Universidad de SanAndrés

TRABAJO DE GRADUACIÓN

LICENCIATURA EN FINANZAS

“Uso de información en Twitter como herramienta para predecir el mercado financiero”

Autores:
Castagno, Ivan - 29036
Guananja, Nahuel - 29318

Tutor:
Margaretic, Paula

Marzo 2022, Buenos Aires, Argentina.

Índice

1. Introducción

Decir que las redes sociales han cambiado por completo la existencia no es nada nuevo. Han revolucionado las comunicaciones, hasta llegar a ser el medio de comunicación cotidiano. Antes de que existan y sean utilizadas masivamente, las noticias se divulgaban mediante medios tradicionales como la televisión y los periódicos, y los rumores se esparcían en conversaciones informales en puntos de encuentro como cafeterías o bares. Pero hoy esto ha cambiado, solo basta con un clic o dos para leer los periódicos de cualquier parte del mundo y estos ya no son los que tienen las primicias. Gran parte de las noticias suelen divulgarse primero a través de redes sociales como Twitter, desplazando así a los medios de comunicación tradicionales a un rol secundario.

Las redes sociales son una sofisticada herramienta multidisciplinaria que permite a los individuos crear contenidos, comunicarse entre sí e incluso evadirse de la realidad. Hoy, a través de ellas, se pueden enviar datos de una punta a otra del mundo en cuestión de segundos, hacer presentaciones en línea, vivir en "mundos de juego" paralelos y utilizar imágenes, vídeo, sonido y texto para compartir parte de la vida cotidiana. Las historias personales se hacen públicas y los asuntos locales se convierten en globales en cuestión de segundos.

El auge de las redes sociales ha suscitado un debate sobre cómo la comunicación en línea afecta a las relaciones sociales. Estas se liberan de las ataduras geográficas y reúnen en comunidades temáticas que no están atadas a ningún lugar concreto. La nuestra es una sociedad en red, globalizada y conectada por las nuevas tecnologías. En conclusión, el impacto de las redes sociales es indiscutible y abarca diversas disciplinas.

Una de estas disciplinas, y sobre la que se va a desarrollar este trabajo, son las finanzas. Los debates que antes se daban en ámbitos académicos y o profesionales, hoy en día se dan en las redes sociales, especialmente en Twitter, y los participantes de estos son, principalmente, inversores minoristas y no calificados o profesionales. De esta forma, se genera un fenómeno que democratiza las finanzas, dándole acceso a cualquier inversor o potencial inversor a información que antes solo inversores calificados o profesionales que se encontraban en el ambiente financiero profesional podían acceder.

Por ejemplo, en Twitter es común encontrar usuarios escribiendo sobre recomendaciones de compra o venta de acciones, opiniones y análisis del estado financiero de compañías que cotizan en el mercado, o simplemente comentando que ellos abrieron una posición larga o corta en determinado activo a través de lo que en esta red social es llamado *tweet*. Algunos de estos usuarios, desarrollan un público de inversores que replican sus estrategias, en otras palabras, existen inversores que simplemente deciden que activo comprar o que activo vender debido a la recomendación de otro usuario en la misma red social.

Por otra parte, en esta misma red social se observan *tweets* en los que se mencionan, en conjunto, dos o más activos financieros como, por ejemplo, la acción de Amazon junto con la de Apple o cualquier otro activo financiero. Se observa también en esta red social, la tendencia a mencionar

activos financieros junto a otros activos financieros que en ese mismo momento están teniendo retornos positivos. Por ejemplo, un día que la acción de Apple tiene retornos en la magnitud del 5%, muchos usuarios tienden a mencionar en *tweets*, activos financieros que poco tienen que ver con Apple. Las razones de esto no son claras, pero se podría llegar a inferir que, cómo Apple tiene retornos positivos, el o los activos financieros que se mencionen junto con Apple obtendrán retornos positivos también.

Este fenómeno de las menciones conjuntas de activos financieros en *tweets* permite inferir que, algunos de los usuarios que escriben este tipo de *tweets* creen que mediante esta mención conjunta logran influenciar la decisión de otros inversores que operan y consecuentemente, mediante la masividad del alcance de las redes sociales, generan un efecto en cadena que termina impactando en el mercado financiero.

La masividad y la velocidad en la que se transmiten mensajes a través de *tweets* en Twitter lleva a preguntar si estos pueden llegar a ser la causa, y no la consecuencia, de los movimientos en los mercados financieros.

En este trabajo, a través de la teoría de grafos, se busca encontrar respuestas a este fenómeno mencionado anteriormente. Para el armado de la red, se utilizó una adaptación de un código de python de Ramiro Gálvez. Luego, se extraen las estadísticas resultantes de la aplicación de la teoría de grafos, con el objetivo de responder específicamente a las siguientes preguntas: ¿Cuál es el contenido informativo de la información que circula en Twitter? ¿Es posible predecir a los mercados financieros a través de esta?

2. Revisión de literatura

De acuerdo con la teoría de mercado eficiente, toda la información se refleja inmediatamente en los precios del mercado (Fama 1970). Por lo tanto, en teoría los inversores deberían ser incapaces de obtener una rentabilidad superior a la media del mercado a largo plazo.

Sin embargo, en diversos trabajos previos, se ha observado empíricamente algunas anomalías en los mercados financieros que pueden ser vistas cómo una excepción a la teoría de mercado eficiente, y que la hipótesis nula de mercado eficiente puede ser rechazada en altos niveles de confianza Rosenberg, Reid, Lanstein (1985). Con el paso de los años y el avance tecnológico, en diversos trabajos se intenta encontrar anomalías de mercado a través de estrategias o meras observaciones que involucren datos de redes sociales o datos de alta frecuencia. Agrawal, Azar, Lo y Singh (2018) analizan el contenido informativo del estado del *social sentiment* de las redes sociales StockTwits y Twitter, y concluyen que es probable que el *social sentiment* de StockTwits y Twitter sea consecuencia de los movimientos en los precios; pero a su vez también es plausible que el *social sentiment* afecte a las propias operaciones, ya sea a través de los usuarios de estas redes sociales o a través de los inversores institucionales que llevan a cabo estrategias que utilizan las señales que proporcionan estos sitios.

Por otra parte, en Galvez y Gravano (2017) se analiza si mensajes extraídos de una red social popular de inversores en Argentina tienen poder predictivo sobre el comportamiento futuro de los

precios de las acciones. Los autores construyen y validan una serie de modelos predictivos utilizando técnicas de aprendizaje automático de última generación. Concluyen que el trabajo arroja evidencia que sugiere que los mensajes extraídos de la red social pueden ser utilizados para entrenar modelos que predigan eficazmente el comportamiento futuro de las acciones con mayor eficacia que conjeturas al azar.

En Pagolu, Challa, Panda y Majhi (2016) se aplica el análisis de *social sentiment* y principios de aprendizaje automático supervisado a los *tweets* extraídos de Twitter, y se analiza la correlación entre los movimientos de acciones en el mercado bursátil y el *social sentiment*. Se demuestra que existe una fuerte correlación entre los movimientos en los precios de las acciones y el *social sentiment* extraído de *tweets*.

En Tan, Lee, Tang, Jiang, Zhou y Li (2011) se demuestra que la información sobre las relaciones sociales, extraída a través de teoría de grafos, puede utilizarse para mejorar el análisis del *social sentiment*. Los usuarios que están de alguna manera "conectados" pueden tener opiniones similares; por tanto, el contenido informativo de los grafos puede complementar lo que se puede extraer sobre la opinión de un usuario. Si bien este trabajo no realiza ningún estudio relacionado con los mercados financieros, demuestra la importancia del contenido informativo de la teoría de grafos.

Continuando ya con la teoría de grafos, pero ahora aplicada a las finanzas, en Gross y Siklos (2020) se arma una red de contagio que revela que las instituciones financieras que generan los mayores efectos de contagio dentro del sistema financiero son también las más importantes transmisoras de los efectos de contagio a la economía real.

En Ruiz et al. (2012) se construyen grafos diarios para cada acción basándose en los *retweets* y otras características (por ejemplo, se considera que Alice y Bob están conectados en la red social de YHOO en un día determinado si Alice tuitea sobre YHOO y Bob retuitea su mensaje ese día) y se demuestra que las características de estos grafos se correlacionan con el volumen de operaciones y los rendimientos de las acciones.

