.8% a casi una subida del 3%. También se puede ver como parece ser que este efecto se prolongó durante y hasta el segundo día de la ventana de estimación posterior al día 0. A continuación se estudiará en el Cuadro 13 los t-estadísticos de estos rendimientos anormales para ver si realmente fueron significativos o se pueden explicar como el movimiento normal del mercado.

Se recuerda que para las diferentes Tablas que contienen p-valores que aparecerán a partir de este punto se establece que se marcarán los p-valores inferiores a 0.10 con (\*), inferiores a 0.05 con (\*\*) y iguales a 0 con al menos 4 decimales iguales a 0 con (\*\*\*)).

Event Time	AR	T-statistic	P-value
-10	-0.0177	-2.54	0.0058**
-9	-0.0045	-0.65	0.2600
-8	-0.0045	-0.65	0.2600
-7	-0.0046	-0.66	0.2551
-6	0.0060	0.87	0.8065
-5	-0.0123	-1.77	0.0388**
-4	-0.0174	-2.45	0.0066**
-3	-0.0160	-2.24	0.0130**
-2	-0.0072	-1.03	0.1520
-1	-0.0073	-1.04	0.1488
0	-0.0074	-1.05	0.1456
1	0.0275	3.94	0.0000***
2	0.00268	3.85	0.0000***
3	0.0140	1.99	0.0233
4	-0.0008	-0.11	0.4559
5	0.0063	0.91	0.1809
6	0.0063	0.91	0.1809
7	0.0063	0.91	0.1809
8	0.0029	0.42	0.3400
9	0.0029	0.42	0.3400

Cuadro 13: Abnormal Returns para estudio de evento 1:

A partir de este Cuadro 13 se puede observar como hay valores estadísticamente diferentes al comportamiento normal del mercado antes del evento, seguramente a causa de anteriores alteraciones del mercado. Principalmente, lo que más destaca son los p-valores de 0 para los dos días posteriores al evento. Estos dos valores nos confirman que existe una correlación entre el tuit de Donald Trump y los posteriores movimientos del mercado, en concreto, parece indicar que el tuit tuvo un efecto positivo en el mercado. Esta Tabla 13 también nos permite ver la duración del efecto, se puede ver que el valor del segundo día también es significativo, al contrario que el valor del tercer día. Se podría decir que a partir del tercer día el mercado ya incorpora perfectamente la nueva información aportada por el presidente de los Estados Unidos y el mercado vuelve a su cauce normal guiado por la hipótesis de mercados eficientes, ya que la información ya es simétrica para todos los agentes. Esta información nos podría llevar a rechazar la afirmación (hipótesis nula  $H_0 : AR = 0$ ) que este tuit de Donald Trump no tuvo impacto en el mercado de valores.

A continuación se verá lo que pasa sobre estos mismos rendimientos acumulados.

Los rendimientos acumulados mostrados en la Figura 25 nos permiten tener una imagen más global de lo acontecido en el mercado. Estos rendimientos acumulados muestran una clara tendencia bajista antes del evento y alcista a partir del evento. Finalmente se ve como el mercado al final de la ventana de estudio vuelve a un nivel muy parecido al inicio de dicha ventana. En la siguiente Tabla 14 se ve la significación con el estadístico T de estos rendimientos acumulados.

En el Cuadro 14 se pueden ver los rendimientos acumulados para diferentes subventanas, dentro de la ventana de estimación. La ventana entre los días (-10,-1) de la ventana de evento nos muestra que la tendencia bajista para estos días es claramente diferente a la tendencia media del mercado, obtenida gracias al periodo de estimación. También se puede ver como las diferentes subventanas después del evento son significativas, lo que muestra que la tendencia alcista después del tuit también es significativamente diferente a la tendencia normal del mercado, al menos hasta el quinto día después del tuit. Finalmente se puede observar como el efecto total de la ventana de estudio (-10, 9) días no es significativamente diferente al movimiento normal del mercado. Por lo tanto, se podría decir que finalmente

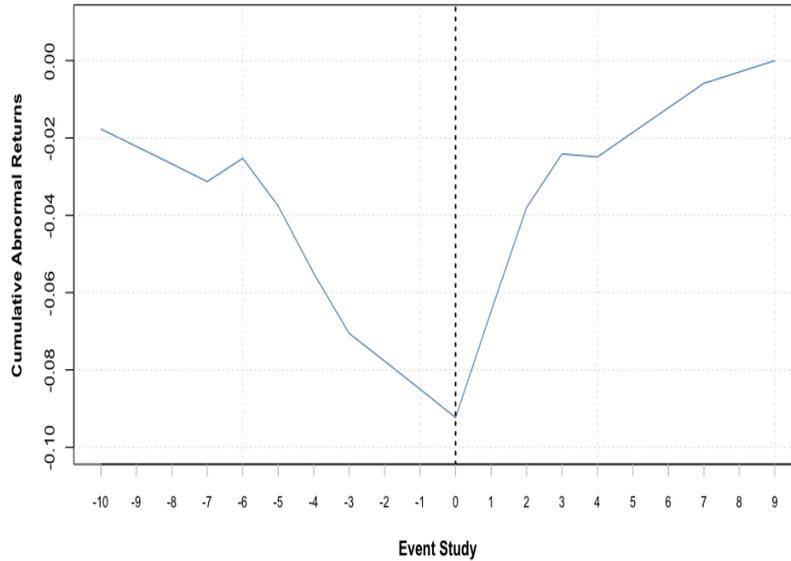


Figura 25: Cumulative Abnormal Returns tuit 1

Ventana de Evento	CAR	T-statistic	P-value
(-10,-1)	-0.0850	-3.86	0.0000***
(0,1)	0.0201	2.04	0.0212**
(0,2)	0.0470	3.90	0.0000***
(0,3)	0.0680	4.88	0.0000***
(0,4)	0.0601	3.86	0.0000***
(0,5)	0.0664	3.89	0.0000***
(-10,9)	0.0000	0	0.5000

Cuadro 14: Cumulative Abnormal Returns para estudio de evento 1:

el mercado ha incorporado esta información. A pesar de esto, los valores significativos anteriores nos hacen pensar que podríamos rechazar la hipótesis fuerte (hipótesis nula  $H_0 : CAR = 0$ ) de que los mercados son perfectamente eficientes ya que se ve como el mercado no es capaz de ajustar la nueva información de forma instantánea.

En este tuit, Donald Trump habla de la financiación de su famoso muro, una valla de seguridad construida por Estados Unidos en su frontera con México. Su objetivo es impedir la inmigración ilegal a los Estados Unidos. Su construcción se inició en 1994, en el gobierno de Bill Clinton, bajo el programa de lucha contra la inmigración ilegal pero tomó más entidad al ser un componente importante de las promesas de campaña del candidato y presidente Donald Trump. En 2017 estaba formado por una barrera física de unos 900 kilómetros de extensión en la zona fronteriza de Tijuana (Baja California) – San Diego (California). El muro incluye tres barreras de contención, iluminación de muy alta intensidad, detectores de movimiento, sensores electrónicos y equipos con visión nocturna conectados a la policía fronteriza estadounidense, así como vigilancia permanente con camionetas todoterreno y helicópteros artillados. Existen otros tramos de muro en los estados de Arizona y Nuevo México.

En su campaña para las elecciones presidenciales de Estados Unidos de 2016, el electo presidente Donald Trump se comprometió a construir un muro que abarcará toda la frontera sur de los Estados Unidos, y en repetidas ocasiones reiteró que México pagaría por su construcción a través del aumento de las tarifas de cruce de frontera y los aranceles del Tratado de Comercio de América del Norte (TLCAN)<sup>27</sup>

<sup>27</sup>El Tratado de Comercio de América del Norte es una zona de libre comercio entre Canadá, Estados Unidos y México. El Tratado permite reducir los costos para promover el intercambio de bienes entre los tres países.

En abril de 2017, Trump tuvo que renunciar a construir el muro en su primer año como presidente, como era su promesa. El Congreso de Estados Unidos no estaba dispuesto a permitir la inclusión de una partida en los presupuestos para la construcción del muro, y el gobierno tuvo que retirar esa partida para facilitar la aprobación del presupuesto. A este obstáculo se añadieron otros, como la resistencia de los propietarios de tierras en la zona fronteriza a ser expropiados

En este tuit de diciembre de 2018, el presidente Donald Trump explica que los fondos para construir dicho muro no deben pertenecer a la partidas presupuestaria de seguridad de los Estados Unidos de América, como se ha comentado, ya se había aclarado por parte del congreso que no se iban a destinar fondos americanos para este hecho. Donald Trump intenta dar más importancia a la construcción del muro, apartándolo del tema de seguridad y declarándolo cómo “Emergencia Nacional”. Además, añade que más dinero se perderá si no se construye el muro, que el dinero que llevará su construcción.

Este tuit es un precedente de lo que finalmente pasaría en 2019, dos meses más tarde cuando el 15 de febrero de 2019 Donald Trump declara el estado de emergencia nacional (mencionado en el tuit como *Shutdown*<sup>28</sup>) para acceder a fondos con los que construir el muro con México. Trump ha afirmado que firma la declaración para proteger al país de una “invasión de drogas y criminales” que llegan desde México y que ha calificado de un grave riesgo para la seguridad nacional.

Relacionándolo con el tema económico, se puede ver como esta noticia fue bien recibida por los inversores americanos, ya que después de este tuit vemos un aumento significativo de los rendimientos del Dow Jones. Como se ha visto en este análisis, la tendencia bursátil era bajista durante los días anteriores y a partir de este tuit, se puede observar un claro cambio de tendencia, hacia una dirección alcista. Este cambio se ve claro, en la respuesta al primer día después del tuit, donde hay un cambio estadísticamente significativo en unos rendimientos que pasan del bajar un 0.8% a subir prácticamente un 3%. Todo parece indicar que este tuit acabo con la incertidumbre que conllevaban por un lado, la falta de seguridad nacional de la que Donald Trump hablaba por la falta del muro, y por otro lado, la incertidumbre financiera que conllevaba la duda sobre de donde se iban a sacar los fondos para su construcción. Parece ser que esta reflexión del Presidente de los Estados Unidos despeja estas dudas de los inversionistas americanos.

De acuerdo con este razonamiento, diferentes artículos nos muestran el papel de la bolsa durante estos cierres de gobierno. Todo parece indicar que, a pesar de lo que sugeriría el sentido común, los mercados financieros no se ven afectados negativamente por los cierres de gobierno (*Shell, 2018*). Se quiere recordar que las conclusiones de este apartado no indican que los tuits de Donald Trump causaran estos movimiento bursátiles, sino que se ha encontrado evidencia significativa para indicar que existe una correlación entre los tuits del presidente y los resultados del Dow Jones.

Para completar este apartado, se realizará el mismo estudio de eventos pero con el segundo tuit escogido que habla de la guerra comercial entre China y Estados Unidos y el papel del gobierno americano sobre el libre comercio y la protección a sus compañías. Se recuerda que estos dos tuits han sido los único conjuntamente con dos más de ser seleccionados para el estudio de eventos por sus propiedades antes y después del evento analizadas a posteriori. Las conclusiones de estos dos tuits son muy significativas pero solo cuatro de los 117 tuits de Trump con contenido financiero han tenido estas propiedades.

En la Figura 26 se puede ver los rendimientos anormales, es decir los rendimientos obtenidos una vez substraída la media de la ventana de estimación para los 20 días de la ventana de evento. Se puede destacar como estos rendimientos son positivos para los días anteriores al evento, y negativos a partir del tercer día del evento. Esta figura parece que podemos distinguir un comportamiento positivo y constante durante los primeros 12 días de la ventana del evento y un comportamiento negativo y más variable durante los últimos 7 días de dicha ventana. A continuación se estudiará en el Cuadro 15 los T-estadísticos de estos rendimientos anormales para ver si realmente fueron significativos o se pueden explicar como el movimiento normal del mercado.

---

<sup>28</sup>En la política de los Estados Unidos, un cierre de la Administración, o cierre del Gobierno, es una situación en la que el gobierno suspende la prestación de todos los servicios públicos, menos los considerados como esenciales. En este caso el cierre parcial se da ante la negativa demócrata de autorizar cinco mil millones de dólares para la construcción del muro fronterizo con México y la oposición del presidente Trump a promulgar leyes que no impliquen dicha suma para el muro. El 25 de enero de 2019, el presidente Trump reabre el gobierno de manera parcial hasta el próximo 15 de febrero, esto como medida límite para llegar a un acuerdo con el senado de los Estados Unidos, dando a conocer que de no llegar a un acuerdo para dicha fecha, él mismo se encargará de los asuntos de seguridad del país.

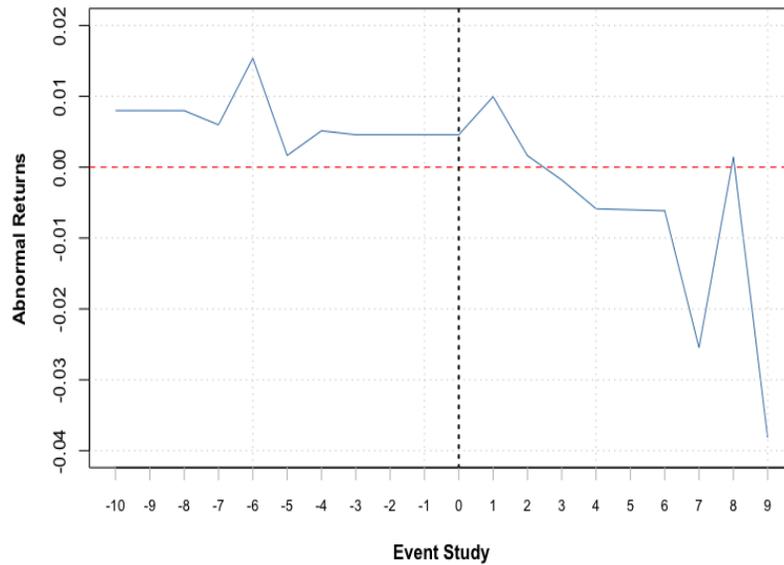


Figura 26: Abnormal Returns tuit 2

Event Time	AR	T-statistic	P-value
-10	0.0080	1.34	0.0905
-9	0.0080	1.34	0.0905
-8	0.0080	1.34	0.0905
-7	0.0060	1.00	0.1583
-6	0.0154	2.58	0.0052**
-5	0.0017	0.28	0.3904
-4	0.0051	0.86	0.1947
-3	0.0046	0.77	0.2212
-2	0.0046	0.77	0.2212
-1	0.0046	0.77	0.2212
0	0.0046	0.77	0.2212
1	0.0100	1.67	0.0489**
2	0.0016	0.27	0.3922
3	-0.0018	-0.30	0.3818
4	-0.0059	-0.98	0.1629
5	0.0060	-1.01	0.1511
6	-0.0061	-1.03	0.1513
7	-0.0255	-4.27	0.0000***
8	0.0014	0.24	0.4056
9	-0.0382	-6.41	0.0000***

Cuadro 15: Abnormal Returns para estudio de evento 2:

A partir de este Cuadro 15 se puede observar como hay valores estadísticamente diferentes al comportamiento normal del mercado para el sexto día antes del evento, seguramente a causa de otras anteriores alteraciones del mercado. Principalmente, lo que más destaca son los p-valores significativos para el día justo después del evento, cabe recordar que el tuit tuvo lugar fuera del horaria de mercado del día 0, por lo tanto su efecto se refleja en el día 1, y para los días 7 y 9 después del evento. Estos valores nos hacen pensar que existe una relación entre el tuit de Donald Trump y los posteriores movimientos del mercado, en concreto, parece indicar que el tuit tuvo un efecto positivo en el mercado justo para el día posterior, pero negativo y más fuerte para una semana después del anuncio. Tal como se ha visto en apartados anterior gracias al análisis de Granger, se ha comprobado que el efecto de los tuits de Donald Trump puede

tener un efecto sobre el Dow Jones con hasta 8 días de retardo, por lo que esta información no debería sorprender al lector. También se ha visto en el análisis de Granger como existe una dualidad temporal en los efectos de los tuits de Trump, en el sentido que los efectos de un tuit pueden ser positivos a corto plazo y negativos a largo plazo, como pasa en este caso.

A continuación se verá lo que pasa sobre estos mismos rendimientos acumulados.

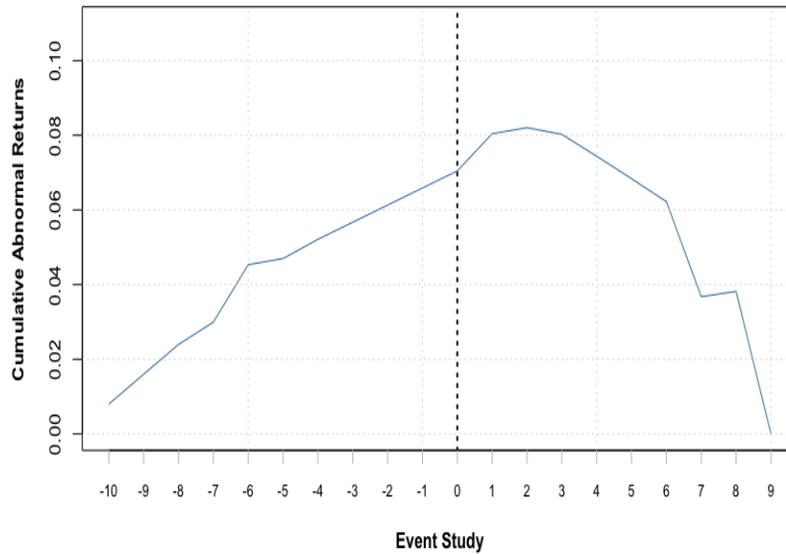


Figura 27: Cumulative Abnormal Returns tuit 2

Los rendimientos acumulados mostrados en la Figura 27 nos permiten tener una imagen más global de lo acontecido en el mercado. Estos rendimientos acumulados muestran una clara tendencia alcista antes del evento que se mantiene hasta tres días después del evento, a partir del tercer día después del evento, se ve una bajada muy acelerada. Finalmente se ve como el mercado al final de la ventana de estudio vuelve a un nivel muy parecido que el del inicio de dicha ventana. En el siguiente cuadro se ve la significación con el estadístico T de estos rendimientos acumulados.

Ventana de Evento	CAR	T-statistic	P-value
(-10,-1)	0.0659	3.50	0.0003**
(0,1)	0.0145	1.72	0.0433**
(0,2)	0.0162	1.57	0.0589
(0,3)	0.0143	1.20	0.1156
(0,4)	0.0085	0.63	0.2646
(0,5)	0.0025	0.17	0.4326
(0,6)	-0.0037	-0.23	0.4091
(0,7)	-0.0291	-1.72	0.0433**
(-10,9)	0.0000	0	0.5000

Cuadro 16: Cumulative Abnormal Returns para estudio de evento 2:

En el Cuadro 16 se pueden ver los rendimientos acumulados para diferentes subventanas, dentro de la ventana de estimación. La ventana entre los días (-10,-1) del evento nos muestra que la tendencia alcista para estos días es claramente diferente a la tendencia media del mercado, obtenida gracias al periodo de estimación. También se puede ver como de las diferentes subventanas después del evento, solo son significativas la ventana correspondiente al primer día del evento, con un coeficiente positivo significativo. Esto muestra que el efecto inmediato del tuit fue positivo

y estadísticamente diferente de 0 para el primer día. Por otro lado, se puede ver como la bajada que se produce entre los días 0 y 6 no es significativamente diferente al rendimiento normal de mercado, pero si que lo es cuando añadimos el séptimo día, por lo que parece indicar que las repercusiones del tuit podrían ser negativas a partir de la semana posterior al tuit. Finalmente se puede observar como el efecto total de la ventana de estudio (-10, 9) días no es significativamente diferente al movimiento normal del mercado. Como resultado, se podría decir que finalmente el mercado ha incorporado esta información. A pesar de esto, los valores significativos anteriores nos hacen pensar que podríamos rechazar la hipótesis fuerte (hipótesis nula  $H_0 : CAR = 0$ ) de que los mercados son perfectamente eficientes ya que se ve como el mercado no es capaz de ajustar la nueva información de forma instantánea.

En este tuit de Febrero de 2018, El presidente Donald Trump establece que quiere evitar las políticas que dificultan a los países extranjeros hacer negocios con los Estados Unidos, precisamente, quiere afirmar que su administración esta intentando lo contrario. El presidente durante su mandato ha impuesto aranceles de millones de dolares a diferentes bienes y ha iniciado disputas comerciales con los principales socios comerciales americanos, pero justamente en este tuit intenta inducir lo contrario. Un ejemplo de estos aranceles son los impuestos sobre el acero y el aluminio en el mundo entero, justificados por Donald Trump como obligación de seguridad nacional. Estas imposiciones han dificultado las negociaciones comerciales entre las compañías automovilísticas europeas y las empresas americanas.

La consecuencia que más ha afectado a la economía americana ha sido la guerra comercial con China que han provocado estos tipos de aranceles.<sup>29</sup> Trump incluso ha anunciado aranceles sobre prácticamente el 100% de los bienes importados desde China. Además, los socios comerciales de Estados Unidos no se han quedado al margen, tomando represalias con aranceles dañinos sobre productos estadounidenses como bourbon, motocicletas y productos agrícolas, lo que obligó al gobierno a proporcionar millones en ayuda a los agricultores.

De hecho, las empresas estadounidenses se han visto obligadas a cambiar sus operaciones o encontrar nuevas fuentes para sus componentes, o abogar por exenciones de las políticas comerciales. Asimismo, los compradores extranjeros de productos estadounidenses se han ido a otros mercados. Con cada ronda de aranceles Trump ha intimidado a más empresas estadounidenses para que se vuelvan proteccionistas. Para muchos estadounidenses, los costes superiores resultantes de sus aranceles significan que ya no pueden competir con empresas extranjeras en el mercado estadounidense o mundial.

La verdad es que si la intención de Donald Trump con esta tuit era aumentar la confianza de las compañías americanas en las políticas comerciales de su presidencia, no se puede decir que fuera un éxito. Si que es verdad, que el día inmediatamente posterior al tuit, los rendimientos bursátiles del Dow Jones fueron estadísticamente positivos, más que el rendimiento normal del mercado, pero si miramos el conjunto entero del evento, se puede ver como a partir del tercer días después del tuit, tuvo lugar una fuerte caída en el Dow Jones. Una caída importante durante la semana posterior al evento. Por lo que parece, aunque inicialmente este tuit fue bien recibido por los inversionistas, al final parece ser que estos no se creyeron las palabras del presidente, sino que se acogieron a los actos de este, y sus actos no eran muy prometedores para los incentivos de estas empresas. Cabe recordar que estas políticas proteccionistas de Donald Trump, incorporando aranceles a los productos extranjeros hace que los que producen los bienes específicos objeto de los aranceles generalmente se benefician de la protección que confieren. Los beneficios generados para los que laboran en los sectores que compiten con las exportaciones deben equilibrarse con las pérdidas sufridas por otros grupos de trabajadores. En primer lugar, muchos se desempeñan en fábricas que utilizan bienes importados como inputs en sus procesos de producción. Por un lado, el aumento del costo de los artículos importados debido a los aranceles afecta el proceso productivo y suele generar la pérdida de puestos de trabajo. Segundo, cuando Estados Unidos impone aranceles de forma unilateral generalmente sus socios comerciales hacen lo propio con carácter de represalia, lo que puede limitar la producción de exportaciones estadounidenses y, en última instancia, perjudicar a los trabajadores de esos sectores industriales. En consecuencia, puede decirse en términos generales que los aranceles de Trump han ayudado a algunos trabajadores y perjudicado a otros. Esta situación no tiene nada de sorprendente, la política comercial casi siempre genera importantes efectos distributivos y todo cambio que se le efectúe implica la decisión de beneficiar a ciertos grupos en detrimento de otros. No obstante, cuando los economistas han procurado sumar los efectos netos de los aranceles de Trump respecto del empleo, en general las mejoras logradas en los sectores que compiten con las exportaciones parecen haber quedado más que compensadas por las pérdidas en las industrias que utilizan bienes importados y cuyas exportaciones son objeto de represalias (*Gertz y Kharas, 2016*).

---

<sup>29</sup>La guerra comercial entre China y los Estados Unidos es un conflicto comercial iniciado en marzo de 2018, después de que el presidente de los Estados Unidos, Donald Trump anunciase la intención de imponer aranceles de 50 000 millones de dólares a los productos chinos bajo el artículo 301 de la Ley de Comercio de 1974, argumentando un historial de «prácticas desleales de comercio», robo de propiedad intelectual y una transferencia forzada de tecnología americana a China. En represalia, el gobierno de la República Popular China impuso aranceles a más de 128 productos estadounidenses, incluyendo en particular la soja, una de las principales exportaciones de Estados Unidos a China.

## 8. Predicción y Machine Learning: Support Vector Machines

El objetivo de este apartado es realizar un modelo de clasificación no lineal como son las Support Vector Machines para poder predecir si para cualquier día  $t$  el mercado bursátil americano, seguimos con el Dow Jones, va a tener un rendimiento positivo o negativo, es decir, si ese día subirá o bajará el mercado. Además, se quiere observar y contrastar las conclusiones obtenidas en el apartado anterior de causalidad de Granger. Como se ha visto, se han encontrado retrasos de los indicadores sentimentales que tienen poder predictivo sobre si el precio subirá o bajará del mercado. En este apartado se quiere contrastar dicha información. Volviendo un poco a las raíces del trabajo, se recuerda que estos indicadores sentimentales están contruidos gracias a un análisis sentimental de los tuits del presidente Donald Trump, como resultado, se esta intentando ver si los tuits de Donald Trump tienen influencia sobre el mercado americano.

Para ello, se construirán dos modelos diferentes, en se intentará predecir si el rendimiento sera positivo o negativo sólo con la información de los rendimientos bursátiles de los días anteriores. En el otro modelo también se añadirá información sobre el único retardo significativo que se ha encontrado en la sección de Granger sobre la variable  $binary.DJIA_t$  y que afecta a toda la presidencia Trump completa.

Se quiere comentar antes de iniciar esta sección, que se han hecho varias pruebas con juegos diferentes de variables sentimentales y retardos para el modelo SVM. Por ejemplo, tener en cuenta el Ratio de Negatividad para los tres días anteriores y no sólo para el tercer día anterior. Finalmente, los resultados más ambiciosos han sido los considerados para esta sección.

### 8.1. Support Vector Machines

Las Máquinas de Vector Soporte (SVMs) son un conjunto de algoritmos de aprendizaje supervisados que desarrollan métodos relacionados con los problemas de clasificación y regresión. Como la mayoría de lo métodos de clasificación supervisada, los datos de entrada, (puntos relacionados con los días bursátiles a clasificar, en nuestro caso) son vistos como un vector  $p$ -dimensional, listado de  $p$  puntos. Dado un conjunto de puntos como un subconjunto de un espacio mayor, en el que cada uno de esto puntos pertenece a una de dos posibles categorías (modalidades de la variable respuesta), en nuestro caso si ese día la bolsa subirá (1) o bajará (0), se define el algoritmo de SVM como la construcción de un modelo capaz de predecir si un nuevo día, del cual no se conoce si acabará siendo vela verde o vela roja<sup>30</sup> (subiendo o bajando en términos financieros) pertenecerá a una categoría o a la otra.

SVM es un modelo, que a partir de un conjunto de datos de entrenamiento, entre los cuales se pueden etiquetar las diferentes clases de la variable respuesta y representar estas muestra como puntos en el espacio para tratar de separar las diferentes clases mediante el espacio más amplio posible. Entonces, cuando los datos de test se ponen en correspondencia con el modelo creado, puedan ser clasificados en función de su proximidad. Este es el concepto de “separación óptima”, donde reside la característica principal de las SVM. Estos algoritmos buscan el hiper-plano que tenga la máxima distancia (a partir de ahora a esta distancia la llamaremos margen), con los puntos que estén situados más cerca de si mismo. De esta forma los puntos del vector que estén etiquetados en una categoría estarán en un lado del hiper-plano, y los puntos que se encuentren en la otra categoría están al otro lado del hiper-plano. El funcionamiento de estos modelos se detallará en el Anexo E con un ejemplo visual con dos variables que facilitará su comprensión.

A través del siguiente esquema, se muestra el procedimiento seguido por las SVM:

---

<sup>30</sup>Las velas en trading indican la puja de los precios del mercado, su valor en la apertura y en el cierre de cada sesión bursátil. Cuando el precio de apertura es menor que el de cierre la vela se pinta de verde o blanco, lo que indica que la cotización subió (vela alcista), y si es al contrario, que el precio de apertura sea mayor que el de cierre, la vela se refleja en rojo o negro, lo que significa que la cotización bajó (vela bajista). Cuando ambos precios son iguales se le conoce como vela neutral.

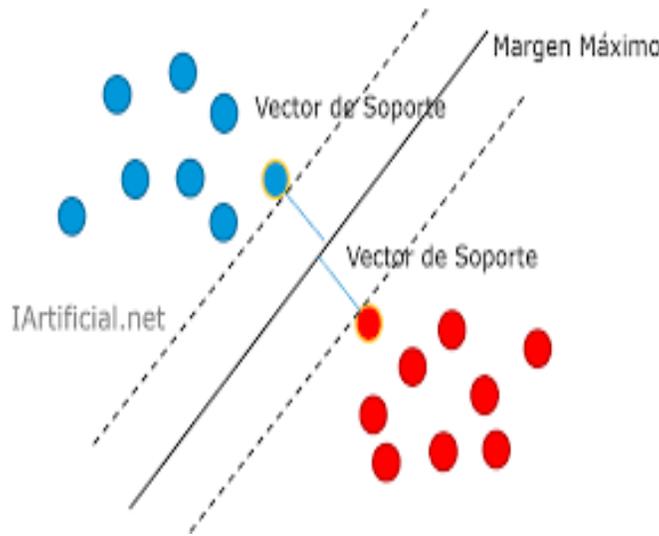


Figura 28: Funcionamiento de las SVM  
Fuente: Heras, s.f.

En general, las formas más sencillas de realizar la separación es mediante una línea recta, un plano recto o un hiperplano N-dimensional, pero los universos a clasificar no suelen presentar la situación ideal de dos dimensiones como pasa en el ejemplo gráfico anterior de la Figura 28, sino que un algoritmo SVM acostumbra a tratar con más de dos variables predictoras, curvas de separación no lineales o casos donde los conjuntos de datos no son completamente separables. La representación con funciones *Kernel* ofrece una solución a este problema, proyectando la información en un espacio de características de mayor dimensión, el cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal, como podría ser la utilización de un Kernel Radial, Gaussiano etc.

Un paso previo al método de clasificación, es la partición del conjunto de datos en dos conjuntos más pequeños que se utilizarán con diferentes propósitos. Estos dos subconjuntos de datos son los llamados de entrenamiento (*Train*) y los de test (*Test*). El subconjunto de datos de entrenamiento se utilizará para estimar los parámetros del modelo y el modelo estimado. Cada registro de la base de datos tiene que aparecer en uno de los dos subconjuntos. El concepto es el de entrenar el modelo con un conjunto de datos independientes de los datos con los que se realizará la validación del modelo.

Como resultado de aplicar un método de clasificación como la Support Vector Machines, se pueden cometer dos errores. En el caso de nuestra variable respuesta, se trata de una variable respuesta binaria, la cual toma dos valores:

- Binary = 0: El mercado bursátil baja respecto el valor de cierre del día anterior.
- Binary = 1: El mercado bursátil sube respecto el valor de cierre del día anterior.

Por lo tanto, habrá 1's que se clasifiquen incorrectamente como 0's y 0's que se clasifiquen incorrectamente como 1's. A partir de este recuento, se puede crear el siguiente cuadro de clasificación. Este se utilizará en los próximos apartados para medir la calidad de los modelos.

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	$P_{1,1}$	$P_{1,2}$
= 1	$P_{2,1}$	$P_{2,2}$

Cuadro 17: Tabla de confusión

En este Cuadro 17 los  $P_{1,1}$  (True Negative: Verdaderos Negativos) y  $P_{2,2}$  (True Positive: Verdaderos Positivos) corresponden a las predicciones correctas, mientras que los valores de  $P_{1,2}$  (False Positive: Falsos Positivos) y  $P_{2,1}$  (False Negative: Falsos Negativos) serán predicciones erróneas. La validez del modelo será la suma de las predicciones correctas entre el total de predicciones.

$$Accuracy(\%) = 100 * \frac{P_{1,1} + P_{2,2}}{n}$$

También se definen las siguientes medidas que se calcularán para cada método de la siguiente forma:

$$ErrorRate(\%) = 100 * \frac{P_{2,1} + P_{1,2}}{n}$$

Donde  $n$  es el número total de predicciones realizadas.

Se define la precisión (*precision*) del modelo como el número de registros recuperados que son relevantes (1's). Nótese que su valor aumenta a medida que el número de falsos positivos disminuye.

$$Precision(\%) = 100 * \frac{P_{2,2}}{P_{2,2} + P_{1,2}}$$

Se define la *recall* del modelo como la fracción de registros relevantes que han sido recuperados. Cuanto más cercano a 1, mejor estarán definidas las distintas clases existentes ya que su valor aumenta a medida que disminuye el número de falsos negativos.

$$Recall(\%) = 100 * \frac{P_{2,2}}{P_{2,2} + P_{2,1}}$$

Se define la *F-measure* como la medida de precisión que tiene un test. Se utiliza la determinación de un valor único ponderando la precisión ( $P$ ) y el *recall* ( $R$ ):

$$F - measure(\%) = \frac{2 * P * R}{P + R} = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

A continuación, se definirá el protocolo de validación. En primer lugar se especificará un 30% de la totalidad de los datos como datos de test. Esta muestra de test será una colección de los últimos 416 días (30% de 1386 días = 416) bursátiles (incluidas las interpolaciones) que incluyen todos los días desde el 9 de Septiembre de 2019 a, el último día del cual se tienen datos, el 6 de Noviembre de 2020<sup>31</sup>. La muestra de *test* se utilizará al final del proceso cuando los modelos óptimos para cada condición se hayan recogido para testear la bondad de ajuste de estos modelos óptimos en condiciones independientes a los datos de *training*.

Tipos de Datos	Training	Test	Total
Número de días bursátiles	970 (70 %)	416 (30 %)	1386 (100 %)

Cuadro 18: Distribución de los datos: SVM

De los 970 días bursátiles restantes, días considerados como muestra de *training* se dividirán aleatoriamente en datos de *learning* y datos de *validation*. Con los datos de *learning* se realizarán varios modelos con diferentes parámetros y con los datos de *validation* se analizará cual es el mejor modelo sin variables de opinión y cual es el mejor modelo con variables de opinión.

Tipos de Datos	Learning	Validation	Test	Total
Número de días bursátiles	776 (56 %)	194 (14 %)	416 (30 %)	1386 (100 %)

Cuadro 19: Distribución de los datos con *learning*: SVM

Tal y como se ha dicho anteriormente, en este proceso se estimará mediante los datos de *learning* y se escogerá el modelo óptimo de la siguiente manera:

<sup>31</sup>Se han cogido 1386 días disponible y no los 1389 disponible originales entre el 17/01/2017 al 06/11/2020 porque al utilizar variables retardadas de sentimiento con retardo de hasta 3 días, no se tiene información completa para los primeros tres días.