Por último, en Sharma y Habib (2019) se estudian las interacciones entre las acciones del mercado de bursátil de India a un nivel de alta frecuencia a través de grafos. Para ello, se eligen intervalos de 30 segundos y se estudia el comportamiento de las 89 acciones negociadas con mayor frecuencia de este mercado en el 2014. Luego, se analiza la matriz de correlación para estudiar la aleatoriedad y se concluye que los coeficientes de correlación entre pares de acciones no siguen comportamientos aleatorios. Por último, se comparan los coeficientes de correlación por pares con sus respectivos puntos de información mutua (PMI). El análisis muestra la existencia de no linealidad en los datos de alta frecuencia, fenómeno que el punto de información mutua logra captar muy bien. Por lo tanto, se concluye que los grafos basados en el punto de información mutua capturan mejor la dinámica entre las acciones a un nivel de alta frecuencia que los grafos basados en el coeficiente de correlación.

3. ¿Qué es una red social o grafo?

Antes de empezar a comentar la metodología utilizada en el trabajo para realizar el análisis propuesto, es importante explicar que es una red social o grafo. Una red social es una estructura compuesta por tres componentes principales: los nodos -o vértices-; las relaciones que conectan estos nodos -las líneas-; y, por último, la estructura de la red que muestra la forma en la que estos se relacionan. Existen diferentes tipos de redes ya que cualquiera de estos componentes puede variar.

En cuanto a los nodos, se puede encontrar dos tipos principales de redes: las unimodales y las multimodales. Las primeras se caracterizan por tener todos los vértices de un mismo tipo, mientras que la segunda permite tener dos o más tipos distintos de vértices. Un ejemplo de las unimodales sería simplemente una red social compuesta por distintas personas. Para el caso de las multimodales se puede pensar que hay personas, y que también están las empresas en las que estas personas trabajan. De esta manera, coexisten dos tipos distintos de nodos en la misma red social.

En cuanto a las relaciones entre los nodos, pueden ser de dos formas distintas: no dirigidas o dirigidas. La línea que conecta los distintos nodos en las no dirigidas se conoce como arista, mientras que en las dirigidas se la conoce como arco. Este arco es direccional y tiene una flecha para mostrar esta relación.

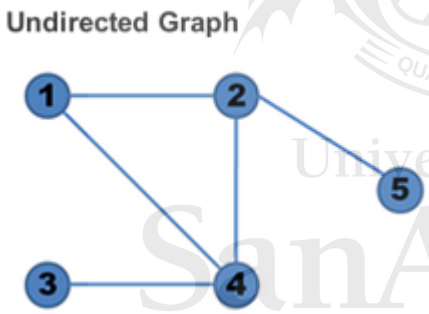


Figura 3.1: ejemplo red no dirigida

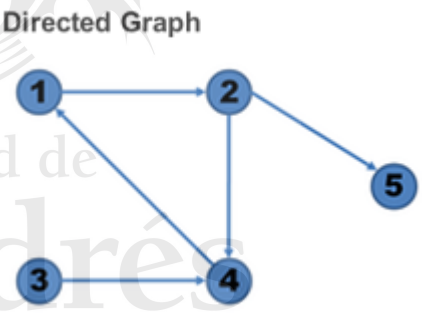


Figura 3.2: ejemplo red dirigida

También existen otras dos diferencias al momento de analizar las aristas, ya que estas pueden ser ponderadas o no ponderadas. En el caso de las no ponderadas -también conocidas como dicotómicas o binarias- las aristas pueden tomar únicamente el valor uno o cero, indicando de esta manera si existe una relación entre los nodos o no. En cambio, en el caso de las relaciones ponderadas, las aristas pueden tomar infinitos valores.

Esto depende de la escala que se utilice, ya que por ejemplo se podrían asignar valores de cero a infinito o valores entre cero y uno. Es interesante notar que, por este motivo, las relaciones ponderadas permiten obtener más información de una red social al entender el grado en el que los nodos están conectados.

4. Metodología

4.1 Recolección de datos

En primer lugar, se necesitó conseguir los datos necesarios para poder realizar el análisis. Para esto, se usó un *scraper* armado en python que permite obtener todos los *tweets*, entre dos fechas fijadas por el usuario, donde se mencionan los *tickers* elegidos de cada empresa. Estos *tweets* se guardan en una estructura diaria, para luego poder analizar la variación diaria de las redes sociales. La fecha de inicio de la muestra de datos es el 01/11/2018 y la última fecha es el 01/11/2020, para trabajar con 2 años completos de información. Se descargaron todos los *tweets* donde aparecía mencionada al menos una empresa perteneciente al S&P 500.

4.2 Armado de red

Luego de descargar todos los *tweets* en los que se mencionan los *tickers* de las empresas que componen el S&P 500, se extraen los *cashtags* ($AMZN, $AAPL, $MSFT, etc) de cada *tweet*. Es importante notar que el hecho de buscar los *tweets* en los que aparecen los *tickers* de las acciones del índice no limita únicamente a trabajar con estas acciones, ya que cada *tweet* puede tener a su vez otros *cashtags* de empresas que no pertenezcan al índice, de criptomonedas u otros activos financieros. Este es un ejemplo, al buscar un *tweet* donde aparezca "$AAPL" se pueden encontrar otros *tickers* que no necesariamente deben pertenecer al S&P 500. Esto genera un efecto de expansión de la red, al aumentar considerablemente la cantidad de nodos y aristas en cada día.

Figura 4.1: ejemplo de *tweet* con *cashtags*

Una vez que se guardaron todos los *cashtags* encontrados, el siguiente paso es armar el grafo -o red- para intentar encontrar relaciones entre estos *tickers*. La red social que se utilizó para el análisis

propuesto es unimodal en cuanto a los nodos y no dirigida ponderada con respecto a las relaciones entre los vértices. Como se comentó con anterioridad al explicar el concepto de red social, el ponderador puede variar dependiendo la necesidad para el análisis a realizar. En este caso, se decidió relacionar los nodos de cada *ticker* utilizando el *Positive Pointwise Mutual Information* (PPMI) entre estos dos, que es una extensión del *Pointwise Mutual Information* (PMI).

4.3 *Positive Pointwise Mutual Information*

En esta sección se explicará con mayor detalle el significado del PMI y PPMI, y el motivo por el cual se tomó la decisión de utilizarlo para relacionar los nodos de la red social o grafo. El PMI se usa en gran medida en aplicaciones relacionadas al procesamiento de lenguaje, ya que por ejemplo permite encontrar relaciones entre palabras. La idea es contar la cantidad de veces que una palabra aparece junto a otra, y relacionarla con la probabilidad de que ambas palabras aparezcan separadas por puro azar. De esa manera, permite encontrar relaciones entre palabras utilizadas, para entender y predecir mejor el lenguaje natural a través de la computadora.

Teniendo en cuenta esto, se decidió aplicar el PMI al caso de análisis de apariciones de *cashtags.* La idea detrás de esto se basa en calcular si dos *cashtags* X e Y están apareciendo juntos, en un momento dado del tiempo, una mayor cantidad de veces de lo que se esperaría que aparecieran aleatoriamente. De esta manera, se podrían encontrar relaciones importantes entre distintas empresas o activos financieros, al ver que durante cierto día -o periodo del tiempo- se nombran juntos una mayor cantidad de veces que lo esperado.

La fórmula utilizada para calcular el valor del PMI aplicado al análisis en cuestión es la siguiente:

$$pmi(\$X, \$Y) = log_2\left(\frac{\#tweets\ con\ \$X\ y\ \$Y * \#tweets\ totales}{\#tweets\ \$X * \#tweets\ \$Y}\right)$$

y el PPMI no es más que el PMI, pero limitado a valores positivos, donde los valores negativos pasan a valer cero. Su fórmula es la siguiente, y este valor será el ponderador de las relaciones entre los *cashtags* en la red social:

$$ppmi(\$X, \$Y) = \max(pmi(\$X, \$Y), 0)$$

4.4 Estadísticas a nivel de red social

Una vez creada la red social para cada día entre las fechas seleccionadas, se calculan y almacenan ciertas estadísticas útiles a nivel de red, que luego se utilizarán para ver si encontramos algún resultado interesante.