En primer lugar se creará un modelo lineal, es decir, que la separación entre las dos clases será un hiper-plano de  $N$  dimensiones. Se probará para diferentes valores de  $C$ , con el objetivo de encontrar el valor de  $C$  que optimiza el modelo. Se entiende el parámetro  $C$  como la tolerancia a permitir puntos en el lado incorrecto del hiper-plano. Un valor bajo de  $C$  permite más puntos mal clasificados, en cambio, un valor de  $C$  alto, conllevará un margen más estrecho. Reducir la  $C$  hace perder eficiencia del separador lineal pero hace ganar estabilidad al modelo, en el sentido que la influencia de cada punto esta limitada por  $C$ . Por lo tanto, existe un *trade off* entre margen y error de entrenamiento. En el siguiente gráfico, se puede ver la influencia de  $C$ :

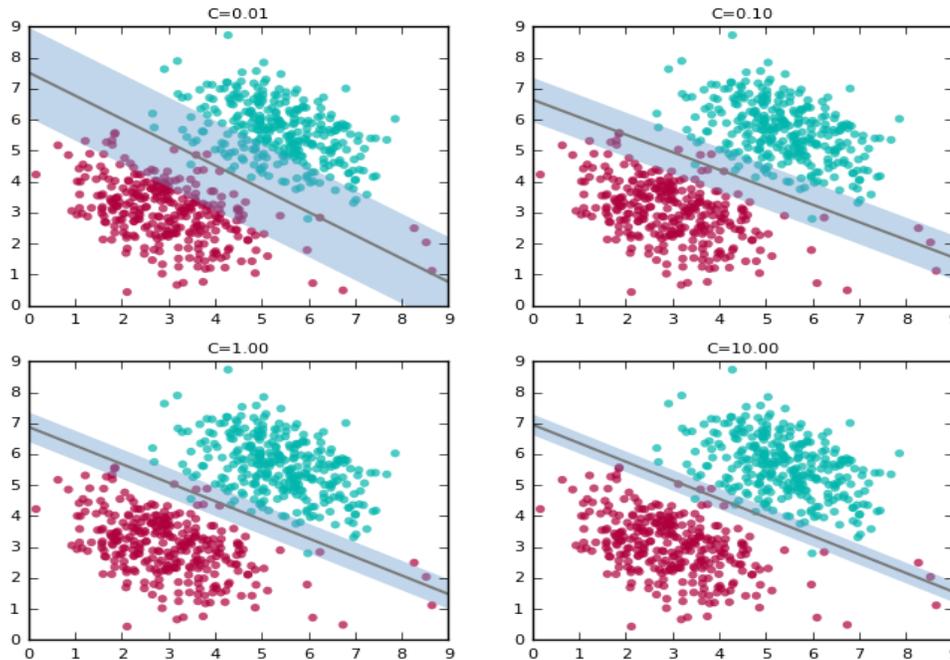


Figura 29: Efecto del parámetro  $C$  sobre las SVM  
Fuente: Ghose, 2017

En esta Figura 29 se puede ver como después de crear diferentes hiper-planos para diferentes valores de  $C$ , estos se corresponden con diferentes amplitudes del margen. A mayor  $C$ , más estrecho el el margen. El procedimiento usado se quedará con aquel valor de  $C$  que minimice el *error-rate* para los datos de validación.

Seguidamente, se volverá a construir el modelo pero ahora la separación entre las clases no tiene porque ser lineal. Este modelo SVM con Kernel Radial depende de  $C$  (como en el caso anterior) pero también de  $\Gamma$ , la cual mide la complejidad del modelo SVM. Un valor de  $\Gamma$  demasiado bajo podría hacer que nuestro modelo fuera demasiado simple, y que, por lo tanto no recogiese bien toda la información de los datos. En cambio, un valor de  $\Gamma$  demasiado alto nos llevaría a un modelo sobre ajustado que no funcionaria bien en *test*. En este caso se probaran diferentes valores de  $C$  y de  $\Gamma$  para obtener el modelo que mejor se ajuste a los datos. Esto ha sido posible gracias a la función de  $R$ : *"tune.SVM"* que permite ajustar diferentes modelos según un abanico de parámetros y finalmente escoger el mejor modelo, más tarde se hablará de esta función. De la misma forma que en el caso anterior, se crearan diferentes modelos para acabar escogiendo aquel con los valores de  $C$  y  $\Gamma$  que minimicen el *error-rate*. La división entre *learning* y *validation* se realizará de la misma manera que el caso anterior.

Además, también se tendrá en cuenta el concepto de *cross-validation*. La validación cruzada o *cross-validation* es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica. Es una técnica muy utilizada en proyectos de inteligencia artificial para validar modelos generados. Más adelante, se explicará como se ha realizado la *cross-validation* en esta sección.

## 8.2. Support Vector Machines sin indicadores sentimentales

En primer lugar, hace falta remarcar que para la creación de estos modelos se utilizarán solo valores retrasados de los rendimientos como predictores de la variable binaria que establece si el precio ha subido o bajado durante ese día. En esta caso intentaremos predecir si la bolsa subirá o bajará para un día concreto basándonos en los rendimientos bursátiles de los tres días anteriores.

$$I_0 = \{DJIA_{t-3,2,1}\}$$

A continuación, se calculará el hiper-plano de N dimensiones que mejor separe los puntos para diferentes valores de C. Para cada uno de estos hiper-planos, se calculará la tasa de error. El resumen de los resultados se puede ver en la Figura 30.

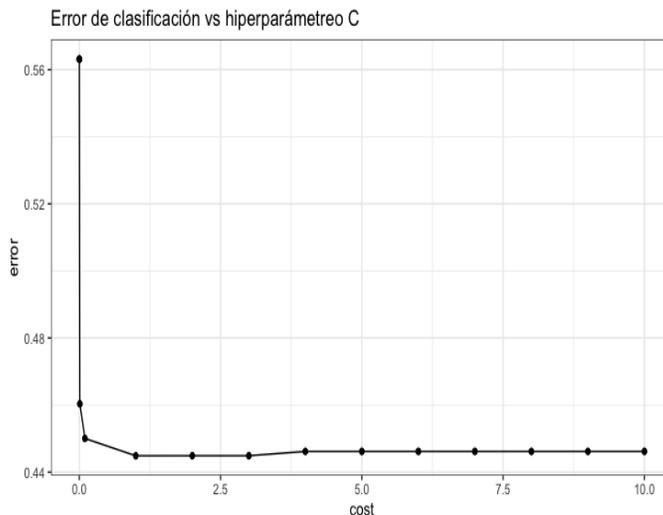


Figura 30: Efecto del parámetro C sobre las SVM en Kernel Lineal

En la Figura 30 se observa el papel del parámetro C respecto al error de clasificación. Se puede ver como para valores de C muy bajos el error es mayor, a medida que aumentamos C el error va disminuyendo. A pesar de eso, el hecho de aumentar C también hace aumentar el margen, es decir aumenta la tolerancia del modelo. Por ello, se ve como el valor de C que minimiza el error de clasificación se encuentra cerca de  $C = 1$ . Esto ha sido posible gracias a la función de *R* `tune.SVM` que te permite ajustar diferentes modelos en función de un abanico de parámetros i finalmente te permite escoger el mejor modelo según el *error-rate*.

Seguidamente, se muestra el código de *R* de dos funciones. Una para encontrar los parámetros óptimos en los modelos de validación; como se ha explicado anteriormente con el parámetro C, y por otro lado la función para construir estos modelos SVM. Estas dos funciones simplemente se muestra como ejemplo para uno de los casos de modelos SVM contruidos, el resto de funciones para los diferentes modelos y parámetros se pueden encontrar en el Anexo E.

```
svm_cv <- tune("svm", binary ~ lag1 + lag2 + lag3,data= DJIA .SVM[learning,] ,
  Kernel="linear",scale=TRUE,
  ranges = list(cost = c(0.00001,0.0001,0.001,0.01,0.1,1:10 )),class.
  weights = c("0" = 1.25, "1" = 1), tunecontrol = tune.control(
  cross=10))

svm1 <- svm(binary ~ lag1 + lag2 + lag3,data= DJIA .SVM ,Kernel="linear",scale=
  TRUE,subset=learning, cost = 1,
  class.weights = c("0" = 1.25, "1" = 1))
```

En la primera función se busca el mejor parámetro de coste entre diferentes parámetros que van desde 0.00001 a 10 gracias al campo *range::cost*. También se puede ver como en el campo *class.weights* se indica como nuestra variable respuesta *binary.DJIA<sub>t</sub>* tiene un 25% más de 1's que de 0's, por lo que aquí le indicamos que es importante que detecte los 0's también, aumentado su peso en el modelo, esto es simplemente una pequeña corrección para calibrar la variable respuesta. Además, se puede ver como en el campo *scale* se deciden escalar las variables predictoras ya que tienen dimensiones diferentes. De la misma manera, se observa como estos modelos se construyen con los datos de *learning* y con un *Kernel* lineal.

Además, por defecto la función *tune* de R realiza una *cross-validation* de 10 iteraciones, cada una con una partición diferente de los datos hecha de forma aleatoria. En cada una de las 10 iteraciones de este tipo de validación se realiza un cálculo de error. El resultado final lo obtenemos a partir de realizar la media aritmética de los 10 valores de errores obtenidos con diferentes particiones de los datos. De esta forma garantizamos que se escogen los parámetros óptimos independientemente de la partición de los datos. Los valores de la Figura 30 corresponden a las medias de esta validación cruzada de 10 iteraciones, por lo tanto, cada punto es la media de 10 modelos diferentes donde estos modelos han sido creados con particiones diferentes entre *learning* y *validation* de los datos.

Una vez ya conocemos como funcionan los modelos a estimar, se estima el primer modelo sin variables sentimentales con los datos de *learning* y un separador lineal.

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	146	199
= 1	122	309

Cuadro 20: Tabla de confusión Kernel Lineal Learning sin variables sentimentales

En el Cuadro 20 se puede ver la matriz de confusión para este primer modelo creado.

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	42	54
= 1	18	80

Cuadro 21: Tabla de confusión Kernel Lineal Validation sin variables sentimentales

En el Cuadro 21 se puede ver la matriz de confusión para un modelo de separación lineal sin variable sentimentales pero ahora con los datos de *validation*.

Las conclusiones de este primer modelo en *learning* y *validation* se pueden ver en el siguiente Cuadro 22.

Datos	Accuracy	Error Rate	Precision	Recall	F-measure
Learning	58.63 %	41.46 %	54.47 %	42.31 %	47.63 %
Validation	62.88 %	37.11 %	70.00 %	43.75 %	53.85 %

Cuadro 22: Medidas de bondad Kernel Lineal sin variables sentimentales

En el Cuadro 22 se pueden ver las diferentes medidas de bondad, explicadas anteriormente para un modelo SVM de separación lineal sin variables sentimentales. Se recuerda que los datos de *learning* y *validation* son una muestra aleatoria de los datos de *training* del 80% i del 20% respectivamente de los datos de *training*. Se puede ver como la *accuracy* del modelo de *validation* es mayor que en *learning*, por lo que el modelo acierta más en *validation* ya que esto es lo que mide la *accuracy* del modelo. En los otros estadísticos como *error-rate*, *precision*, *recall* y *F-measure* también funciona mejor el modelo de *validation*. El mismo análisis se realizará pero ahora con un separador de Kernel Radial.

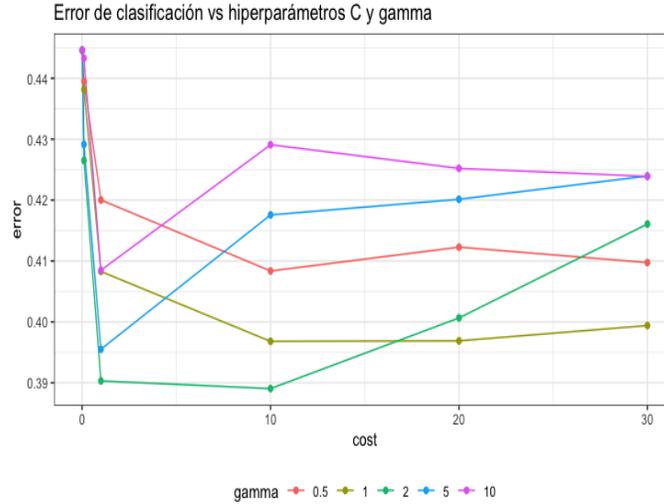


Figura 31: Efecto del parámetro C y Gamma sobre las SVM en Kernel Radial

El separador de Kernel Radial, como se ha comentado anteriormente, permite hacer separaciones no lineales de la variable respuesta. Dicho separador depende de dos parámetros, por un lado, el parámetro  $C$  y por el otro el parámetro  $Gamma$  que mide la complejidad del modelo. En la Figura 31 se puede ver el comportamiento de estos dos parámetros respecto al error de clasificación para nuestro modelo de Kernel Radial. Se observa que el error mínimo, *cross-validado* se consigue con un valor de  $C = 10$  y de  $Gamma = 2$ . Estas referencias serán las que se utilizarán para los modelos de Kernel Radial de esta subsección.

En el Cuadro 23 se puede ver la matriz de confusión para este modelo creado con separador Radial.

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	280	69
= 1	125	306

Cuadro 23: Tabla de confusión Kernel Radial Learning sin variables sentimentales

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	60	36
= 1	36	62

Cuadro 24: Tabla de confusión Kernel Radial Validation sin variables sentimentales

En el Cuadro 24 se puede ver la matriz de confusión para un modelo de separación lineal sin variable sentimentales pero ahora con los datos de *validation*.

Las conclusiones de este segundo modelo en *learning* y *validation* se pueden ver en el siguiente cuadro.

Datos	Accuracy	Error Rate	Precision	Recall	F-measure
Learning	75.52 %	24.48 %	69.14 %	81.16 %	74.66 %
Validation	62.89 %	37.11 %	62.50 %	62.50 %	62.50 %

Cuadro 25: Medidas de bondad Kernel Radial sin variables sentimentales

En el Cuadro 25 se pueden ver las diferentes medidas de bondad, explicadas anteriormente para un modelo SVM de separación lineal sin variables sentimentales. Se puede ver como la *accuracy* del modelo de *validation* es inferior que en *learning*, por lo que el modelo acierta más en *learning* ya que esto es lo que mide el modelo. En los otros estadísticos como *Error-rate*, *precision*, *recall* y *F-measure* también funciona mejor el modelo de *learning*.

Lo que interesa ahora, es comparar los resultados obtenidos con un separador Lineal y un separador Radial en los modelos sin indicadores sentimentales, para escoger el modelo óptimo con variables procedentes únicamente de los mercados financieros. Para ello, compararemos los resultados obtenidos bajo la muestra de validación en los dos casos ya que esta muestra no ha formado parte de la construcción del modelo. En este caso se puede ver los valores de *accuracy* son superiores en el modelo Radial (62.88 % vs 62.89 %), aunque la diferencia sea mínima. Por construcción si la *accuracy* del modelo Lineal es inferior, mayor será su *error-rate* respecto al modelo Radial. La precisión (70.00 % vs 62.50 %) es favorable al modelo Lineal pero no así el *recall* (43.75 % vs 62.50 %) y el *f-measure* (53.85 % vs 62.50 %) donde los valores del modelo Radial son superiores.

En este caso en concreto donde se intenta predecir si la bolsa subirá o bajará respecto a la información del análisis sentimental de los tuits del presidente Trump se considera que estadísticos como *precisión*, *recall* o su ponderación *f-measure* que se centran en representar el número de registros recuperados únicamente, es decir en aquellos valores que la variable respuesta tiene valor de 1 y nuestros modelos es capaz de clasificar como 1's no son medidas claras sobre la calidad de la clasificación. El objetivo de este proyecto no es simplemente acertar cuando va subir el mercado sino también queremos darle mucha importancia a acertar cuando el mercado va a bajar.<sup>32</sup> Por estos motivos se ha decidido centrarnos en la *accuracy* en la selección de modelo óptimo de cada apartado.

Como resultado de esta metodología se ha decidido seleccionar el modelo de Kernel Radial con parámetros de coste  $C = 10$  y  $\text{Gamma} = 2$  como modelo que mejor funciona con únicamente las variables financieras retrasadas tres días para intentar predecir si el mercado bursátil bajará o subirá.

### 8.3. Support Vector Machines con indicadores sentimentales

En este apartado se añadirá el único retardo, de una de las 4 variables sentimentales, significativo para el análisis de Granger sobre la variable  $\text{binary.DJIA}_t$ . Por lo tanto los modelos que se estimaran en esta subsección serán de la siguiente forma:

$$I_1 = \{DJIA_{t-3,2,1}, \text{Ratio.Neg}_{t-3}\}$$

Se intentará predecir si el mercado subirá o bajará el día  $t$  en función de lo que haya hecho el mercado los días anteriores y en función de la proporción de tuits negativos de Donald Trump respecto al total de tuits de tres días antes. En este caso también se propuso añadir no solo la información del ratio de negatividad para tres días antes, sino para un día y dos días antes pero se concluyó que no aportaban capacidad predictiva al modelo. En esta sub-sección, se utilizará el mismo proceso que en el apartado anterior, se empezará definiendo el valor de coste óptimo para un modelo SVM de separación Lineal y se verán sus resultados. A continuación se escogerán los valores óptimos de  $C$  y de  $\text{Gamma}$  para un modelo SVM de separación Radial y se verán sus resultados. Finalmente gracias a los datos de validación, se escogerá el modelo que mejor explica la relación entre estas variables financieras y sentimentales para predecir el comportamiento del mercado.

A continuación, se calculará el hiper-plano de  $N$  dimensiones que mejor separe los puntos para diferentes valores de  $C$ . Para cada uno de estos hiper-planos, se calculará la tasa de error media procedente de la *cross-validation*. El resumen de los resultados se puede ver en la Figura 32.

En la Figura 32 se observa el papel del parámetro  $C$  respecto al error de clasificación. Se puede ver como para valores de  $C$  muy bajos el error es mayor, a medida que aumentamos  $C$  el error va disminuyendo. A pesar de eso, el hecho de aumentar  $C$  también hace aumentar el margen, es decir aumenta la tolerancia del modelo. Por ello, se ve como el valor de  $C$  que minimiza el error de clasificación en este caso, se encuentra cerca de  $C = 1$ . Se recuerda que estos valores han sido obtenidos gracias a un proceso de validación cruzada.

En el Cuadro 26 se puede ver la matriz de confusión para este primer modelo creado añadiendo variables sentimentales.

En el Cuadro 27 se puede ver la matriz de confusión para un modelo de separación lineal con variable sentimentales pero ahora con los datos de *validation*.

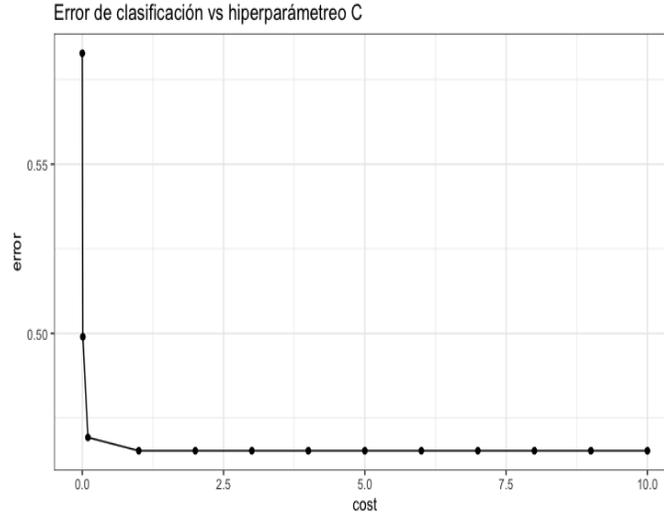


Figura 32: Efecto del parámetro C sobre las SVM en Kernel Lineal con variables sentimentales

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	210	135
= 1	210	221

Cuadro 26: Tabla de confusión Kernel Lineal Learning con variables sentimentales

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	56	40
= 1	42	56

Cuadro 27: Tabla de confusión Kernel Lineal Validation con variables sentimentales

Datos	Accuracy	Error Rate	Precision	Recall	F-measure
Learning	55.54 %	44.46 %	50.00 %	60.87 %	54.90 %
Validation	57.73 %	42.27 %	57.14 %	58.33 %	57.73 %

Cuadro 28: Medidas de bondad Kernel Lineal con variables sentimentales

Las conclusiones de este primer modelo en *learning* y *validation* se pueden ver en la siguiente Tabla.

En el Cuadro 28 se pueden ver las diferentes medidas de bondad, explicadas anteriormente para un modelo SVM de separación lineal con variables sentimentales. Se puede ver como la *accuracy* del modelo de *validation* es mayor que en *learning*, por lo que el modelo acierta más en *validation*. En los otros estadísticos como *error-rate*, *precision* y *F-measure* también funciona mejor el modelo de *validation*. El mismo análisis se realizará pero ahora con un separador de Kernel Radial.

El separador de Kernel Radial como se ha comentado anteriormente, permite hacer separaciones no lineales de la variable respuesta. Dicho separador depende de dos parámetros, por un lado, el parámetro *C* y por el otro, el parámetro *Gamma* que mide la complejidad del modelo. En la Figura 33 se puede ver el comportamiento de estos dos parámetros respecto al error de clasificación para nuestro modelo de Kernel Radial. Se observa que el error mínimo

<sup>32</sup>Por eso se hace la corrección de *class.weights* dentro de la función constructora de SVM, se quiere dar la misma importancia al hecho de acertar los 1's que los 0's en este modelo.

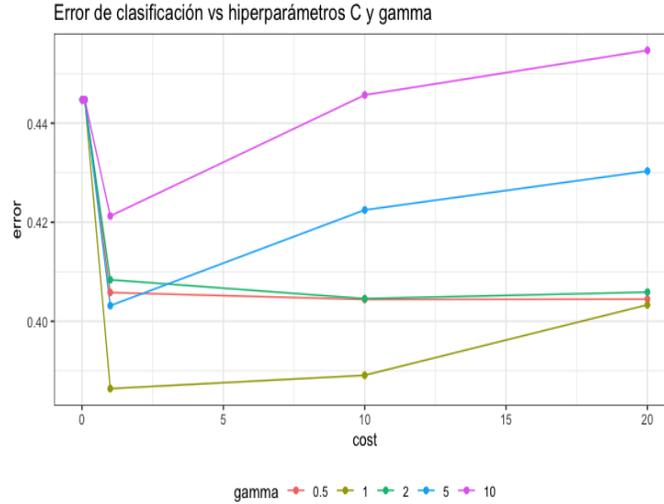


Figura 33: Efecto del parámetro C y Gamma sobre las SVM en Kernel Radial con variables sentimentales

se consigue con un valor de  $C = 1$  y de  $\Gamma = 1$ . Estas referencias serán las que se utilizarán para los modelos de Kernel Radial de esta subsección.

En el Cuadro 29 se puede ver la matriz de confusión para este modelo creado con separador Radial.

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	228	117
= 1	109	322

Cuadro 29: Tabla de confusión Kernel Radial Learning con variables sentimentales

Valor Real / Valor Estimado	$\hat{binary} = 0$	$\hat{binary} = 1$
= 0	44	25
= 1	52	73

Cuadro 30: Tabla de confusión Kernel Radial Validation con variables sentimentales

En el Cuadro 30 se puede ver la matriz de confusión para un modelo de separación radial con variables sentimentales pero ahora con los datos de *validation*.

Las conclusiones de este segundo modelo en *learning* y *validation* se pueden ver en el siguiente cuadro.

Datos	Accuracy	Error Rate	Precision	Recall	F-measure
Learning	70.88 %	29.12 %	67.66 %	66.09 %	66.86 %
Validation	60.31 %	39.69 %	63.77 %	45.83 %	53.33 %

Cuadro 31: Medidas de bondad Kernel Radial con variables de sentimentales

En el Cuadro 31 se pueden ver las diferentes medidas de bondad, explicadas anteriormente para un modelo SVM de separación lineal sin variables sentimentales. Se puede ver como la *accuracy* del modelo de *validation* es inferior que en *learning*, por lo que el modelo acierta más en *learning* ya que esto es lo que mide el modelo. En los otros estadísticos como *Error-rate*, *precision*, *recall* y *F-measure* también funciona mejor el modelo de *learning*.

Lo que interesa ahora, es comparar los resultados obtenidos con un separador Lineal y un separador Radial en los modelos con indicadores sentimentales, para escoger el modelo óptimo con variables procedentes de los tuits de Donald Trump. Para ello, compararemos los resultados obtenidos bajo la muestra de validación en los dos casos ya que esta muestra no ha formado parte de la construcción del modelo. En este caso se puede ver los valores de *accuracy* son superior en el modelo Radial (57.73% vs 60.31%), por construcción si la *accuracy* del modelo Lineal es inferior, mayor será su *error-rate* respecto al modelo Radial. La precisión (57.14% vs 63.77%) también es favorable al modelo Radial pero no así el *recall* (58.33% vs 45.83%) y el *f-measure* (57.73% vs 53.33%) donde los valores del modelo Lineal son superiores.

Como resultado de la metodología para selección del modelo óptimo entre modelos con *Kernel* distinto, descrita en el apartado anterior, se ha decidido seleccionar el modelo de Kernel Radial con parámetros de coste  $C = 1$  y  $\text{Gamma} = 1$  como modelo que mejor funciona con las variables financieras retrasadas hasta tres días y la variable Ratio de Negatividad retrasado tres días para intentar predecir si el mercado bursátil bajará o subirá.

#### 8.4. Comparación de los modelos con y sin indicadores sentimentales

Por último, una vez se han escogido los dos modelos óptimos para las dos condiciones experimentales, en esta caso un modelo de Kernel Radial con parámetros  $C = 10$  y  $\text{Gamma} = 2$  para el caso con sólo variables financieras y un modelo de Kernel Radial con parámetros  $C = 1$  y  $\text{Gamma} = 1$  para la situación donde se incorpora la información procedente del análisis de Granger.

A continuación, se verá cual de los dos modelos funciona mejor bajo los datos de *Test*, de esta manera se podrá contrastar la información concluida del apartado de Granger que nos indicaba que alguna variables sentimentales, en este caso, el tercer retardo del ratio de negatividad, contienen información predictiva sobre los rendimientos del Dow Jones o si por lo contrario, la información extraída de los tuits de Donald Trump con el análisis sentimental realizado no ayuda a predecir si el Dow Jones subirá o bajará para un día concreto.

Se recuerda que esta muestra de *test* contiene un 30% de los datos, 416 observaciones que se corresponden a los últimos 416 días en muestra, des de el 9 de Septiembre de 2019 al 6 de Noviembre de 2020. Debido a que esta es una muestra independiente de la que se ha utilizado para estimar el modelo, se considera que se tendrá una buena estimación del error de generalización del método seleccionado. La generalización es la capacidad de obtener buenos resultados con datos nuevos. Para medir la generalización, medimos el error del modelo en el conjunto de testeo. Es decir, usamos los datos de entrenamiento para que el modelo aprenda y luego le pedimos al modelo que nos de los resultados correspondientes a los datos de testeo.

Para el modelo sin indicadores sentimentales bajo los datos de test se obtiene la siguiente matriz de confusión.

Valor Real / Valor Estimado	$\hat{\text{binary}} = 0$	$\hat{\text{binary}} = 1$
= 0	88	87
= 1	80	161

Cuadro 32: Tabla de confusión Kernel Radial Test sin variables sentimentales

Para el modelo con indicadores sentimentales bajo los datos de test se obtiene la siguiente matriz de confusión.

Valor Real / Valor Estimado	$\hat{\text{binary}} = 0$	$\hat{\text{binary}} = 1$
= 0	76	99
= 1	69	172

Cuadro 33: Tabla de confusión Kernel Radial Test con variables sentimentales

A partir de estas tablas ya se puede observar como el modelo con indicadores sentimentales acierta más cuando el mercado va a subir pero también vemos como falla más cuando afirma que el mercado subirá pero finalmente este no lo hace. El modelo sin variables sentimentales acierta más cuando el mercado tendrá un movimiento bajista que el modelo con variables sentimentales. Parece ser que el modelo con variables sentimentales tiende a sobre-estimar los días alcistas en el mercado.

Datos	Accuracy	Error Rate	Precision	Recall	F-measure
Sin variables sentimentales	59.86 %	40.14 %	52.38 %	50.28 %	51.31 %
Con variables sentimentales	59.62 %	40.38 %	52.41 %	43.43 %	47.50 %

Cuadro 34: Medidas de bondad de Test sin y con variables de sentimentales

Finalmente, en el Cuadro 34 se ve la comparación de los resultados en ambas situaciones experimentales.

Los resultados procedentes de el Cuadro 34 indican un valor de *accuracy* muy parecido para ambos modelos, 59.86 % de acierto para el modelo sin variables sentimentales en *test* y un 59.62 % de acierto para el modelo con variables sentimentales en *test*, por lo que el *error-rate*, en este caso error de generalización, es superior en el modelo con indicadores sentimentales (40.38 %) que en el model sin indicadores sentimentales (40.14 %) Los valores para los estadísticos de *precision*, *recall* y *f-measure* también son muy parecidos entre los dos modelos, pero ligeramente superiores para el modelo sin variables sentimentales.

Se puede concluir que este modelo de Support Vector Machine es neutral a la incorporación de esta variable sentimental, en el sentido que no hay indicios para decir que mejore o empeore la *accuracy* del modelo. Ambos modelos con o sin variable sentimentales tienen porcentajes de error muy parecidos. Observando al detalle las matrices de confusión, parece ser que el modelo sentimental tiende a sobre-estimar las subidas del mercado, haciendo que acierte más cuando el mercado realmente sube pero también fallando más cuando el mercado realmente baja. Además, el modelo escogido como modelo óptimo incorporando información sentimental no haya sido un modelo Lineal, sino Radial, se puede relacionar con la dualidad de efectos que se ha visto que causan las variables sentimentales, siendo su efecto algunas veces positivo sobre los rendimientos cuando Trump está de buen “humor” y otras veces hemos visto que cuando Trump está de mal humor, esto también tiene consecuencias positivas sobre los rendimientos.

Estos resultados señalan que la incorporación de la variable, significativa en el análisis de Granger, que mide el ratio de negatividad para cada día respecto a los tuits de Donald Trump no mejora la capacidad predictiva de un modelo SVM con los retardos de la información financiera para predecir el comportamiento del mercado para un día cualquiera  $t$ . Estos resultados no afirman que los tuits de Donald Trump no tienen ninguna importancia en los rendimientos del Dow Jones, simplemente rechazan la hipótesis que el ratio de negatividad retardado 3 días mejoré la capacidad de predecir si el Dow Jones subirá o bajará. En todo este trabajo, se han encontrado diferentes motivos para creer que las conclusiones derivadas de un análisis sentimental de los tuits del presidente Trump, mediante la creación de indicadores sentimentales puede llevar a establecer una ligera relación entre dichos indicadores y los rendimientos del Dow Jones, por lo que un estudio más exhaustivo de los efectos de la incorporación de otras variables sentimentales con otros retardos podría hacer mejorar significativamente los valores de predicción respecto a los modelos de SVM sin variables sentimentales.

## 9. Conclusiones

Los medios sociales han sido utilizados cada vez más en los informes financieros de rendimientos, este hecho genera oportunidades y desafíos para el sector de la inversión. El *microblogging* se está erigiendo como una fuente vital de datos para su análisis, disponible para cualquiera que quiera aprovecharla. Se ha discutido recientemente sobre los tuits del presidente Trump y su foco de influencia en el mundo financiero, y sobre como estos tuits pueden hacer mover el mercado bursátil americano y otros mercados internacionales. El objetivo de este trabajo es ver de una forma rigurosa y completa los efectos de estos tuits en los mercados bursátiles americanos, trabajando con el índice del Dow Jones, el S&P 500 y su índice de volatilidad, el VIX.

Se recuerda que aunque ahora se presentarán las diferentes conclusiones de los diferentes apartados, los resultados de cada apartado son vitales para la realización de la porción posterior. No son las conclusiones individuales de cada apartado en lo que se basa la riqueza de este trabajo, sino en la relación e implicación de todas las secciones para poder llegar a un desenlace final que se discutirá más adelante.

La aplicación del análisis textual previo sobre los tuits del presidente Trump, ha permitido conocer la forma única que tiene el presidente Trump de comunicarse en sus redes sociales. La representación gráfica en nubes de palabras nos ha facilitado encontrar las palabras o conjunto de ellas más utilizadas en sus tuits. Como resultado se ha visto que había un conjunto de términos que en otros campos tienen una connotación neutral y que en sus tuits concretos les aplicaba un significado negativo. Esto ha permitido modificar el diccionario léxico añadiendo estos vocablos y su connotación en términos de Trump. Algunas de estas palabras han sido: muro, caza de brujas o México.

A partir del diccionario léxico modificado con las conclusiones del análisis textual, se ha realizado un análisis sentimental para asignar un sentimiento entre las categorías: “positivo”, “neutro” y “negativo” para cada tuit y de la misma forma clasificar cada día en las mismas categorías según la valoración de sus tuits. Las conclusiones de este análisis sentimental han mostrado cómo, al contrario de lo que inicialmente se esperaba, los tuits de Trump son mayoritariamente positivos (56 % vs un 39 % de tuits negativos y un 5 % de tuits neutros). Este resultado está muy influido por la modificación explicada del diccionario léxico que ha permitido convertir muchos tuits que antes se consideraban como neutros a otra categoría. Además, se ha podido ver cuál ha sido el estado de ánimo de Donald Trump durante todo su mandato. Se ha visto como los indicadores de opinión o sentimentales mostraban como el presidente Trump inició su mandato de forma optimista con un índice de polaridad bastante positivo, con un repunte de positividad en septiembre de su primer año como presidente. Su segundo año como presidente se inicia con una caída en el estado de ánimo, ya que estos indicadores de polaridad bajan considerablemente, aunque hay otro repunte de tuits positivos en octubre de ese 2018. 2019 es el año más pesimista de Donald Trump, con unos indicadores bastante bajos y con los únicos dos picos mensuales donde el número de tuits negativos supera al número de tuits positivos para enero y marzo de 2019, seguramente a causa de las tensiones comerciales por su política exterior con China y la discusión sobre la construcción del muro frontera con México. El año 2020 es el año más volátil, empieza con récord absoluto de sentimiento positivo en los índices de polaridad, pero estos caen abruptamente a partir del estallido de la pandemia del COVID-19. Finalmente, estos índices consiguen recuperar-se una vez entrada en la segunda mitad del año 2020. De esta forma se ha podido establecer el “humor” de Donald Trump para su presidencia, en base a sus tuits.

Se hace énfasis en la metodología utilizada para este estudio, porque se considera que el enfoque realizado para discernir el efecto de los tuits de Donald Trump en los mercados bursátiles americanos ya puede ser considerado una conclusión en sí. En razón de que en este trabajo se puede ver un técnica de indagación apropiada para los efectos de las redes sociales en los mercados financieros.