Para esto se armó un *script* de python que calcula las estadísticas diariamente de los grafos y las convierte en series de tiempo, para poder analizar con respecto a distintas variables, como pueden ser los retornos de los activos, volumen operado, volatilidad u otras variables económicas o financieras que puedan ser de interés. En el presente trabajo nosotros comparamos contra volumen, ya que encontramos que había correlación. Las estadísticas que se incluyeron son las siguientes:

- Número de nodos
- Número de aristas
- PPMI promedio
- Eficiencia global
- Centralidad de vector propio (*Eigenvector centrality*)
- Coeficiente de agrupamiento promedio (*Average Clustering coefficient*)
- Cantidad de *Tweets* positivos
- Cantidad de *Tweets* negativos
- Cantidad de *Tweets* totales

A continuación, se explicarán algunas de estas estadísticas. En primer lugar, el número de nodos total de la red indica todas las acciones o activos distintos que se mencionaron a lo largo de un día en *twitter*. El número de aristas indica todas las conexiones que hubo entre estas acciones mencionadas. Las conexiones se dan cuando se nombran varios activos dentro del mismo *tweet*. La medida de centralidad de la red social mide la importancia de un nodo en la red y hay diferentes métodos para calcularlo. En este caso, utilizamos la medida de *Eigenvector centrality*. El *coeficiente de agrupamiento* mide cuan agrupados están los nodos de un grafo. Esta medida es a nivel de nodos, por lo que se calculó el promedio para poder tener la estadística a nivel de red.

Para definir la cantidad de *tweets* positivos y negativos, creamos una variable de *sentiment* a través de un algoritmo clasificador *NaiveBayes*. El algoritmo analiza los textos de cada *tweet* y devuelve un resultado de *polarity*. A continuación, mostramos algunos ejemplos:

- Paula is a great professor, her lectures are amazing. Polarity = 0.7
- $TSLA earnings look really bad. Polarity = -0.69
- Today is Wednesday. Polarity = 0

Luego de obtener el resultado de *polarity* de cada uno de los 30.000.000 *tweets*, definimos un umbral para el cual consideramos un *tweet* como positivo, negativo o neutral. Para el caso de *tweets* positivos tomamos el valor del percentil 75, de 0.375. Para los *tweets* negativos tomamos el valor del percentil 10, de -0.15. Los neutrales los definimos como los *tweets* con polarity entre -0.15 y 0.375. Entonces la división quedó de esta forma:

- Polarity >= 0.375 *tweet* positivo.

- -0.15 < Polarity < 0.375 *tweet* neutral.
- Polarity <= -0.15 *tweet* negativo.

5. Evolución de las estadísticas en el tiempo

En esta sección se mostrarán los gráficos de variación diaria de cada estadística utilizada para analizar la red. En la siguiente sección se explicará con mayor detalle, pero la línea vertical en los gráficos marca el día 6 de marzo de 2020, que fue el día previo al comienzo del crash bursátil y el día del caso 100 de COVID dentro de USA. A continuación, se puede ver la evolución de la cantidad de nodos y aristas de la red.

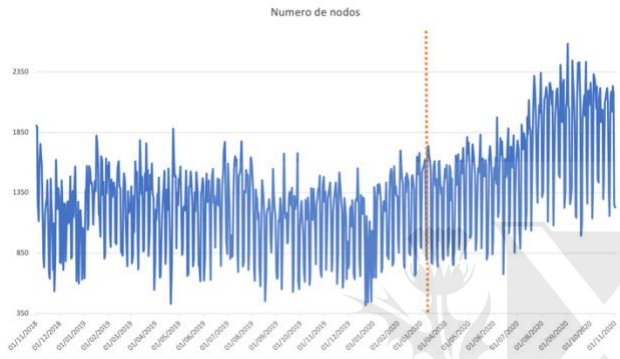


Figura 5.1: Evolución de la cantidad de nodos

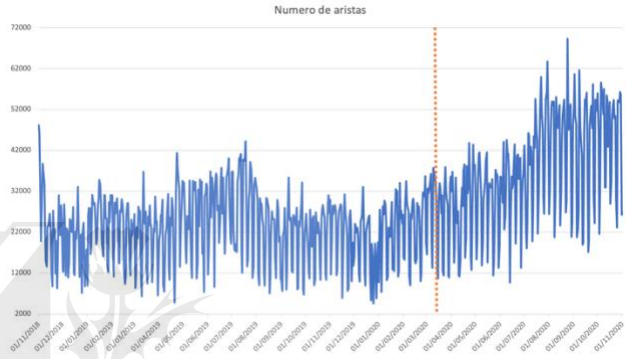


Figura 5.2: Evolución de la cantidad de aristas

Claramente se observa un cambio en la tendencia cercano a la fecha de la división, a partir de la cual empieza a aumentar la cantidad de nodos, y en consecuencia por la expansión de la red, también aumenta la cantidad de aristas. Los siguientes gráficos muestran la evolución del PPMI promedio y del grado promedio. En cuanto al primero, no se ve a simple vista un cambio significativo a partir de la fecha de corte; pero en cuanto al grado promedio, se puede observar un leve aumento hacia el futuro.

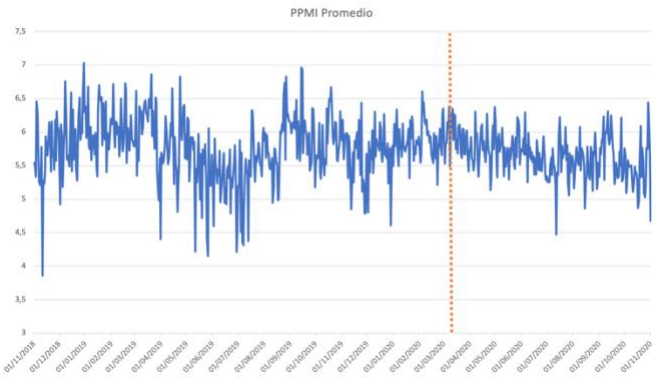


Figura 5.3: Evolución del PPMI promedio

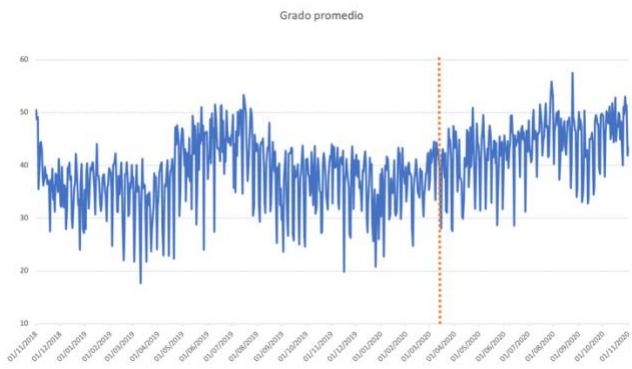


Figura 5.4: Evolución del grado promedio

Los valores de la eficiencia global se mantienen prácticamente variando en el mismo rango durante los dos años de observación; por lo que no es relevante destacar el gráfico. A continuación, se observa la evolución de la centralidad de vector propio y del coeficiente de agrupamiento promedio.

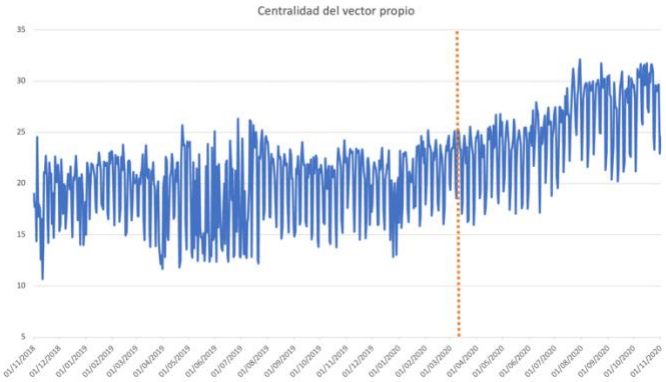


Figura 5.5: Evolución de la centralidad del vector propio

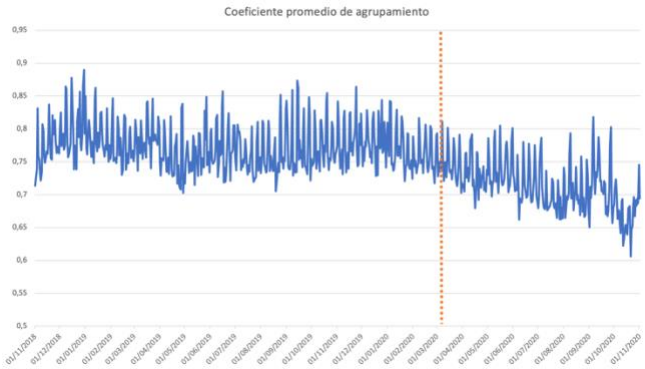


Figura 5.6: Evolución del coeficiente promedio de agrupamiento

Como se puede ver en el primer gráfico, la centralidad del vector propio tiene un comportamiento similar al observado en la evolución de la cantidad de nodos y de aristas. Luego de la fecha de corte, comienza una tendencia alcista de los valores que se mantiene hasta el ultimo día de la muestra. En el caso del coeficiente promedio de agrupamiento, se puede observar un comportamiento inverso; luego del 6 de marzo de 2020, comienza una tendencia bajista que también se observa hasta el 1 de noviembre de 2020, ultimo día de la muestra.