De los datos financieros por si solos, se ha concluido que la presidencia Trump no sigue la teoría de ciclos electorales que establece que normalmente los dos primeros años electorales son los más flojos, y en la última mitad del ciclo electoral, a causa de la proximidad de las elecciones, tenemos tiempos más prósperos para la economía. En la presidencia Trump, se observa como el primer año el mercado sube un 32 %, baja un 5 % el segundo año, para rebotar con una subida del 19 % en 2019 y finalizar con una bajada del 2.5 % hasta octubre de 2020, sobretodo a causa del impacto del COVID-19 en la economía.

Esta información procedente de los tuits de Donald Trump se ha combinado con los rendimientos diarios de los valores del Dow Jones, S&P 500 y VIX. Se ha realizado un análisis de causalidad para observar si los valores de los indicadores de opinión contienen información predictiva de los rendimientos de los mercados.

El único efecto persistente para toda la presidencia ha sido el efecto del Ratio de Negatividad que mide la proporción de tuits negativos respecto al total de tuits con opinión. Se ha encontrado que el valor retardado 3 días de esta variable es estadísticamente significativo como predictor de una variable binaria que mide si el Dow Jones sube o baja cada

día respecto al anterior. Esta relación se ha estudiado al detalle en un modelo logit, ya que la variable respuesta es binaria. Aumentar el Ratio de Negatividad 1 punto, es decir pasar de un día donde no hayan tuits negativos a que todos los tuits sean negativos aumenta la probabilidad de que el rendimiento del tercer día siguiente sea positivo en un 73 %.

Los retardos de hasta 9 días del Ratio de Negatividad tienen información predictiva sobre los valores de los rendimientos del Dow Jones para el año 2017 en concreto. Esta relación es positiva y estadísticamente significativa solo para el retardo del día 8, por lo tanto, para el año 2017 cuando aumenta el Ratio de Negatividad, también aumentan los rendimientos bursátiles al cabo de 8 días. Más concretamente, aumentar el ratio de negatividad en 1 punto, aumenta los rendimientos esperados en un 1.74 %.

El modelo con retardos de hasta 10 días del Ratio de Positividad, del Ratio de Negatividad y del Índice de Polaridad 1 tienen información predictiva sobre la variable binaria del S&P 500 para el año 2018. Se ha visto como aumentar los índices de Polaridad 1 o el Ratio de Positividad hace más probable que al cabo de 6 y 10 días el rendimiento bursátil del S&P 500 sea negativo. Al contrario pasa con el 5 día, donde vemos que la relación es positiva para las variables del Ratio de Positividad y del Índice de Polaridad 1. Un aumento en la proporción de tuits positivos hace aumentar la probabilidad de tener un rendimiento positivo al cabo de 5 días. Al revés pasa para el Ratio de Negatividad, por construcción, aumentar el número relativo de tuits negativos hace disminuir la probabilidad de obtener rendimientos positivos al cabo de 5 días y hace aumentar la posibilidad de obtener rendimientos positivos al cabo de 6 y 10 días. Por lo tanto, se puede ver como incluso en periodos temporales cortos existe esta dualidad de relación entre las variables sentimentales y financieras.

Hasta el momento se había comprobado si la información procedente de los tuits de Donald Trump y que refleja su estado de “humor” era capaz predecir la volatilidad del mercado. Es decir, ¿Cuándo Trump esta de mal “humor” aumenta la volatilidad del mercado? Se ha visto que no habían indicios estadísticos para pensar en ninguna relación entre las variables sentimentales y el índice de volatilidad del S&P 500, el VIX. Después, para completar el análisis se ha querido dar respuesta a la siguiente pregunta: ¿Cuándo el “humor” de Trump es más volátil según sus tuits, es el mercado también más volátil? Mediante el calculo de la desviación típica diaria de la puntuación sentimental de los tuits del presidente y los valores del VIX se ha visto que tampoco hay ninguna relación entre los cambio de “humor” del presidente y la volatilidad del mercado medida por el VIX.

Como conclusión general los tuits de Trump si que afectan al mercado ya que existe un efecto persistente en toda la presidencia Trump. La relación entre el número relativo de tuits negativos y la variable binaria del DJIA. Gracias al análisis de causalidad se ha visto que los valores retardados hasta tres días de esta variable sentimental tienen información predictiva de la variable que mide si los rendimientos son superiores o inferiores a 0. Cuando se ha estudiado esta relación utilizando el modelo adecuado Logit se ha visto que la pieza clave de esta relación era el retardo del tercer día. De esta manera, se ha investigado y declarado que aumentar este Ratio de Negatividad hace que el rendimiento del tercer día siguiente tenga más posibilidades de ser positivo. Esta conclusión es la idea de partida para la modelización con SVM donde se ha comparado un modelo de SVM con solo variables financieras con un modelo SVM donde se le ha añadido información procedente de las variables sentimentales. También se destaca la dualidad de las consecuencias estadísticamente palpables procedentes de las variables sentimentales. En este sentido, cuando el “humor” de Trump es bueno, reflejado en la positividad de sus tuits, este precede a movimientos diarios alcistas del mercado a corto plazo y negativo a largo plazo.

Respecto a la implementación de modelos de Support Vector Machines para predecir la tendencia de un rendimiento, se ha visto como para el modelo sin variables sentimentales, la mejor manera de predecir ha sido mediante un Kernel Radial, seguido muy de cerca por un Kernel Lineal. Al añadir la información procedente de las variables sentimentales, el mejor modelo para predecir las tendencias diarias ha sido claramente el modelo Radial, lo que corrobora lo dicho anteriormente. La relación entre el “humor” de Trump y la economía no es lineal. Finalmente se ha visto como este modelo de Support Vector Machines es neutral a la incorporación de esta variable sentimental, es decir, la precisión del modelo no varía. Ambos modelos con o sin variables sentimentales tienen porcentajes de error muy parecidos.

Estos resultados señalan que la incorporación de la variable, significativa en el análisis de Granger, que mide el Ratio de Negatividad para cada día, no mejora la capacidad predictiva de un modelo SVM con los retardos de la información financiera para predecir el comportamiento del mercado para un día cualquiera  $t$ . Estos resultados no afirman que los tuits de Donald Trump no tengan ninguna importancia en los rendimientos del Dow Jones, simplemente rechazan la hipótesis que el Ratio de Negatividad retardado 3 días mejoré la capacidad de predecir si el Dow Jones subirá o bajará.

La investigación sobre los efectos concretos de los tuits de carácter financiero de Donald Trump, mediante un estudio de eventos ha permitido examinar las consecuencias concretas de dos tuits de Donald Trump. De los 14.074 tuits disponibles sólo 117 de estos se han catalogado como tuits financieros. De estos 117 tuits sólo 4 han mostrado propiedades estadísticas para que resultara interesante estudiar sus repercusiones bursátiles.

Por un lado, se ha estudiado como un tuit de Donald Trump sobre la financiación del famoso “Muro” que actúa como frontera con México ha sido capaz de realizar movimientos anormales dentro del mercado. En este caso se ha visto que los días anteriores a este tuit el mercado americano del Dow Jones llevaba una caída progresiva de 10 días y des de este tuit el mercado americano se recupero progresivamente hasta volver a los niveles normales en ese periodo. Obviamente, no se puede decir que fuera el tuit de Donald Trump el que salvará los mercados pero si que se ha explicado la intuición económica de como este tuit podría afectar a los mercados americanos y se ha encontrado evidencia significativa para indicar que existe un correlación entre este tuit y los resultados del Dow Jones.

Por otro lado, el estudio de eventos sobre un tuit de Donald Trump de 2020, en relación a las políticas comerciales americanas con los demás países, modificó la tendencia bursátil del Dow Jones. Antes de este tuit el mercado estaba en subida y después de este tuit bajo progresivamente. Más en concreto, se encontró que el primer día bursátil después de este tuit tenía un rendimiento positivo anormal para ese momento del mercado. Sin embargo, la tendencia general después de este tuit fue una caída, significativa y importante, que no obedece al comportamiento normal del mercado. Se vuelve a ver en este caso, la dualidad de consecuencias que tienen los tuits de Trump, dependiendo de la temporalidad escogida.

En líneas generales se puede ver que los tuits de Donald Trump, ciertamente, tienen efecto sobre la economía americana. Esta afirmación se basa en que se ha encontrado una relación significativa entre su porcentaje de tuits negativos y la tendencia bursátil de los tres siguientes días. A partir de las conclusiones de este proyecto, se puede hablar de efecto de los tuits de Trump en el índice bursátil del Dow Jones específicamente. No se han encontrado relaciones significativas para el índice S&P 500, ni tampoco se cree que los tuits de Trump afecten a la volatilidad del mercado. El efecto de los tuits de Trump tiende a ser diferente en el corto y en el largo plazo. Se ha mirado el efecto de tuits concretos sobre el mercado del DJIA y se ha visto que algunos de sus tuits tenían un efecto palpable sobre los rendimientos de este índice. Por lo tanto, se puede decir que el “humor” del presidente, reflejado en sus tuits, influye en el mercado durante toda su presidencia, pero también que tuits específicos sobre temas económico-financieros causan comportamientos no normales en el mercado.

A diferencia de trabajos parecidos, se han utilizado diferentes mercados financieros y no se ha centrado únicamente en las repercusiones de un mercado. Además, se ha añadido posibles repercusiones sobre la volatilidad del mercado, a parte de sobre los rendimientos del mercado. Asimismo, se ha contemplado un horizonte temporal muy extenso, todos los trabajos con los que se pueden encontrar similitudes trabajan con series temporales de máximo 15 meses. En este proyecto en cambio, se quieren obtener resultados significativos para una presidencia en completo, prácticamente 4 años. Existen muchos estudios que intentan predecir los rendimientos bursátiles a través del estado de ánimo general, reflejado en un volumen inmenso de tuits y su clasificación sentimental, mientras que, en este proyecto se ha intentado obtener resultados parecidos a estos trabajos pero únicamente con un solo usuario de Twitter. Por último, esta investigación se basa en los efectos de los 14.074 tuits de Donald Trump, pero también se ha puesto el foco en analizar que pasa con los tuits de carácter financiero de Trump, por lo que se han visto los efectos del estado de ánimo de Donald Trump sobre la economía y también los efectos de las palabras que dice Trump respecto al sector financiero, sobre los propios mercados.

Las limitaciones de este trabajo se basan en la propia construcción del proyecto, como se ha comentado con anterioridad, este proyecto se basa en una construcción progresiva del trabajo, en el sentido de que los resultados de cada sección son vitales para el siguiente apartado. Esto crea una limitación en la misión y es que las elecciones recogidas en cada sección tienen influencia en los resultados finales. Por ejemplo, se podría ver cuales serían los resultados si se hubiera escogido otro diccionario léxico. Otra limitación, es que el trabajo se basa en que los indicadores sentimentales recogen o reflejan en buena proporción el estado de ánimo del presidente a partir de sus tuits, los resultados de este proyecto podrían variar si se hubieran escogido otros indicadores. También comentar que se escogieron 4 indicadores diferentes para paliar este problema. Por lo que hace a la mejora de predicción incorporando variables sentimentales, los resultados han sido neutros al respecto, pero se cree que con otros indicadores o utilizando otro conjunto de valores retardados los resultados podrían ser diferentes y en última instancia, se podría concluir que la incorporación de variables sentimentales mejora la clasificación de futuros días bursátiles entre negativos o positivos. Además se cree que los efectos de los tuits de Donald Trump en casos o tuits concretos son más fáciles de controlar en periodos de minutos o horas después del tuit. Con la información financiera por hora se cree que se hubieran podido localizar más efectos de los tuits del presidente, que con los valores diarios con los que se ha trabajado.

Se podrían implementar futuras líneas de investigación para medir el poder que tienen políticos o responsables de grandes corporaciones en los mercados financieros. Con el incremento de popularidad de Twitter, se ha visto que el poder que tienen estas personas en sus manos es increíble. Este trabajo se ha centrado en los efectos generales del estado de “humor” del presidente Trump, pero otra línea de trabajo sería medir que efectos tiene para un empresa ser mencionada en un tuit de Donald Trump. Recientemente, Kylie Jenner, modelo estadounidense, hizo un tuit negativo sobre la compañía y red social Snapchat, ese mismo día la compañía perdió más de mil millones de euros en la capitalización del mercado. Esto muestra que con una importante base de seguidores, se puede afectar positivamente o negativamente a muchas organizaciones y sobretodo se puede influir sobre el comportamiento de muchas personas. Como resultado, las empresas e incluso los gobiernos deberían considerar monitorizar y realizar un análisis crítico sobre el papel de las redes sociales en los mercados financieros.

Otra futura línea de exploración sería considerar la implementación del análisis de sentimientos en base a modelos de inteligencia artificial que aprendan de un dataset de entrenamiento. Para que, después de aprender a clasificar palabras en base a unos datos ya clasificados, sea capaz de asignar puntuación sentimental al dataset con el que se trabajará. Esto puede ser muy beneficioso en casos como el lenguaje natural procedente de las redes sociales y en casos como este donde se quiere analizar la forma de hablar de una persona en concreto, ya que al entrenar el modelo con datos procedentes de la misma o una fuente parecida, esta se adaptará de forma más correcta a la forma de hablar propia de este lenguaje específico.

Por último, sería interesante investigar la misma relación: información procedente de los tuits de Trump y datos financieros pero en el sentido opuesto. Buscar si hay una relación de causalidad de las variables financieras hacia los tuits de Donald Trump. Es decir, ¿está el “humor” de los tuits de Trump causado por los buenos o malos ciclos económicos? Esta posibilidad se considera como un posible trabajo posterior relacionado con este estudio.

Después de la realización de este proyecto y para cerrar él mismo; a pesar de que no es un conocimiento específico ni de la carrera de estadística, ni de la carrera de economía, sino que forma parte troncal de las dos, se quiere considerar y poner en valor el papel del pensamiento crítico. En este trabajo se han podido ver los efectos de los pensamientos de un personaje público y sus consecuencias en el mundo real, en el mundo donde vivimos todos. Es alarmante que alguien tenga tal poder en sus manos y somos nosotros los que podemos evitar que así sea.

## Índice de figuras

1.	Valores de cierre para el índice DJIA del 17-01-2017 al 06-11-2020 . . . . .	16
2.	Rendimientos del índice DJIA del 18-01-2017 al 06-11-2020 . . . . .	17
3.	Boxplot e histograma de los rendimientos del índice DJIA del 18-01-2017 al 06-11-2020 . . . . .	18
4.	Fuente de procedencia de los tuits del presidente Trump . . . . .	20
5.	Evolución de los tuits del presidente Trump por mes . . . . .	20
6.	Boxplot del número RT y FAV por tuit del presidente Trump . . . . .	21
7.	Relación entre el número RT y de FAV por tuit del presidente Trump . . . . .	22
8.	Wordcloud de los unigramas de los tuits de Donald Trump . . . . .	24
9.	Wordcloud de los Bigramas de los tuits de Donald Trump . . . . .	25
10.	Distribución de los tuits según polaridad . . . . .	30
11.	Boxplot e histograma de los tuits según polaridad . . . . .	31
12.	Índice Polaridad 1 por mes . . . . .	32
13.	Índice Polaridad 2 por mes . . . . .	32
14.	Ratio de Negatividad y de Positividad por mes . . . . .	33
15.	Serie temporal del índice de polaridad 1 y el índice DJIA . . . . .	35
16.	Serie temporal del índice de polaridad 2 y el índice DJIA . . . . .	35
17.	Serie temporal del ratio de positividad y el índice DJIA . . . . .	36
18.	Serie temporal del ratio de negatividad y el índice DJIA . . . . .	37
19.	Ejemplo del funcionamiento del análisis de Granger . . . . .	39
20.	Efecto del tercer retardo de Negatividad sobre DJIA.Binary en modelo logit . . . . .	45
21.	Timeline del Event Study . . . . .	48
22.	Tuit 1 para el estudio de eventos . . . . .	51
23.	Tuit 2 para el estudio de eventos . . . . .	52
24.	Abnormal Returns DJIA tuit 1 . . . . .	52
25.	Cumulative Abnormal Returns tuit 1 . . . . .	54
26.	Abnormal Returns tuit 2 . . . . .	56
27.	Cumulative Abnormal Returns tuit 2 . . . . .	57
28.	Funcionamiento de las SVM . . . . .	60
29.	Efecto del parámetro C sobre las SVM . . . . .	62
30.	Efecto del parámetro C sobre las SVM en Kernel Lineal . . . . .	63
31.	Efecto del parámetro C y Gamma sobre las SVM en Kernel Radial . . . . .	65
32.	Efecto del parámetro C sobre las SVM en Kernel Lineal con variables sentimentales . . . . .	67
33.	Efecto del parámetro C y Gamma sobre las SVM en Kernel Radial con variables sentimentales . . . . .	68
34.	Valores de cierre para el índice S&P 500 del 17-01-2017 al 06-11-2020 . . . . .	82
35.	Rendimientos del índice S&P 500 del 18-01-2017 al 06-11-2020 . . . . .	82
36.	Valores de cierre para el índice VIX del 17-01-2017 al 06-11-2020 . . . . .	83
37.	Rendimientos del índice VIX del 18-01-2017 al 06-11-2020 . . . . .	83
38.	SVM para dos variables predictoras y Kernel Lineal . . . . .	93
39.	SVM para dos variables predictoras y Kernel Radial . . . . .	94

## Índice de cuadros

1.	Revisión de la literatura del efecto de las redes sociales sobre los mercados financieros . . . . .	13
2.	Estadísticos descriptivos univariantes de las diferentes variables del DJIA . . . . .	15
3.	Estadísticos descriptivos univariantes para la variable rendimientos . . . . .	18
4.	Estadísticos descriptivos univariantes para las variables $rt$ y $fav$ . . . . .	21
5.	Ejemplos de algunas palabras del diccionario léxico añadidas . . . . .	26
6.	Ejemplos de algunas palabras del diccionario léxico utilizado . . . . .	27
7.	Ejemplos de algunos enfatizadores del diccionario léxico utilizado . . . . .	27
8.	Muestra de tuits clasificados por diccionario léxico . . . . .	29
9.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para las variables de opinión y el DJIA . . . . .	40
10.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para las variables de opinión y el DJIA .binary . . . . .	41
11.	Coefficiente Exponenciados del modelo logarítmico . . . . .	45
12.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para la desviación típica y la media del sentimiento promedio de cada tuit y el VIX . . . . .	46
13.	Abnormal Returns para estudio de evento 1: . . . . .	53
14.	Cumulative Abnormal Returns para estudio de evento 1: . . . . .	54
15.	Abnormal Returns para estudio de evento 2: . . . . .	56
16.	Cumulative Abnormal Returns para estudio de evento 2: . . . . .	57
17.	Tabla de confusión . . . . .	60
18.	Distribución de los datos: SVM . . . . .	61
19.	Distribución de los datos con <i>learning</i> : SVM . . . . .	61
20.	Tabla de confusión Kernel Lineal Learning sin variables sentimentales . . . . .	64
21.	Tabla de confusión Kernel Lineal Validation sin variables sentimentales . . . . .	64
22.	Medidas de bondad Kernel Lineal sin variables sentimentales . . . . .	64
23.	Tabla de confusión Kernel Radial Learning sin variables sentimentales . . . . .	65
24.	Tabla de confusión Kernel Radial Validation sin variables sentimentales . . . . .	65
25.	Medidas de bondad Kernel Radial sin variables sentimentales . . . . .	65
26.	Tabla de confusión Kernel Lineal Learning con variables sentimentales . . . . .	67
27.	Tabla de confusión Kernel Lineal Validation con variables sentimentales . . . . .	67
28.	Medidas de bondad Kernel Lineal con variables sentimentales . . . . .	67
29.	Tabla de confusión Kernel Radial Learning con variables sentimentales . . . . .	68
30.	Tabla de confusión Kernel Radial Validation con variables sentimentales . . . . .	68
31.	Medidas de bondad Kernel Radial con variables de sentimentales . . . . .	68
32.	Tabla de confusión Kernel Radial Test sin variables sentimentales . . . . .	69
33.	Tabla de confusión Kernel Radial Test con variables sentimentales . . . . .	69
34.	Medidas de bondad de Test sin y con variables de sentimentales . . . . .	70
35.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para las variables de opinión y el S&P 500 . . . . .	84
36.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para las variables de opinión y el S&P 500.binary . . . . .	84
37.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para las variables de opinión y el VIX . . . . .	84
38.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para las variables de opinión y los rendimientos del DJIA para el año 2017 . . . . .	85
39.	Significación estadística (P-valores) de la correlación bivariante de causalidad de Granger para las variables de opinión y el S&P 500.binary para el año 2018 . . . . .	87

## Referencias

- [1] Johan Bollen, Huina Mao y Xiao-Jun Zeng. “Twitter Mood Predicts the Stock Market”. En: *Journal of Computational Science* 2 (oct. de 2010). DOI: 10.1016/j.jocs.2010.12.007.
- [2] Xue Zhang, Hauke Fuehres y Peter A. Gloor. “Predicting Asset Value through Twitter Buzz”. En: *Advances in Collective Intelligence 2011*. Ed. por Jörn Altmann, Ulrike Baumöl y Bernd J. Krämer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, págs. 23-34. ISBN: 978-3-642-25321-8.
- [3] Beatriz Navarro. *Cómo los tuits de Trump influyen en la bolsa*. Sep. de 2019. URL: <https://www.lavanguardia.com/economia/20190910/47273448654/trump-tuits-china-jpmorgan-bolsas.html>.
- [4] Constantin Colonescu. “The Effects of Donald Trump’s Tweets on US Financial and Foreign Exchange Markets”. En: *Athens Journal of Business & Economics* 4.10 (2018), págs. 375-388. DOI: <http://dx.doi.org/10.30958/ajbe.4-4-2>.
- [5] William B. Cavnar y John M. Trenkle. “N-Gram-Based Text Categorization”. En: *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. 1994, págs. 161-175.
- [6] Ilya Sutskever y col. “On the importance of initialization and momentum in deep learning”. En: *Proceedings of the 30th International Conference on Machine Learning*. Ed. por Sanjoy Dasgupta y David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun de 2013, págs. 1139-1147. URL: <http://proceedings.mlr.press/v28/sutskever13.html>.
- [7] Tomás Mikolov y col. “Distributed Representations of Words and Phrases and their Compositionality”. En: *CoRR* abs/1310.4546 (2013). arXiv: 1310.4546. URL: <http://arxiv.org/abs/1310.4546>.
- [8] Xiang Zhang, Junbo Zhao y Yann LeCun. “Character-level Convolutional Networks for Text Classification”. En: *Advances in Neural Information Processing Systems*. Ed. por C. Cortes y col. Vol. 28. Curran Associates, Inc., 2015, págs. 649-657. URL: <https://proceedings.neurips.cc/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf>.
- [9] Luciano Barbosa y Junlan Feng. “Robust Sentiment Detection on Twitter from Biased and Noisy Data”. En: vol. 2. Ene. de 2010, págs. 36-44.
- [10] Duyu Tang y col. “Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification”. En: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, jun. de 2014, págs. 1555-1565. DOI: 10.3115/v1/P14-1146. URL: <https://www.aclweb.org/anthology/P14-1146>.
- [11] Gabriele Ranco y col. “The Effects of Twitter Sentiment on Stock Price Returns”. En: *PloS one* 10 (sep. de 2015), e0138441. DOI: 10.1371/journal.pone.0138441.
- [12] Panagiotis Papaioannou y col. “Can social microblogging be used to forecast intraday exchange rates?” En: *NETNOMICS: Economic Research and Electronic Networking* 14 (oct. de 2013). DOI: 10.1007/s11066-013-9079-3.
- [13] Gabriel Pui Cheong Fung, Jeffrey Xu Yu y Wai Lam. “News Sensitive Stock Trend Prediction”. En: *Advances in Knowledge Discovery and Data Mining*. Ed. por Ming-Syan Chen, Philip S. Yu y Bing Liu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, págs. 481-493. ISBN: 978-3-540-47887-4.
- [14] M.-A Mittermayer. “Forecasting Intraday Stock Price Trends with Text Mining Techniques”. En: feb. de 2004, 10 pp. ISBN: 0-7695-2056-1. DOI: 10.1109/HICSS.2004.1265201.
- [15] Jae Min y Young-Chan Lee. “Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters”. En: *Expert Systems with Applications* 28 (mayo de 2005), págs. 603-614. DOI: 10.1016/j.eswa.2004.12.008.
- [16] Christian Dunis y col. “Forecasting IBEX-35 moves using support vector machines”. En: *Neural Computing and Applications* 23 (jul. de 2012). DOI: 10.1007/s00521-012-0821-9.
- [17] Huina Mao, Scott Counts y Johan Bollen. *Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data*. 2011. arXiv: 1112.1051 [q-fin.ST].

- [18] Anshul Mittal y Arpit Goel. “Stock prediction using twitter sentiment analysis”. En: *Stanford University, CS229 (2011)* 15 (2012). URL: <http://cs229.stanford.edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis.pdf>.
- [19] Yin Ni y col. “A novel stock evaluation index based on public opinion analysis”. En: *Procedia Computer Science* 147 (2019). 2018 International Conference on Identification, Information and Knowledge in the Internet of Things, págs. 581-587. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2019.01.212>. URL: <http://www.sciencedirect.com/science/article/pii/S1877050919302315>.
- [20] Jasmina Smailović y col. “Predictive sentiment analysis of tweets: A stock market application”. En: *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Springer. 2013, págs. 77-88.
- [21] Yang Yu, W. Duan y Q. Cao. “The impact of social and conventional media on firm equity value: A sentiment analysis approach”. En: *Decis. Support Syst.* 55 (2013), págs. 919-926.
- [22] Timm O Sprenger y col. “Tweets and trades: The information content of stock microblogs”. En: *European Financial Management* 20.5 (2014), págs. 926-957.
- [23] Huina Mao, Scott Counts y Johan Bollen. *Quantifying the effects of online bullishness on international financial markets*. Statistics Paper Series 9. European Central Bank, 2015. URL: <https://EconPapers.repec.org/RePEc:ecb:ecbsps:20159>.
- [24] Bo Pang y Lillian Lee. “Opinion Mining and Sentiment Analysis”. En: *Foundations and Trends in Information Retrieval* 2 (ene. de 2008), págs. 1-135. DOI: 10.1561/1500000011.
- [25] Alexander Pak y Patrick Paroubek. “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”. En: vol. 10. Ene. de 2010.
- [26] Anna Iglesias Campeny. “Text Mining y Riesgo reputacional: un caso aplicado a la Banca Española”. En: (2019).
- [27] Peter Willett. “The Porter stemming algorithm: then and now”. En: *Program* (2006).
- [28] Tyler Rinker y Vitalie Spinu. *trinker/sentimentr: version 0.4.0*. Ver. v0.4.0. Dic. de 2016. DOI: 10.5281/zenodo.222103. URL: <https://doi.org/10.5281/zenodo.222103>.
- [29] Maurizio Naldi. “A review of sentiment computation methods with R packages”. En: *CoRR* abs/1901.08319 (2019). arXiv: 1901.08319. URL: <http://arxiv.org/abs/1901.08319>.
- [30] Peter Turney. “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”. En: *Computing Research Repository - CORR* (dic. de 2002), págs. 417-424. DOI: 10.3115/1073083.1073153.
- [31] Y. Hirsch y J.A. Hirsch. *Stock Trader’s Almanac 2004*. Almanac Investor Series. Wiley, 2003. ISBN: 9780471477549. URL: <https://books.google.es/books?id=viaTuAAACAAJ>.
- [32] *Granger causality*. Ene. de 2021. URL: [https://en.wikipedia.org/wiki/Granger\\_causality](https://en.wikipedia.org/wiki/Granger_causality).
- [33] S. Ranganathan, K. Nakai y C. Schonbach. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. Elsevier Science, 2018. ISBN: 9780128114322. URL: <https://books.google.es/books?id=rs51DwAAQBAJ>.
- [34] Stephen J. Brown y Jerold B. Warner. “Measuring security price performance”. En: *Journal of Financial Economics* 8.3 (1980), págs. 205-258. ISSN: 0304-405X. DOI: [https://doi.org/10.1016/0304-405X\(80\)90002-1](https://doi.org/10.1016/0304-405X(80)90002-1). URL: <http://www.sciencedirect.com/science/article/pii/0304405X80900021>.
- [35] Krishan Rayarel. “The Impact of Donald Trump’s Tweets on Financial Markets”. En: (2018).
- [36] Jeffery A Born, David H Myers y William J Clark. “Trump tweets and the efficient Market Hypothesis”. En: *Algorithmic Finance* 6.3-4 (2017), págs. 103-109.
- [37] Pamela P Peterson. “Event studies: A review of issues and methodology”. En: *Quarterly journal of business and economics* (1989), págs. 36-66.
- [38] Adam Shell. *Government shutdown: History suggests stock market can weather storm*. Dic. de 2018. URL: <https://eu.usatoday.com/story/money/2018/12/21/government-shutdown-impact-dow-jones-industrial-average/2386933002/>.

- [39] Geoffrey Gertz y Homi Kharas. *The New Global Middle Class: A Cross-Over from West to East*. Jul. de 2016. URL: <https://www.brookings.edu/research/the-new-global-middle-class-a-cross-over-from-west-to-east/>.
- [40] Jose Martinez Heras. *support vector machines archivos*. URL: <https://www.iartificial.net/tag/support-vector-machines/>.
- [41] Abhishek Ghose. *An Introduction to Support Vector Machines - DZone AI*. Ago. de 2017. URL: <https://dzone.com/articles/support-vector-machines-tutorial>.

## Apéndice A Código para la extracción de los tuits de Donald Trump

```
get_trumptwitterarchive <- function(years = NULL) {
  ## default to years 2008 through current
  if (is.null(years) || isTRUE(years) || identical(tolower(years), "all")) {
    years <- seq(2008, as.integer(format(Sys.Date(), "%Y")))
  }
  stopifnot(is.numeric(years))
  ## get data for each year
  tta <- lapply(years, trumptwitterarchive_)
  ## make status IDs data frame with tta as attribute
  ids <- lapply(tta, "[[", "id_str")
  ids <- tibble::as_tibble(
    list(status_id = unlist(ids)),
    validate = FALSE
  )
  ## list of data (element = year)
  attr(ids, "data") <- tta
  attr(ids, "years") <- years
  ids
}

trumptwitterarchive_data <- function(data, years = NULL) {
  if (!"data" %in% names(attributes(data))) {
    stop("Archive data not found", call. = FALSE)
  }
  tta <- attr(data, "data")

  if (!is.null(years)) {
    ## if years to subset are provided
    data_years <- attr(data, "years")
    ## if no years attr or if years length differs return w/ warning
    if (is.null(data_years) || length(data_years) != length(years)) {
      warning(
        "Length of years attribute differs from length of data. Returning all extracted data",
        call. = FALSE
      )
    } else {
      tta <- tta[data_years %in% years]
    }
  }
  ## tidy things up a bit
  tta <- tta[lengths(tta) > 0L]
  if (is.list(tta)) {
    tta <- lapply(tta, tibble::as_tibble)
    tta <- do.call("rbind", tta)
    if ("created_at" %in% names(tta)) {
      tta$created_at <- rtweet::format_date(tta$created_at)
    }
  }
  tta
}

trumptwitterarchive_ <- function(year, fromJSON = TRUE) {
```

```

## build and send request
url <- paste0(
  "http://trumptwitterarchive.com/",
  "data/realdonaldtrump/",
  year,
  ".json"
)
op <- getOption("encoding")
on.exit(options(encoding = op))
options(encoding = "UTF-8")
## response object
r <- tryCatch(httr::GET(url), error = function(e) return(NULL))
if (is.null(r)) {
  warning(
    paste("http request for", year, "data failed."),
    call. = FALSE
  )
  return(data.frame())
}
## check html status
httr::warn_for_status(r)
## if fromJSON then convert to list otherwise return response object
if (fromJSON) {
  r <- httr::content(r, "text", encoding = "UTF-8")
  ## if html return empty data frame
  if (grepl("^<!DOCTYPE", r)) {
    r <- data.frame()
  } else {
    r <- jsonlite::fromJSON(r)
  }
}
r
}

status_ids <- get_trumptwitterarchive()

data <- trumptwitterarchive_data(status_ids)