Por último, se muestran los gráficos de la evolución de la cantidad de *tweets* positivos, negativos y totales.

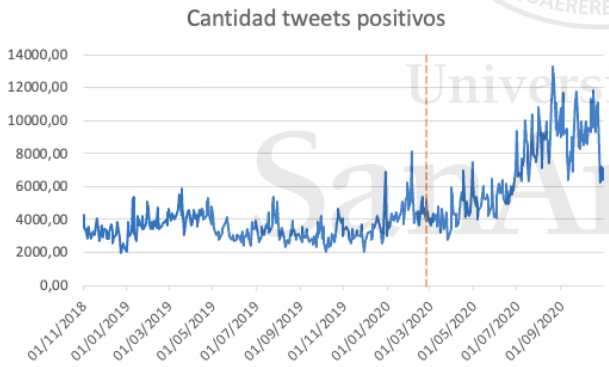


Figura 5.7: Cantidad *tweets* positivos

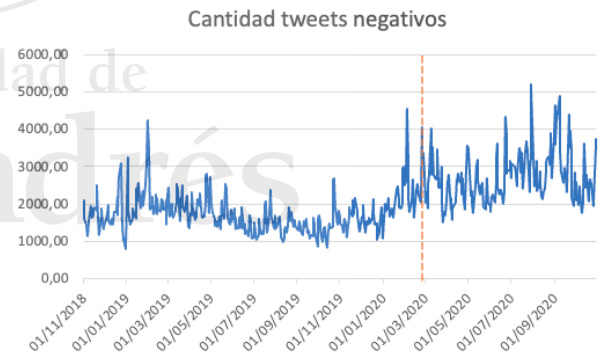


Figura 5.8: Cantidad *tweets* negativos

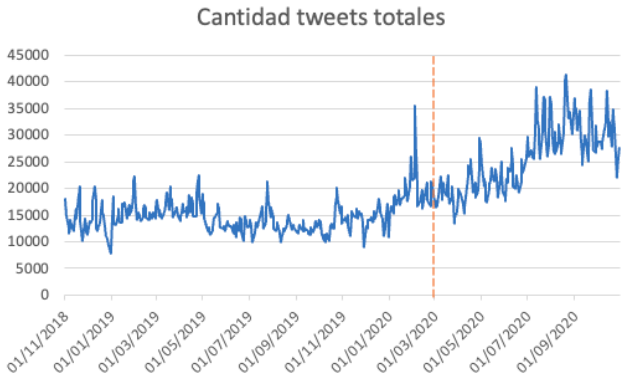


Figura 5.8: Cantidad *tweets* totales

En estos 3 casos se puede ver que hay un aumento en la cantidad de *tweets* luego de la fecha de división, aunque este crecimiento es mayor en la cantidad de positivos y totales. También se puede ver a continuación, a modo de ejemplo, los grafos adaptados para visualización de las fechas: 1 de noviembre de 2018, 10 de marzo de 2020 y 11 de noviembre de 2020.

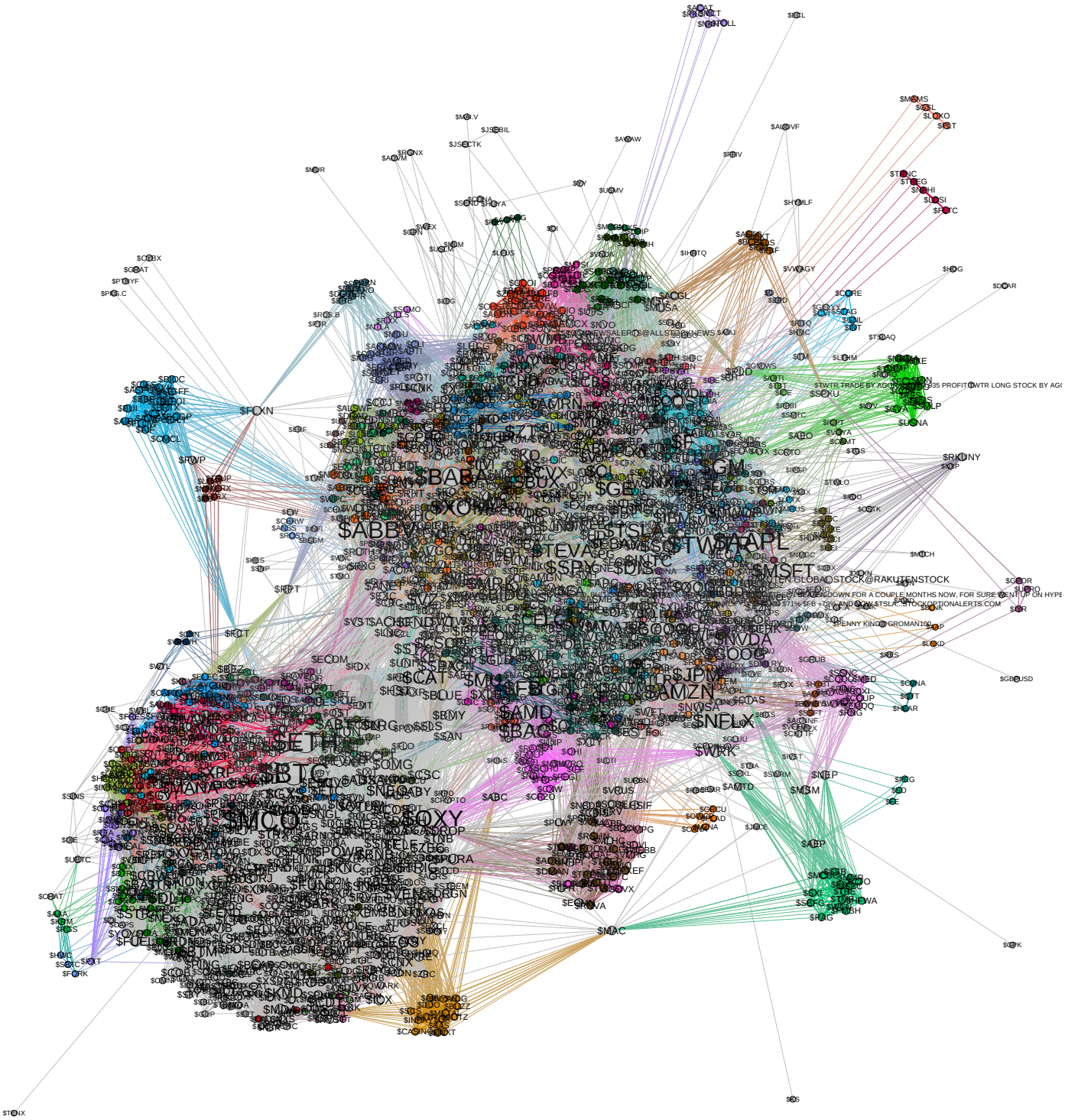


Figura 5.10: Grafo del día 1 de noviembre de 2018

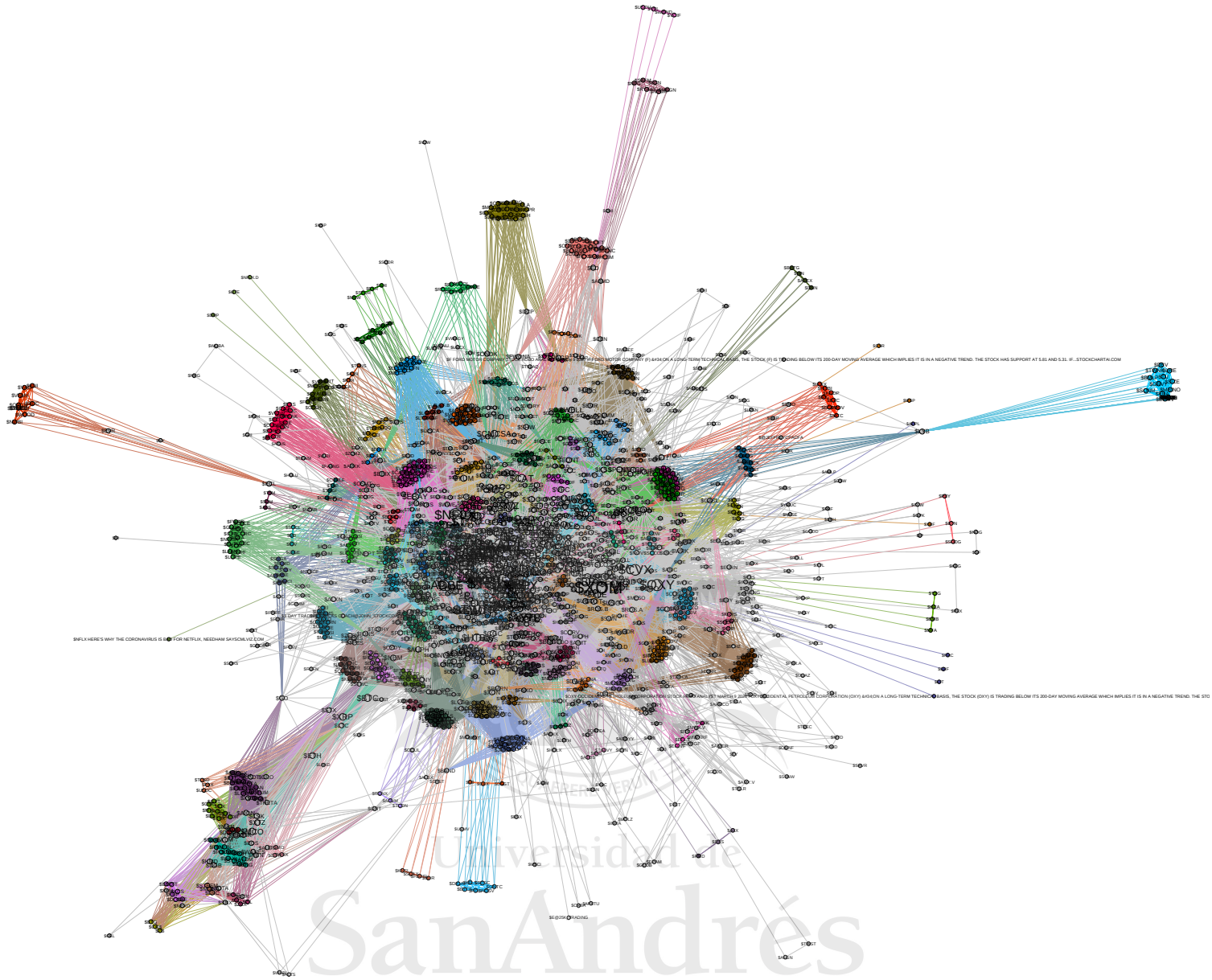


Figura 5.11: Grafo del día 10 de marzo de 2020

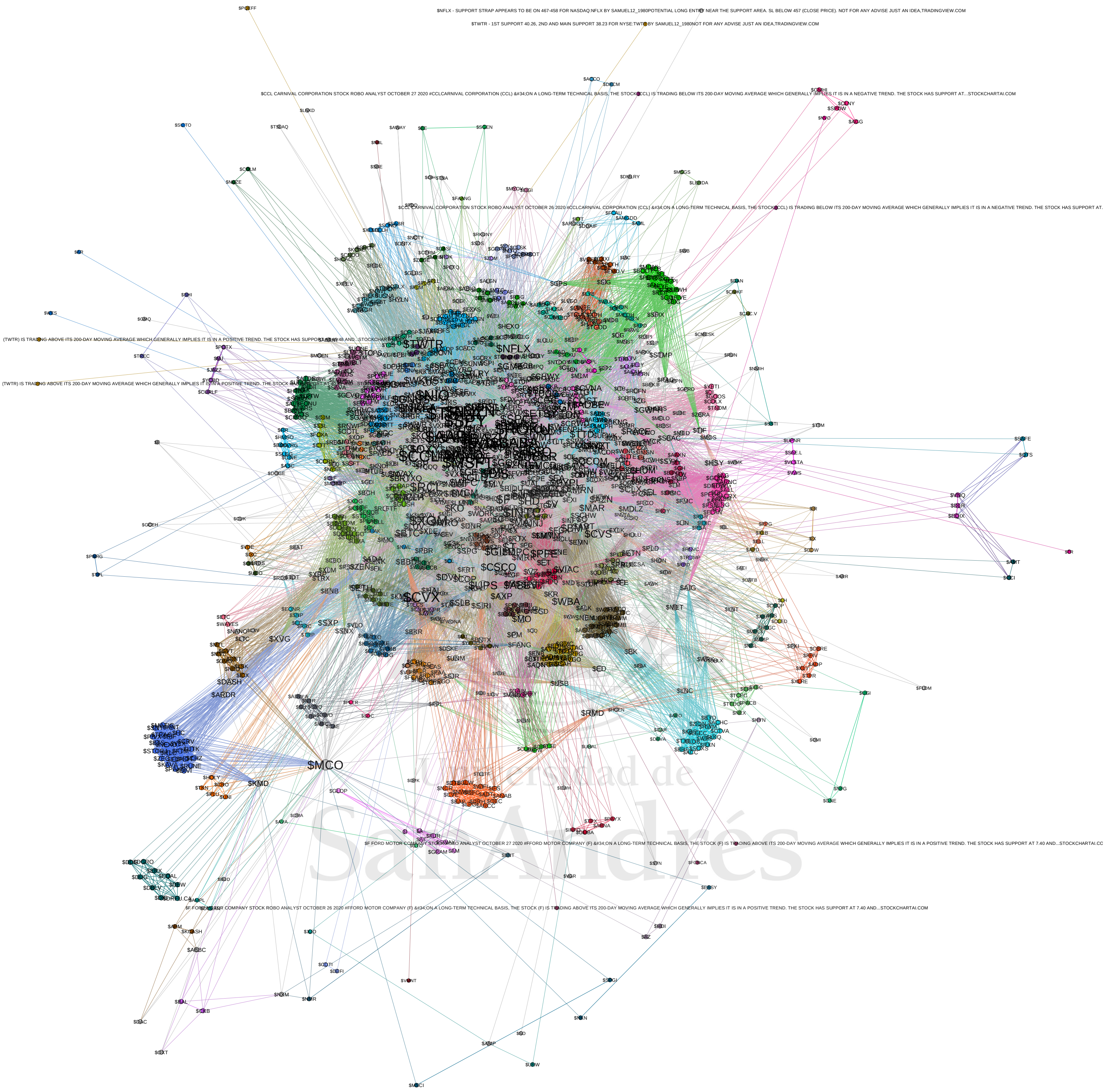


Figura 5.12: Grafo del día 1 de noviembre de 2020

6. Correlación lineal

6.1 Análisis de correlaciones durante el periodo estudiado completo

Luego de obtener y almacenar las estadísticas de las redes diariamente, ya se tiene armada la serie de tiempo de los valores para poder analizar. Se compara contra los retornos y el volumen del índice S&P 500, ya que se está trabajando con redes sociales conformadas por las acciones de ese índice.

Los siguientes son los resultados de correlacionar cada una de las estadísticas de la red respecto a los retornos y el volumen del S&P 500:

Periodo completo (Pearson)	
Contra retornos S&P500	
Estadistica	Correlacion
Numero de nodos	-0,0068
Numero de aristas	0,0105
PPMI promedio	-0,0808
Eficiencia global	0,0405
Centralidad de EV	0,0119
Coef. promedio de clustering	-0,0466
# tweets positivos	0,0254
# tweets negativos	-0,1547
# tweets totales	-0,0100

Figura 6.1.1: Correlación de retornos S&P 500 vs estadísticas

Periodo completo (Pearson)	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	0,1442
Numero de aristas	0,1534
PPMI promedio	-0,0405
Eficiencia global	0,2976
Centralidad de EV	0,2696
Coef. promedio de clustering	-0,2731
# tweets positivos	0,1669
# tweets negativos	0,4093
# tweets totales	0,2768

Figura 6.1.2: Correlación de volumen S&P 500 vs estadísticas

Se puede observar como la correlación respecto a los retornos es baja, por lo tanto, se podría pensar que estas redes sociales (o las estadísticas con las que se trabaja) no tienen relación clara con los retornos.

En cambio, al correlacionar las estadísticas de las redes respecto al volumen operado día a día, se pueden observar mejoras en los resultados. Por este motivo, nuestros modelos y resultados los realizamos con la variable del volumen. La centralidad del vector propio y la eficiencia global tiene una correlación con el volumen de casi 0,3. La cantidad de tweets positivos, negativos y totales tienen correlaciones altas llegando a 0,41. Se puede ver que exceptuando el PPMI promedio, las demás estadísticas tienen correlaciones superiores al 0,15.