```

## Apéndice B Descripción de datos financieros

### B.1 S&P 500

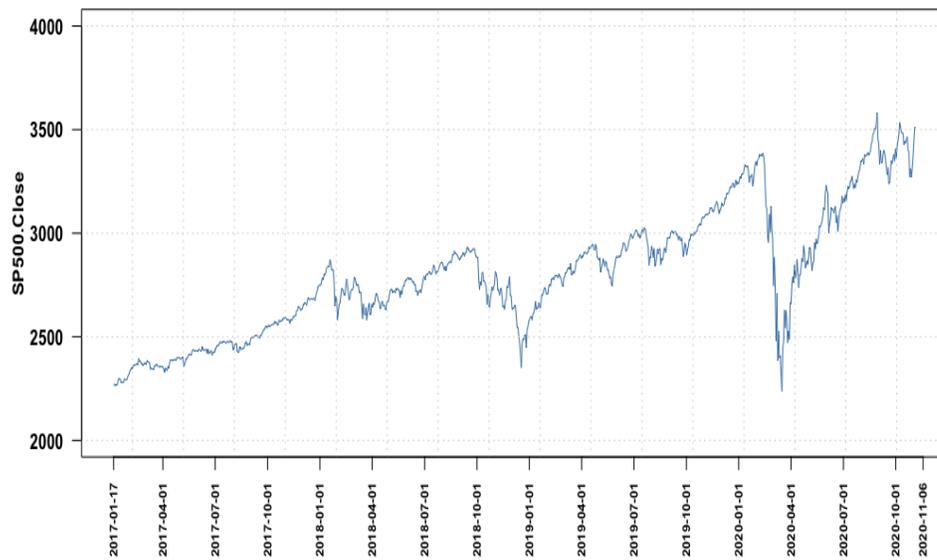


Figura 34: Valores de cierre para el índice S&P 500 del 17-01-2017 al 06-11-2020

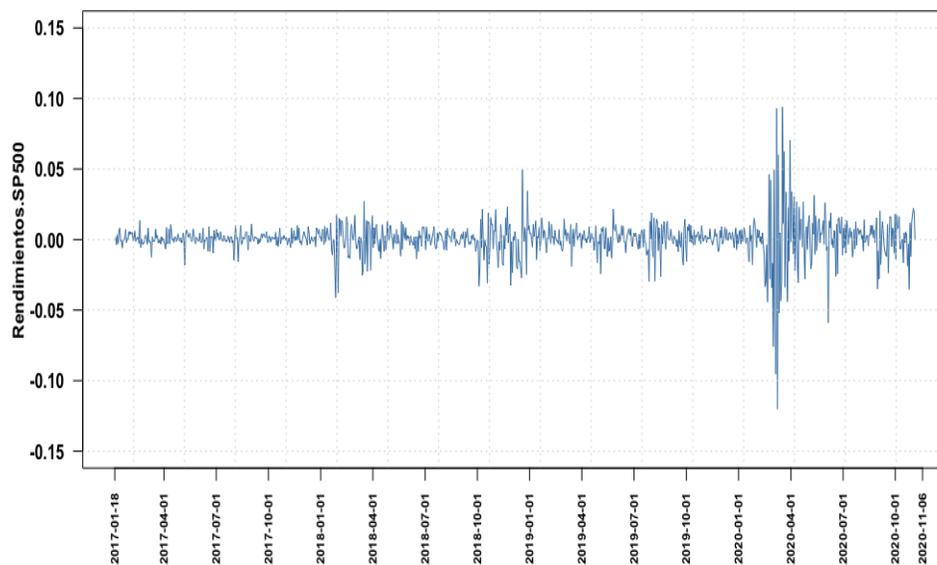


Figura 35: Rendimientos del índice S&P 500 del 18-01-2017 al 06-11-2020

## B.2 VIX

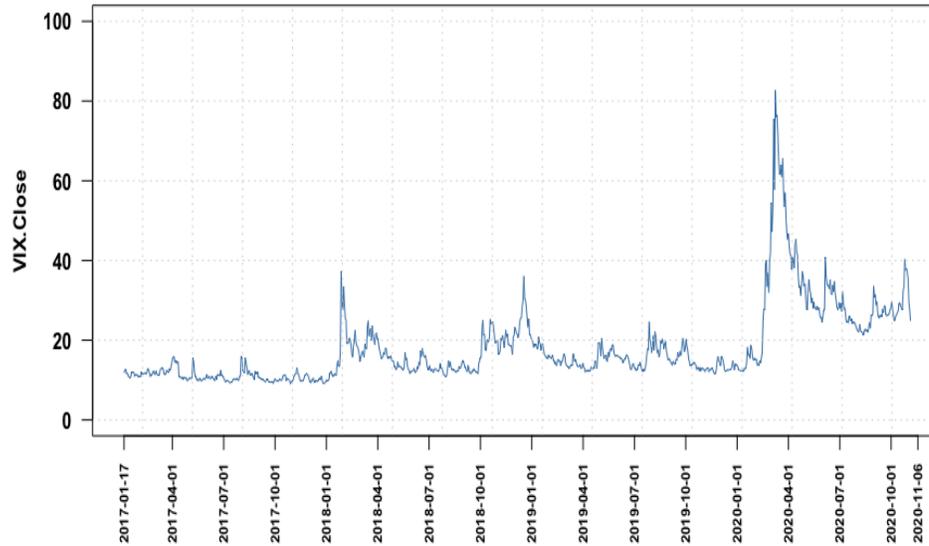


Figura 36: Valores de cierre para el índice VIX del 17-01-2017 al 06-11-2020

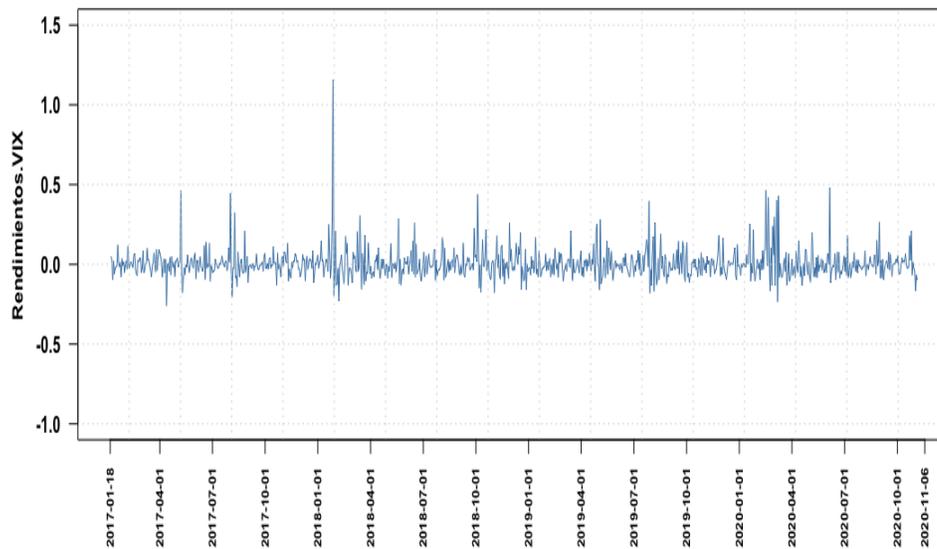


Figura 37: Rendimientos del índice VIX del 18-01-2017 al 06-11-2020

## Apéndice C Causalidad de Granger

### C.1 S&P 500

S&P 500: Lag	Ind. Pol.1	Ind.Pol.2	Ratio Neg.	Ratio Pos.
1 día	0.81	0.42	0.82	0.81
2 días	0.76	0.71	0.72	0.81
3 días	0.86	0.80	0.84	0.89
4 días	0.98	0.92	0.98	0.99
5 días	0.99	0.92	0.98	0.99
6 días	0.88	0.56	0.87	0.89
7 días	0.90	0.58	0.89	0.91
8 días	0.85	0.43	0.84	0.86
9 días	0.95	0.68	0.95	0.95
10 días	0.97	0.71	0.97	0.97

Cuadro 35: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el S&P 500

S&P 500.binary : Lag	Ind. Pol.1	Ind.Pol.2	Ratio Neg.	Ratio Pos.
1 día	0.60	0.79	0.59	0.62
2 días	0.30	0.79	0.27	0.35
3 días	0.18	0.91	0.16	0.21
4 días	0.42	0.95	0.38	0.48
5 días	0.50	0.97	0.45	0.57
6 días	0.15	0.72	0.13	0.17
7 días	0.19	0.80	0.17	0.22
8 días	0.17	0.77	0.13	0.19
9 días	0.28	0.93	0.26	0.32
10 días	0.36	0.97	0.35	0.38

Cuadro 36: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el S&P 500.binary

### C.2 VIX

VIX: Lag	Ind. Pol.1	Ind.Pol.2	Ratio Neg.	Ratio Pos.
1 día	0.89	0.30	0.92	0.86
2 días	0.73	0.56	0.67	0.80
3 días	0.94	0.57	0.91	0.97
4 días	0.97	0.74	0.95	0.97
5 días	0.99	0.86	0.98	0.99
6 días	0.26	0.24	0.24	0.28
7 días	0.34	0.30	0.33	0.37
8 días	0.38	0.30	0.36	0.40
9 días	0.46	0.30	0.46	0.47
10 días	0.47	0.27	0.47	0.48

Cuadro 37: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el VIX

### C.3 Test por años: 2017, 2018, 2019 y 2020

DJIA .2017 : Lag	Ind. Pol.1	Ind.Pol.2	Ratio Neg.	Ratio Pos.
1 día	0.75	0.45	0.71	0.81
2 días	0.87	0.84	0.65	0.99
3 días	0.93	0.90	0.79	0.98
4 días	0.64	0.38	0.54	0.70
5 días	0.44	0.52	0.41	0.46
6 días	0.53	0.62	0.51	0.53
7 días	0.44	0.55	0.41	0.48
8 días	0.14	0.64	0.10	0.21
9 días	0.14	0.78	<b>0.04**</b>	0.23
10 días	0.15	0.87	0.18	0.14

Cuadro 38: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y los rendimientos del DJIA para el año 2017

Modelo Lineal para entender la relación entre el resultado del test de Granger de la variable del Ratio de Negatividad retardada y los rendimientos del DJIA para el año 2017 que se puede observar en el cuadro 38.

Listing 3: R output Modelo Lineal con Ratio de Negatividad para 2017

---

```

Call:
lm(formula = df.DJIA_2017$return ~ lagpad(df.DJIA_2017$return,
  1) + lagpad(df.DJIA_2017$return, 2) + lagpad(df.DJIA_2017$return,
  3) + lagpad(df.DJIA_2017$return, 4) + lagpad(df.DJIA_2017$return,
  5) + lagpad(df.DJIA_2017$return, 6) + lagpad(df.DJIA_2017$return,
  7) + lagpad(df.DJIA_2017$return, 8) + lagpad(df.DJIA_2017$return,
  9) + lagpad(df.DJIA_2017$return, 10) + lagpad(df.DJIA_2017$ratio_neg,
  1) + lagpad(df.DJIA_2017$ratio_neg, 2) + lagpad(df.DJIA_2017$ratio_neg,
  3) + lagpad(df.DJIA_2017$ratio_neg, 4) + lagpad(df.DJIA_2017$ratio_neg,
  5) + lagpad(df.DJIA_2017$ratio_neg, 6) + lagpad(df.DJIA_2017$ratio_neg,
  7) + lagpad(df.DJIA_2017$ratio_neg, 8) + lagpad(df.DJIA_2017$ratio_neg,
  9) + lagpad(df.DJIA_2017$ratio_neg, 10))

Residuals:
    Min       1Q   Median       3Q      Max
-0.0177067 -0.0012803 -0.0001778  0.0012992  0.0132408

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.0011255   0.0007804  -1.442   0.1506
lagpad(df.DJIA_2017$return, 1) -0.0439645   0.0701969  -0.626   0.5317
lagpad(df.DJIA_2017$return, 2)  0.1518393   0.0670125   2.266   0.0244 *
lagpad(df.DJIA_2017$return, 3) -0.0280193   0.0665349  -0.421   0.6741
lagpad(df.DJIA_2017$return, 4) -0.0699795   0.0673334  -1.039   0.2998
lagpad(df.DJIA_2017$return, 5) -0.0059953   0.0675453  -0.089   0.9294
lagpad(df.DJIA_2017$return, 6) -0.0767077   0.0659151  -1.164   0.2458
lagpad(df.DJIA_2017$return, 7) -0.0261722   0.0643401  -0.407   0.6846
lagpad(df.DJIA_2017$return, 8)  0.0085102   0.0639286   0.133   0.8942
lagpad(df.DJIA_2017$return, 9)  0.1318329   0.0630916   2.090   0.0378 *
lagpad(df.DJIA_2017$return, 10) 0.0548947   0.0636306   0.863   0.3892
lagpad(df.DJIA_2017$ratio_neg, 1) 0.0005194   0.0008361   0.621   0.5351
lagpad(df.DJIA_2017$ratio_neg, 2) 0.0005762   0.0008349   0.690   0.4909
lagpad(df.DJIA_2017$ratio_neg, 3) 0.0008558   0.0008400   1.019   0.3094
lagpad(df.DJIA_2017$ratio_neg, 4) 0.0013274   0.0008193   1.620   0.1066
lagpad(df.DJIA_2017$ratio_neg, 5) 0.0007783   0.0008198   0.949   0.3434
lagpad(df.DJIA_2017$ratio_neg, 6) -0.0004757   0.0008245  -0.577   0.5646
lagpad(df.DJIA_2017$ratio_neg, 7) -0.0009622   0.0008326  -1.156   0.2490
lagpad(df.DJIA_2017$ratio_neg, 8) 0.0017412   0.0008353   2.085   0.0382 *
lagpad(df.DJIA_2017$ratio_neg, 9) -0.0001053   0.0008287  -0.127   0.8990
lagpad(df.DJIA_2017$ratio_neg, 10) 0.0002426   0.0008019   0.303   0.7625
---
Signif. codes:  0   ***    0.001   **    0.01   *    0.05   .    0.1   1

```

Residual standard error: 0.00315 on 227 degrees of freedom  
 (100 observations deleted due to missingness)  
 Multiple R-squared: 0.09551, Adjusted R-squared: 0.01582  
 F-statistic: 1.199 on 20 and 227 DF, p-value: 0.2568

---

S&P 500.binary.2018: Lag	Ind. Pol.1	Ind.Pol.2	Ratio Neg.	Ratio Pos.
1 día	0.48	0.36	0.48	0.48
2 días	0.51	0.22	0.51	0.51
3 días	0.38	0.24	0.38	0.38
4 días	0.51	0.36	0.51	0.51
5 días	0.53	0.42	0.53	0.53
6 días	0.18	0.35	0.18	0.18
7 días	<b>0.07*</b>	0.19	<b>0.07*</b>	<b>0.07*</b>
8 días	0.14	0.30	0.14	0.14
9 días	0.13	0.18	0.13	0.13
10 días	<b>0.06*</b>	0.24	<b>0.06*</b>	<b>0.06*</b>

Cuadro 39: Significación estadística (P-valores) de la correlación bivalente de causalidad de Granger para las variables de opinión y el S&P 500.binary para el año 2018

Modelo Logit para entender la relación entre el resultado del test de Granger de la variable del Ratio de Negatividad retardada y los rendimientos del S&P 500 para el año 2018 que se puede observar en el cuadro 39.

Listing 4: R output Modelo Logit con Ratio de Negatividad para 2018

```
Call:
glm(formula = df.SP_2018$binary ~ lagpad(df.SP_2018$binary, 1) +
    lagpad(df.SP_2018$binary, 2) + lagpad(df.SP_2018$binary,
    3) + lagpad(df.SP_2018$binary, 4) + lagpad(df.SP_2018$binary,
    5) + lagpad(df.SP_2018$binary, 6) + lagpad(df.SP_2018$binary,
    7) + lagpad(df.SP_2018$binary, 8) + lagpad(df.SP_2018$binary,
    9) + lagpad(df.SP_2018$binary, 10) + lagpad(df.SP_2018$ratio.neg.scaled,
    1) + lagpad(df.SP_2018$ratio.neg.scaled, 2) + lagpad(df.SP_2018$ratio.neg.
    scaled,
    3) + lagpad(df.SP_2018$ratio.neg.scaled, 4) + lagpad(df.SP_2018$ratio.neg.
    scaled,
    5) + lagpad(df.SP_2018$ratio.neg.scaled, 6) + lagpad(df.SP_2018$ratio.neg.
    scaled,
    7) + lagpad(df.SP_2018$ratio.neg.scaled, 8) + lagpad(df.SP_2018$ratio.neg.
    scaled,
    9) + lagpad(df.SP_2018$ratio.neg.scaled, 10), family = binomial(link = "logit"
    ))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1029  -0.9657   0.4355   0.9684   2.0404

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.40059    0.35956  -1.114  0.2652
lagpad(df.SP_2018$binary, 1)  1.25177    0.28931  4.327 1.51e-05 ***
lagpad(df.SP_2018$binary, 2)  0.64796    0.29612  2.188  0.0287 *
lagpad(df.SP_2018$binary, 3) -0.39008    0.30206  -1.291  0.1966
lagpad(df.SP_2018$binary, 4)  0.01810    0.30237  0.060  0.9523
lagpad(df.SP_2018$binary, 5) -0.32415    0.29815  -1.087  0.2769
lagpad(df.SP_2018$binary, 6)  0.08279    0.29922  0.277  0.7820
lagpad(df.SP_2018$binary, 7) -0.63491    0.29489  -2.153  0.0313 *
lagpad(df.SP_2018$binary, 8) -0.30544    0.29347  -1.041  0.2980
lagpad(df.SP_2018$binary, 9)  0.59379    0.29599  2.006  0.0448 *
lagpad(df.SP_2018$binary, 10) -0.24821    0.28707  -0.865  0.3872
lagpad(df.SP_2018$ratio.neg.scaled, 1)  0.00913    0.12480  0.073  0.9417
lagpad(df.SP_2018$ratio.neg.scaled, 2)  0.19028    0.12670  1.502  0.1331
lagpad(df.SP_2018$ratio.neg.scaled, 3)  0.13981    0.12771  1.095  0.2736
lagpad(df.SP_2018$ratio.neg.scaled, 4) -0.11012    0.12807  -0.860  0.3899
lagpad(df.SP_2018$ratio.neg.scaled, 5) -0.22479    0.13583  -1.655  0.0979 .
lagpad(df.SP_2018$ratio.neg.scaled, 6)  0.22256    0.13277  1.676  0.0937 .
lagpad(df.SP_2018$ratio.neg.scaled, 7)  0.12105    0.13553  0.893  0.3717
lagpad(df.SP_2018$ratio.neg.scaled, 8) -0.08461    0.13339  -0.634  0.5259
lagpad(df.SP_2018$ratio.neg.scaled, 9)  0.06598    0.13382  0.493  0.6220
lagpad(df.SP_2018$ratio.neg.scaled, 10) 0.29817    0.13504  2.208  0.0272 *
---
Signif. codes:  0   ***    0.001   **    0.01   *    0.05   .    0.1    1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 396.47 on 285 degrees of freedom
Residual deviance: 337.24 on 265 degrees of freedom
(79 observations deleted due to missingness)
AIC: 379.24

Number of Fisher Scoring iterations: 4
```

Modelo Logit para entender la relación entre el resultado del test de Granger de la variable del Ratio de Positividad retardada y los rendimientos del S&P 500 para el año 2018 que se puede observar en el cuadro 39.