6.2 Análisis de la correlación con división en el tiempo

Luego de haber calculado y analizado las correlaciones de las estadísticas de la red a lo largo de todo el periodo de tiempo estudiado -es decir entre el 1 de noviembre de 2018 y el 1 de noviembre de 2020-, se decidió dividirlo en dos distintos segmentos. Por un lado, el lapso entre el 1 de noviembre de 2018 y el 6 de marzo de 2020; por otro lado, del 9 de marzo de 2020 al 1 de noviembre de 2020.

¿Por qué motivo se decidió cortar el primer segmento el 6 de marzo del 2020? La razón, como se anticipó anteriormente, es que fue el viernes previo al comienzo del *crash* bursátil de 2020; que llevó al índice S&P 500 a caer desde los 2.970 puntos a los mínimos de 2.190 tocados el 23 de marzo,

representando una caída de aproximadamente 26% en tan solo 11 ruedas de operaciones. También fue el día donde se conoció que Estados Unidos había llegado al caso 100 de COVID.

A continuación, se presentan los resultados de las correlaciones contra el volumen operado para ambos segmentos del tiempo:

1/11/2018 a 6/3/2020	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	0,239
Numero de aristas	0,212
PPMI promedio	0,123
Eficiencia global	0,186
Centralidad de EV	0,222
Coef. promedio de clustering	-0,253
# tweets positivos	0,177
# tweets negativos	0,342

9/3/2020 a 1/11/2020	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	-0,619
Numero de aristas	-0,628
PPMI promedio	0,420
Eficiencia global	-0,453
Centralidad de EV	-0,626
Coef. promedio de clustering	0,611
# tweets positivos	-0,662
# tweets negativos	-0,118

Figura 6.2.1: Correlación de volumen S&P 500
vs estadísticas con división del tiempo

Se puede observar el gran cambio ocurrido en las correlaciones simplemente al separar la ventana del tiempo en estos dos segmentos, y es interesante intentar pensar las causas de este efecto. Uno de los motivos podría ser que los inversores, al ver un aumento significativo de la volatilidad e incertidumbre del 9 de marzo de 2020 en adelante, hayan recurrido a utilizar aún más la plataforma *Twitter* como una fuente de información para realizar operaciones en el mercado financiero. Las correlaciones son altas, con valores de 0,61 para el coeficiente promedio de clustering; 0,42 para el PPMI Promedio; -0,63 para la centralidad del vector único (EV); y -0,62 y -0,63 para el numero de aristas y numero de nodos, respectivamente. También se puede observar que para el caso de la cantidad de *tweets* positivos tiene una correlación de -0,66.

6.3 Análisis de la serie de tiempo de la correlación lineal

En esta sección se realizó el calculo de la correlación móvil con una ventana temporal de 60 días. Como se mostró en la sección anterior se obtiene una correlación estática de todo el periodo, en cambio de esta manera se puede observar la correlación como una serie de tiempo e intentar de obtener mayor información sobre los datos con los que se está trabajando.

A continuación, se muestran los gráficos que creemos que son más relevantes. Los gráficos de la izquierda muestran en el eje izquierdo la correlación móvil (línea azul) y en el eje derecho la evolución de la estadística a lo largo del tiempo, como puede ser en el primer caso la evolución del numero de nodos.

Los gráficos de la derecha muestran nuevamente en el eje izquierdo la correlación móvil (línea azul), pero en el eje derecho se gráfica la evolución del volumen operado del S&P 500 (en Billions). Nuevamente se agregó la línea vertical para indicar el día 6/3/2020.

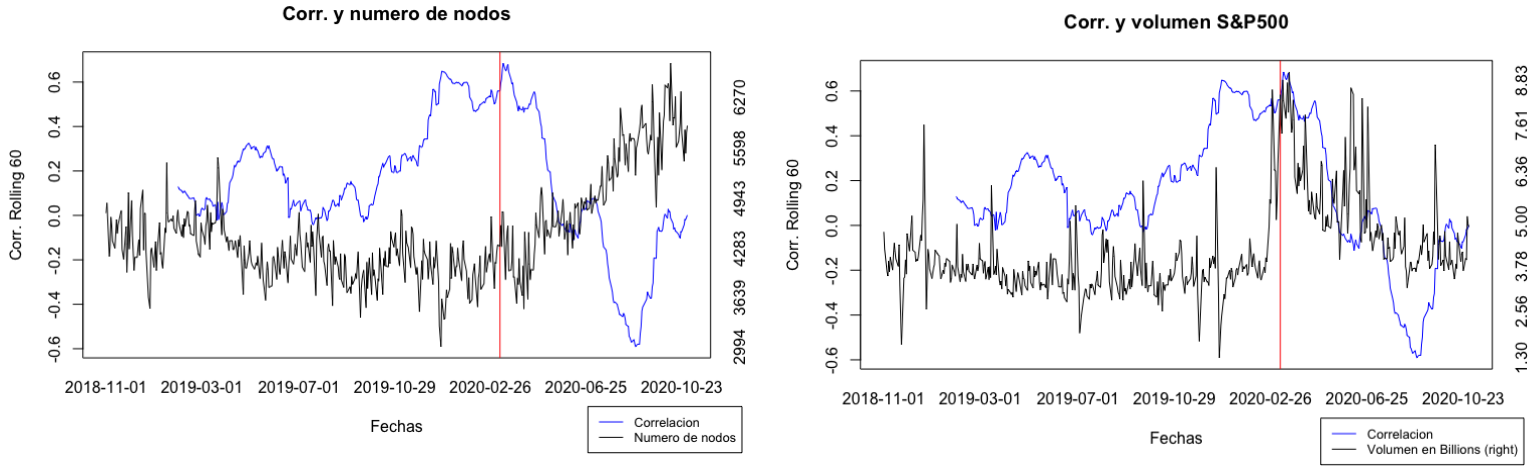


Figura 6.3.1: Correlación móvil de la estadística numero de nodos.

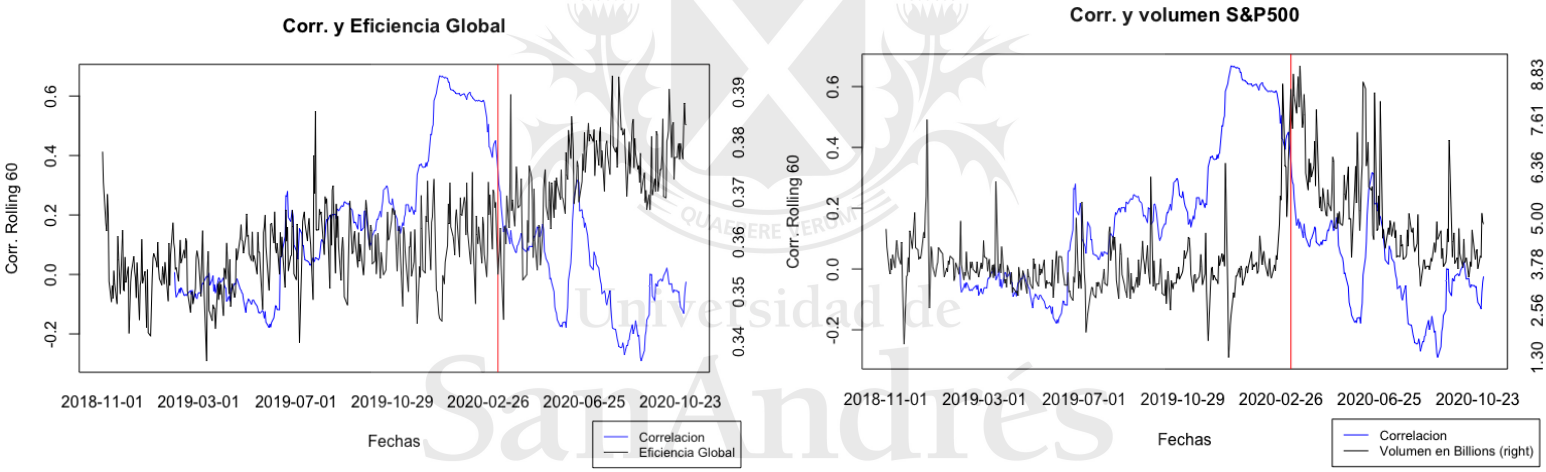


Figura 6.3.2: Correlación móvil de la estadística eficiencia global

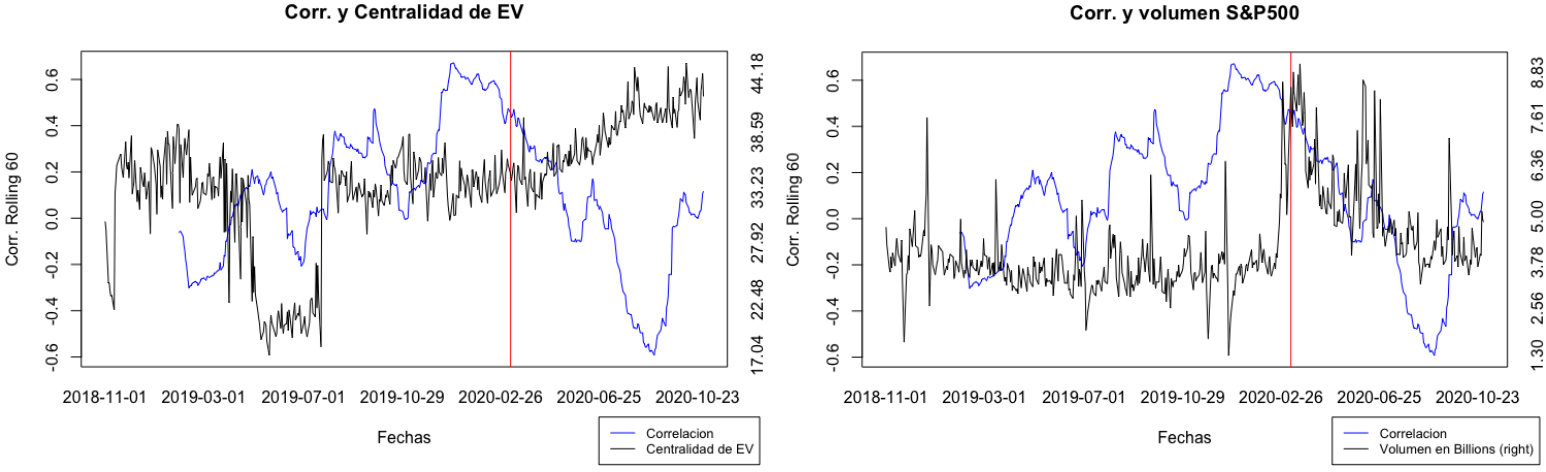
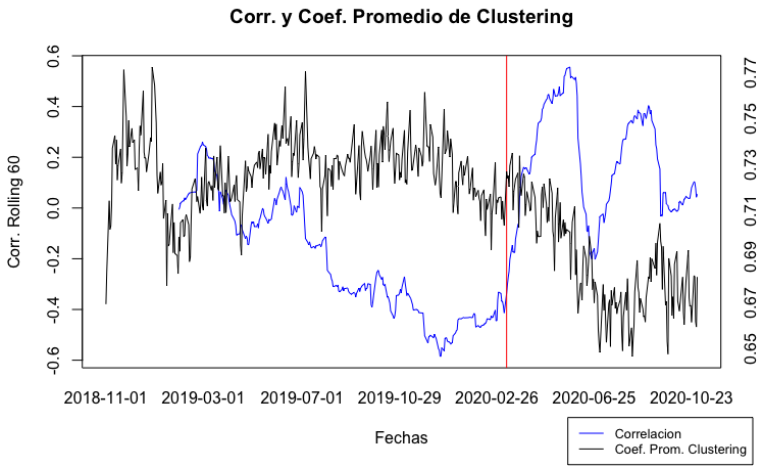


Figura 6.3.3: Correlación móvil de la estadística centralidad de EV



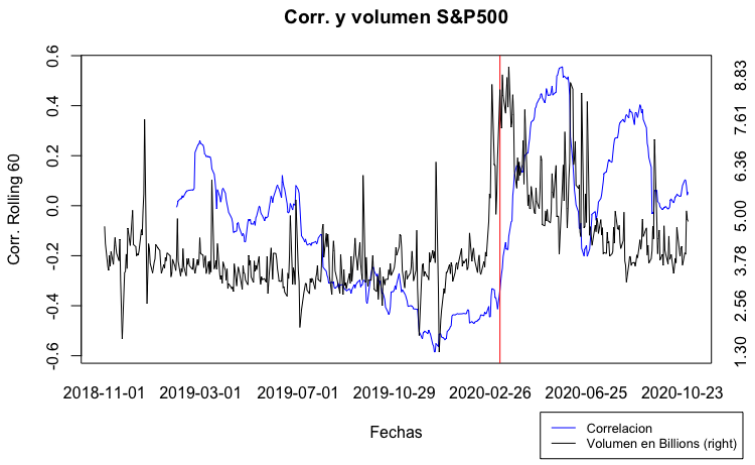
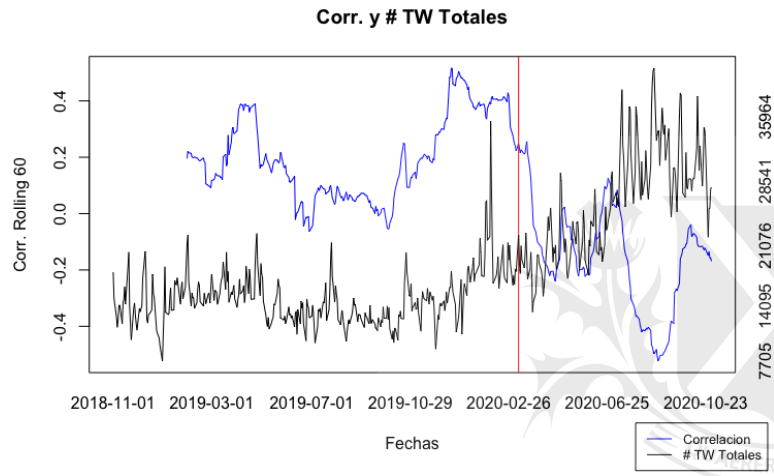


Figura 6.3.4: Correlación móvil de la estadística coeficiente promedio de *clustering*



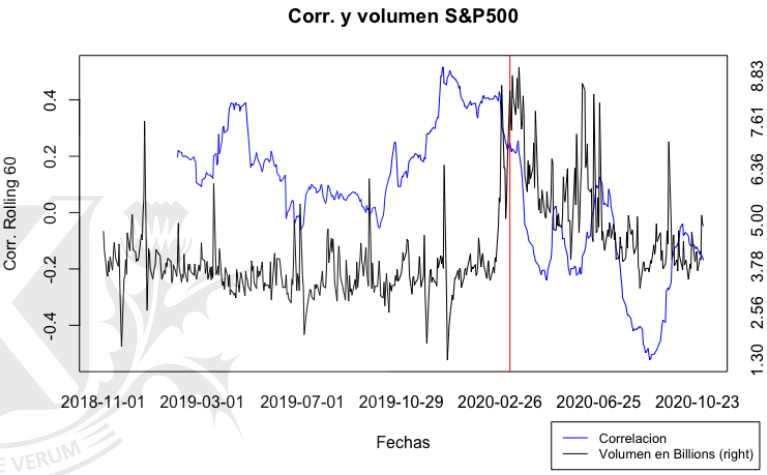


Figura 6.3.5: Correlación móvil de la estadística cantidad de *tweets* totales

7. Correlación no lineal

7.1 Análisis de correlaciones durante el periodo estudiado completo

También se optó por calcular dos medidas alternativas de correlación con el fin de captar la no linealidad de los datos con los que se trabaja. En primer lugar, se calculó el coeficiente de Spearman, que es una medida no paramétrica de la correlación, donde se intenta describir la relación de dos variables a través de una función monótona. A su vez, el coeficiente de Spearman es menos sensible que el de Pearson para valores lejos de los esperado.

El segundo método alternativo que se calculó fue el coeficiente de Kendall, que también es una medida no paramétrica que indica la fuerza de dependencia entre dos variables. A continuación, se muestran los coeficientes de ambos métodos para todas las estadísticas en el periodo completo:

Periodo completo (Spearman)	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	0,3532
Numero de aristas	0,3692
PPMI promedio	-0,1055
Eficiencia global	0,4105
Centralidad de EV	0,4169
Coef. promedio de clustering	-0,4643
# tweets positivos	0,4278
# tweets negativos	0,5225
# tweets totales	0,5145

Periodo completo (Kendall)	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	0,2362
Numero de aristas	0,2452
PPMI promedio	-0,0673
Eficiencia global	0,2721
Centralidad de EV	0,2714
Coef. promedio de clustering	-0,3072
# tweets positivos	0,2784
# tweets negativos	0,3602
# tweets totales	0,3371

Es interesante notar que en todos los casos el coeficiente de Spearman toma un mayor valor que el coeficiente de Kendall y el de Pearson. Aquí se puede observar las diferencias en la tabla comparativa.