Listing 5: R output Modelo Logit con Ratio de Positividad para 2018

```
Call:
glm(formula = df.SP_2018$binary ~ lagpad(df.SP_2018$binary, 1) +
    lagpad(df.SP_2018$binary, 2) + lagpad(df.SP_2018$binary,
    3) + lagpad(df.SP_2018$binary, 4) + lagpad(df.SP_2018$binary,
    5) + lagpad(df.SP_2018$binary, 6) + lagpad(df.SP_2018$binary,
    7) + lagpad(df.SP_2018$binary, 8) + lagpad(df.SP_2018$binary,
    9) + lagpad(df.SP_2018$binary, 10) + lagpad(df.SP_2018$ratio.pos.scaled,
    1) + lagpad(df.SP_2018$ratio.pos.scaled, 2) + lagpad(df.SP_2018$ratio.pos.
    scaled,
    3) + lagpad(df.SP_2018$ratio.pos.scaled, 4) + lagpad(df.SP_2018$ratio.pos.
    scaled,
    5) + lagpad(df.SP_2018$ratio.pos.scaled, 6) + lagpad(df.SP_2018$ratio.pos.
    scaled,
    7) + lagpad(df.SP_2018$ratio.pos.scaled, 8) + lagpad(df.SP_2018$ratio.pos.
    scaled,
    9) + lagpad(df.SP_2018$ratio.pos.scaled, 10), family = binomial(link = "logit"
    ))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1029  -0.9657   0.4355   0.9684   2.0404

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.40059    0.35956  -1.114  0.2652
lagpad(df.SP_2018$binary, 1)  1.25177    0.28931  4.327 1.51e-05 ***
lagpad(df.SP_2018$binary, 2)  0.64796    0.29612  2.188  0.0287 *
lagpad(df.SP_2018$binary, 3) -0.39008    0.30206  -1.291  0.1966
lagpad(df.SP_2018$binary, 4)  0.01810    0.30237  0.060  0.9523
lagpad(df.SP_2018$binary, 5) -0.32415    0.29815  -1.087  0.2769
lagpad(df.SP_2018$binary, 6)  0.08279    0.29922  0.277  0.7820
lagpad(df.SP_2018$binary, 7) -0.63491    0.29489  -2.153  0.0313 *
lagpad(df.SP_2018$binary, 8) -0.30544    0.29347  -1.041  0.2980
lagpad(df.SP_2018$binary, 9)  0.59379    0.29599  2.006  0.0448 *
lagpad(df.SP_2018$binary, 10) -0.24821    0.28707  -0.865  0.3872
lagpad(df.SP_2018$ratio.pos.scaled, 1) -0.00913    0.12480  -0.073  0.9417
lagpad(df.SP_2018$ratio.pos.scaled, 2) -0.19028    0.12670  -1.502  0.1331
lagpad(df.SP_2018$ratio.pos.scaled, 3) -0.13981    0.12771  -1.095  0.2736
lagpad(df.SP_2018$ratio.pos.scaled, 4)  0.11012    0.12807  0.860  0.3899
lagpad(df.SP_2018$ratio.pos.scaled, 5)  0.22479    0.13583  1.655  0.0979 .
lagpad(df.SP_2018$ratio.pos.scaled, 6) -0.22256    0.13277  -1.676  0.0937 .
lagpad(df.SP_2018$ratio.pos.scaled, 7) -0.12105    0.13553  -0.893  0.3717
lagpad(df.SP_2018$ratio.pos.scaled, 8)  0.08461    0.13339  0.634  0.5259
lagpad(df.SP_2018$ratio.pos.scaled, 9) -0.06598    0.13382  -0.493  0.6220
lagpad(df.SP_2018$ratio.pos.scaled, 10) -0.29817    0.13504  -2.208  0.0272 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1  1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 396.47 on 285 degrees of freedom
Residual deviance: 337.24 on 265 degrees of freedom
(79 observations deleted due to missingness)
AIC: 379.24

Number of Fisher Scoring iterations: 4
```

Modelo Logit para entender la relación entre el resultado del test de Granger de la variable del Índice de Polaridad 1 retardada y los rendimientos del S&P 500 para el año 2018 que se puede observar en el cuadro 39.

Listing 6: R output Modelo Logit con Índice de Polaridad 1 para 2018

```
Call:
glm(formula = df.SP_2018$binary ~ lagpad(df.SP_2018$binary, 1) +
    lagpad(df.SP_2018$binary, 2) + lagpad(df.SP_2018$binary,
    3) + lagpad(df.SP_2018$binary, 4) + lagpad(df.SP_2018$binary,
    5) + lagpad(df.SP_2018$binary, 6) + lagpad(df.SP_2018$binary,
    7) + lagpad(df.SP_2018$binary, 8) + lagpad(df.SP_2018$binary,
    9) + lagpad(df.SP_2018$binary, 10) + lagpad(df.SP_2018$index.pol.scaled,
    1) + lagpad(df.SP_2018$index.pol.scaled, 2) + lagpad(df.SP_2018$index.pol.
    scaled,
    3) + lagpad(df.SP_2018$index.pol.scaled, 4) + lagpad(df.SP_2018$index.pol.
    scaled,
    5) + lagpad(df.SP_2018$index.pol.scaled, 6) + lagpad(df.SP_2018$index.pol.
    scaled,
    7) + lagpad(df.SP_2018$index.pol.scaled, 8) + lagpad(df.SP_2018$index.pol.
    scaled,
    9) + lagpad(df.SP_2018$index.pol.scaled, 10), family = binomial(link = "logit"
    ))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1029  -0.9657   0.4355   0.9684   2.0404

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.40059    0.35956  -1.114   0.2652
lagpad(df.SP_2018$binary, 1)  1.25177    0.28931   4.327 1.51e-05 ***
lagpad(df.SP_2018$binary, 2)  0.64796    0.29612   2.188   0.0287 *
lagpad(df.SP_2018$binary, 3) -0.39008    0.30206  -1.291   0.1966
lagpad(df.SP_2018$binary, 4)  0.01810    0.30237   0.060   0.9523
lagpad(df.SP_2018$binary, 5) -0.32415    0.29815  -1.087   0.2769
lagpad(df.SP_2018$binary, 6)  0.08279    0.29922   0.277   0.7820
lagpad(df.SP_2018$binary, 7) -0.63491    0.29489  -2.153   0.0313 *
lagpad(df.SP_2018$binary, 8) -0.30544    0.29347  -1.041   0.2980
lagpad(df.SP_2018$binary, 9)  0.59379    0.29599   2.006   0.0448 *
lagpad(df.SP_2018$binary, 10) -0.24821    0.28707  -0.865   0.3872
lagpad(df.SP_2018$index.pol.scaled, 1) -0.00913    0.12480  -0.073   0.9417
lagpad(df.SP_2018$index.pol.scaled, 2) -0.19028    0.12670  -1.502   0.1331
lagpad(df.SP_2018$index.pol.scaled, 3) -0.13981    0.12771  -1.095   0.2736
lagpad(df.SP_2018$index.pol.scaled, 4)  0.11012    0.12807   0.860   0.3899
lagpad(df.SP_2018$index.pol.scaled, 5)  0.22479    0.13583   1.655   0.0979 .
lagpad(df.SP_2018$index.pol.scaled, 6) -0.22256    0.13277  -1.676   0.0937 .
lagpad(df.SP_2018$index.pol.scaled, 7) -0.12105    0.13553  -0.893   0.3717
lagpad(df.SP_2018$index.pol.scaled, 8)  0.08461    0.13339   0.634   0.5259
lagpad(df.SP_2018$index.pol.scaled, 9) -0.06598    0.13382  -0.493   0.6220
lagpad(df.SP_2018$index.pol.scaled, 10) -0.29817    0.13504  -2.208   0.0272 *
---
Signif. codes:  0   ***    0.001   **    0.01   *    0.05   .    0.1    1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 396.47 on 285 degrees of freedom
Residual deviance: 337.24 on 265 degrees of freedom
(79 observations deleted due to missingness)
AIC: 379.24

Number of Fisher Scoring iterations: 4
```

# Apéndice D Estudio de Eventos aplicado a Twitter

## D.1 Código para el estudio de eventos

```
abnormal <- function(num_tweet,eco_df,df, days_before, days_after, estimation_
  window){

  if(eco_df[num_tweet,]$hour >= 21 ){ #tanca a les 9 utc (hora anglesa) --->
    aquell dia compte

    dates <- seq(from = eco_df[num_tweet,]$anyo_mes
      - days_before + 1, to = eco_df[num_tweet,]$anyo_mes, by = 1)
    #dia inclous

    efecte <- seq(from = eco_df[num_tweet,]$anyo_mes+1, to = eco_df[num_tweet,]$
      anyo_mes + days_after, by = 1)

    abnormal <- vector()
    caar <- vector()

    for (i in 1:(days_after + days_before)){
      abnormal[i] <- df[df$dia == eco_df[num_tweet,]$anyo_mes -
        days_after + i,]$return -
        mean(df[df$dia %in% dates | df$dia %in% efecte,]$return, na.rm = TRUE)
      caar[i] <- sum(abnormal)
    }

  } else {

    dates <- seq(from = eco_df[num_tweet,]$anyo_mes
      - days_before , to = eco_df[num_tweet,]$anyo_mes -1, by = 1)
    #dia NO inclous

    efecte <- seq(from = eco_df[num_tweet,]$anyo_mes, to = eco_df[num_tweet,]$anyo
      _mes + days_after - 1, by = 1)

    abnormal <- vector()
    caar <- vector()

    for (i in 1:(days_after + days_before)){
      abnormal[i] <- df[df$dia == (eco_df[num_tweet,]$anyo_mes -
        days_before + i - 1),]$return -
        mean(df[df$dia %in% dates | df$dia %in% efecte,]$return, na.rm = TRUE)
      caar[i] <- sum(abnormal)
    }

    estimation <- seq(from = eco_df[num_tweet,]$anyo_mes - estimation_window - 10,
      to = eco_df[num_tweet,]$anyo_mes - 10 - 1, by = 1)

    sd1 <- sd(df[df$dia %in% estimation,]$return, na.rm = TRUE)

    t <- abnormal / sd(df[df$dia %in% estimation,]$return, na.rm = TRUE)

    t_caar <- caar / (sqrt(days_after + abs(days_before) + 1)*
      sd(df[df$dia %in% estimation,]$return, na.rm = TRUE))
  }
}
```

```

}

return(list(t.test(df[df$dia %in% dates,]$return,
df[df$dia %in% efecte,]$return), abnormal, caar, t, t_caar, sd1))

}

```

## Apéndice E SVM

### E.1 Funciones utilizadas para la creación de los diferentes modelos SVM

```

# SIN SENTIMENTAL Y Kernel LINEAL:

svm_cv <- tune("svm", binary ~ lag1 + lag2 + lag3,data= DJIA .SVM[learning,] ,
  Kernel="linear",scale=TRUE,
  ranges = list(cost = c(0.00001,0.0001,0.001,0.01,0.1,1:10 )),class.
  weights = c("0" = 1.25, "1" = 1))

svm1 <- svm(binary ~ lag1 + lag2 + lag3,data= DJIA .SVM ,Kernel="linear",scale=
  TRUE,subset=learning, cost = 1,
  class.weights = c("0" = 1.25, "1" = 1))

# SIN SENTIMENTAL Y Kernel Radial:

svm_cv <- tune("svm", binary ~ lag1 + lag2 + lag3,data= DJIA .SVM[learning,],
  Kernel = 'Radial',
  ranges = list(cost = c(0.01, 0.1, 1, 10,20,30),
  gamma = c(0.5, 1, 2, 5, 10)))

svm1 <- svm(binary ~ lag1 + lag2 + lag3,data= DJIA .SVM ,Kernel="Radial",scale=
  TRUE,subset=learning, cost = 10, gamma = 2,
  class.weights = c("0" = 1.25, "1" = 1))

# CON SENTIMENTAL Y Kernel LINEAL:

svm_cv2 <- tune("svm", binary ~ lag1 + lag2 + lag3 + ratio_neg_lag3,data= DJIA .
  SVM[learning,] ,Kernel="linear",scale=TRUE,
  ranges = list(cost = c(0.00001,0.01,0.1,1:10 )),class.weights = c("
  0" = 1.25, "1" = 1))

svm2 <-svm(binary ~ lag1 + lag2 + lag3 + ratio_neg_lag3,data= DJIA .SVM ,Kernel="
  linear",scale=TRUE,
  subset = learning,class.weights = c("0" = 1.25, "1" = 1), cost = 1)

# CON SENTIMENTAL Y Kernel Radial

svm_cv2 <- tune("svm", binary ~ lag1 + lag2 + lag3 + ratio_neg_lag3,data= DJIA .
  SVM[learning,], Kernel = 'Radial',
  ranges = list(cost = c(0.01, 0.1, 1, 10, 20),
  gamma = c(0.5, 1, 2, 5, 10)))

svm2 <- svm(binary ~ lag1 + lag2 + lag3+ ratio_neg_lag3, DJIA .SVM ,Kernel="

```

```
Radial",scale=TRUE,subset=learning, cost = 1, gamma = 1,class.weights = c("0" =
1.25, "1" = 1))
```

## E.2 Ejemplo visual de SVM con dos variables predictoras

Gracias a la utilización de solo dos variables predictoras, se puede ver visualmente en un gráfico como se realiza la separación Lineal y la separación de Kernel Radial sobre dos variables, para predecir la categoría de la variables respuesta.

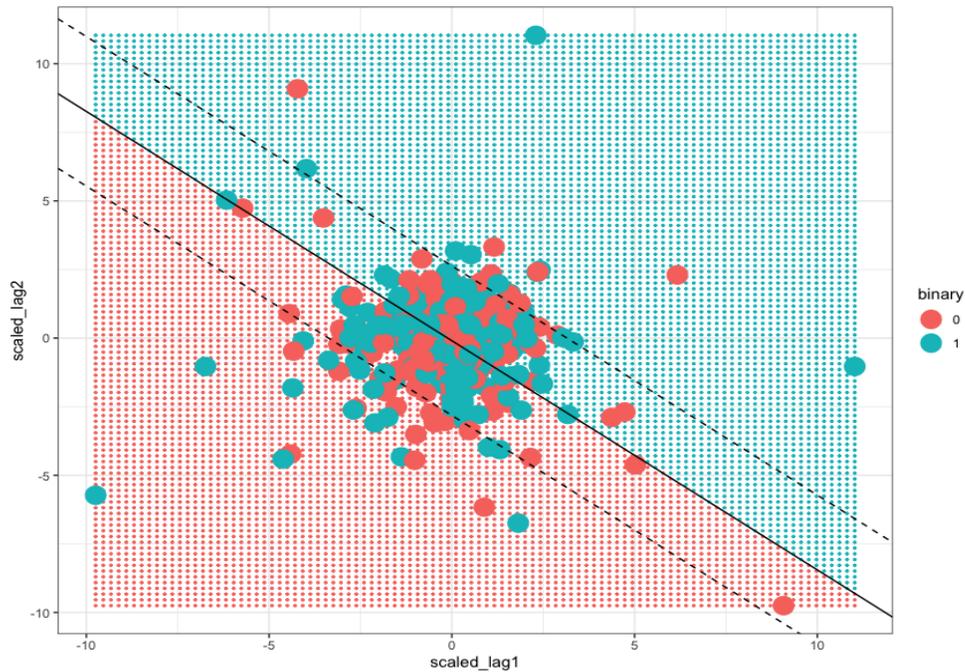


Figura 38: SVM para dos variables predictoras y Kernel Lineal

Se puede ver en la figura 38 como la separación lineal para discernir si los retornos del día  $t$  serán positivos (azul) o negativos (rojo) en función de: El valor numérico del rendimiento del día anterior ( $t_{-1}$ ) y el valor numérico del rendimiento del día anterior al anterior ( $t_{-2}$ ). Esta separación deja muchas zonas mal clasificadas, encontrándose la mayoría de valores a una lado o a otro de la separación lineal de forma equiprobable.

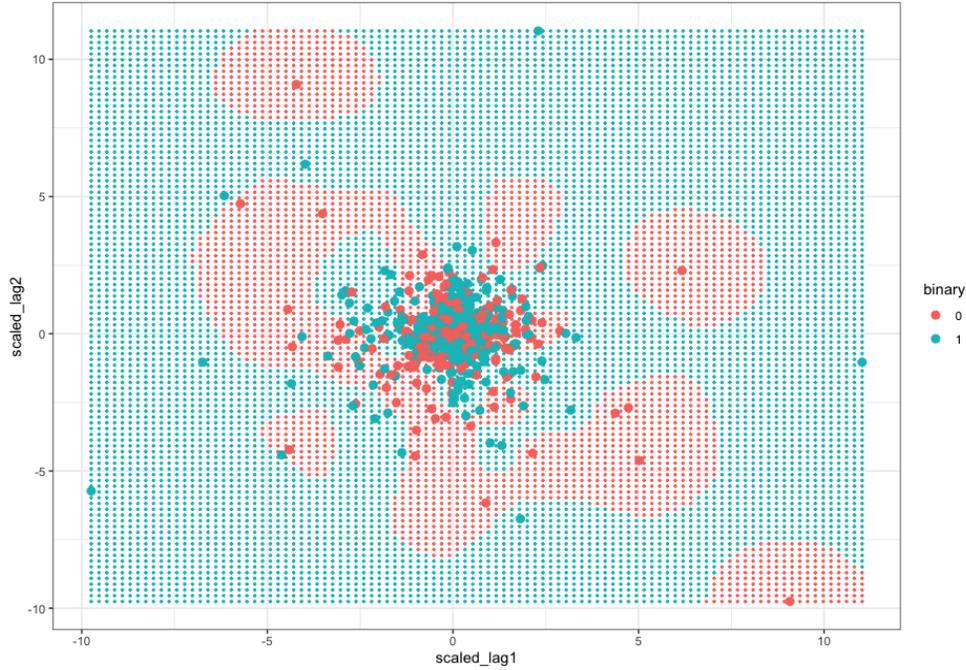


Figura 39: SVM para dos variables predictoras y Kernel Radial

Se puede ver en la figura 39, la separación con un Kernel Radial para discernir si los retornos del día  $t$  serán positivos (azul) o negativos (rojo) en función de: El valor numérico del rendimiento del día anterior ( $t_{-1}$ ) y el valor numérico del rendimiento del día anterior al anterior ( $t_{-2}$ ). Esta separación mejora la clasificación en diferentes zonas de este espacio de dos dimensiones. Se puede ver como el hecho de añadir complejidad al espacio de separación se consigue clasificar de forma más precisa la categoría de la variable respuesta.

## Apéndice F Resto del código en R

En el siguiente enlace se encuentra el código de R utilizado en este estudio por sus diferentes secciones. También se puede encontrar los datos utilizados.

TFG\_TRUMP.