Tabla comparativa para periodo completo		
Estadistica	Corr. Pearson	Corr. Spearman
Numero de nodos	0,1442	0,3532
Numero de aristas	0,1534	0,3692
PPMI promedio	-0,0405	-0,1055
Eficiencia global	0,2976	0,4105
Centralidad de EV	0,2696	0,4169
Coef. promedio de clustering	-0,2731	-0,4643
# tweets positivos	0,1669	0,4278
# tweets negativos	0,4093	0,5225
# tweets totales	0,2768	0,5145

Figura 7.1.3: Tabla comparativa de correlación de volumen S&P 500 con método de Pearson y Spearman.

7.2 Análisis de correlaciones con división en el tiempo

Como se mostró en el caso de las correlaciones lineales, también se realizó el calculo con el método de Spearman y Kendall para la división temporal en el día 6/3/2020, con el fin de ver el cambio en las correlaciones. Los resultados son los siguientes:

1/11/2018 a 6/3/2020 (Spearman)	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	0,265
Numero de aristas	0,211
PPMI promedio	0,182
Eficiencia global	0,133
Centralidad de EV	0,278
Coef. promedio de clustering	-0,268
# tweets positivos	0,185
# tweets negativos	0,284
# tweets totales	0,278

9/3/2020 a 1/11/2020 (Spearman)	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	-0,721
Numero de aristas	-0,724
PPMI promedio	0,390
Eficiencia global	-0,463
Centralidad de EV	-0,707
Coef. promedio de clustering	0,608
# tweets positivos	-0,736
# tweets negativos	-0,121
# tweets totales	-0,695

Figura 7.2.1: Correlación de Spearman con división temporal contra volumen S&P500

1/11/2018 a 6/3/2020 (Kendall)	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	0,182
Numero de aristas	0,145
PPMI promedio	0,125
Eficiencia global	0,090
Centralidad de EV	0,183
Coef. promedio de clustering	-0,180
# tweets positivos	0,126
# tweets negativos	0,197
# tweets totales	0,192

9/3/2020 a 1/11/2020 (Kendall)	
Contra volumen S&P500	
Estadistica	Correlacion
Numero de nodos	-0,505
Numero de aristas	-0,501
PPMI promedio	0,266
Eficiencia global	-0,315
Centralidad de EV	-0,493
Coef. promedio de clustering	0,422
# tweets positivos	-0,521
# tweets negativos	-0,081
# tweets totales	-0,487

Figura 7.2.2: Correlación de Kendall con división temporal contra volumen S&P500

Se puede observar que con el método de Spearman las correlaciones siguen siendo mayores en todos los casos que la de Pearson y Kendall, por lo que se podría concluir que está captando de una mejor forma las relaciones entre las variables al ser un método no paramétrico.

7.3 Análisis de la serie de tiempo de la correlación no lineal de Spearman

En esta sección se realizó nuevamente el calculo de la correlación móvil de Spearman con una ventana temporal de 60 días. A continuación, se muestran los gráficos más relevantes.

Los gráficos de la izquierda muestran en el eje izquierdo la correlación móvil de Spearman (línea azul) y en el eje derecho la evolución de la estadística a lo largo del tiempo.

Los gráficos de la derecha muestran nuevamente en el eje izquierdo la correlación móvil de Spearman (línea azul), pero en el eje derecho se gráfica la evolución del volumen operado del S&P 500 (en Billions). Nuevamente se agregó la línea vertical para indicar el día 6/3/2020 que consideramos importante.

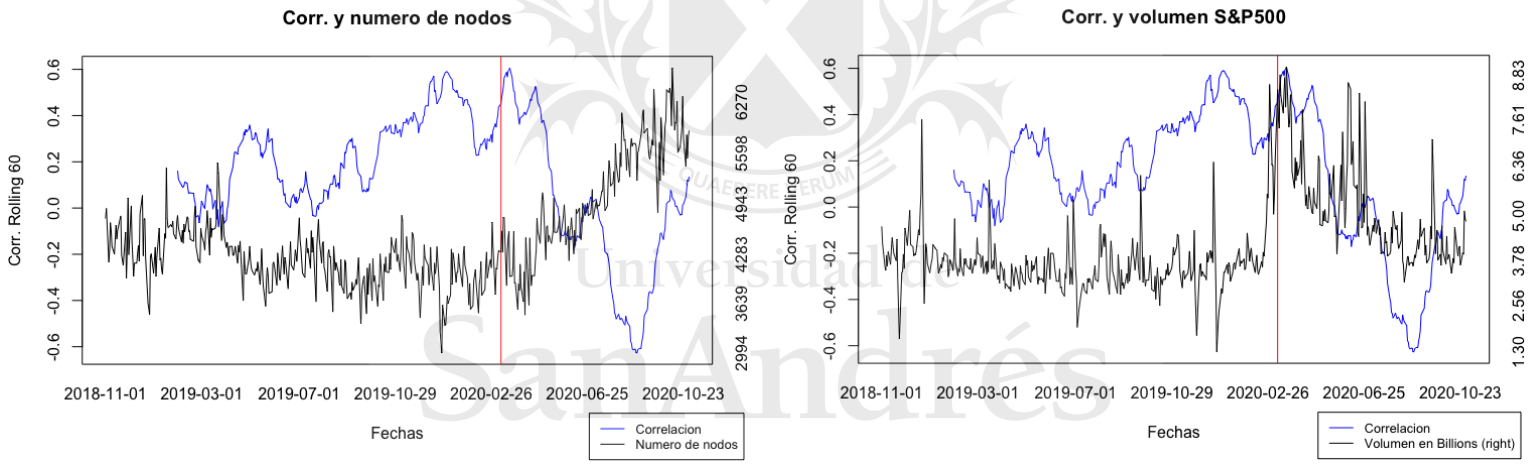


Figura 7.3.1: Correlación móvil de Spearman de la estadística número de nodos.

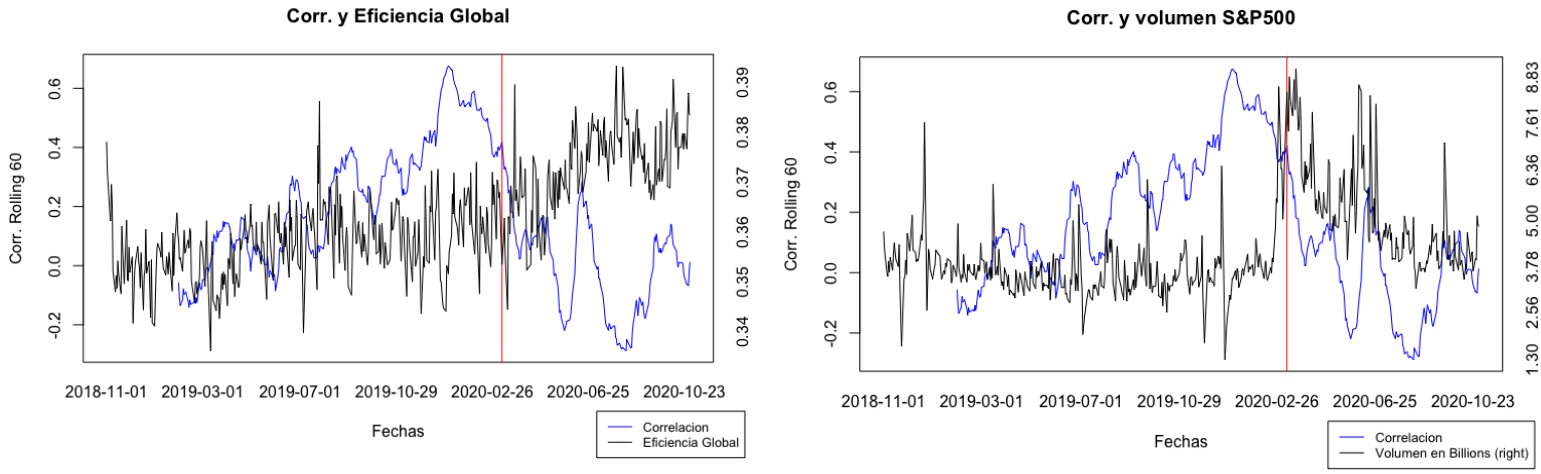


Figura 7.3.2: Correlación móvil de Spearman de la estadística eficiencia global

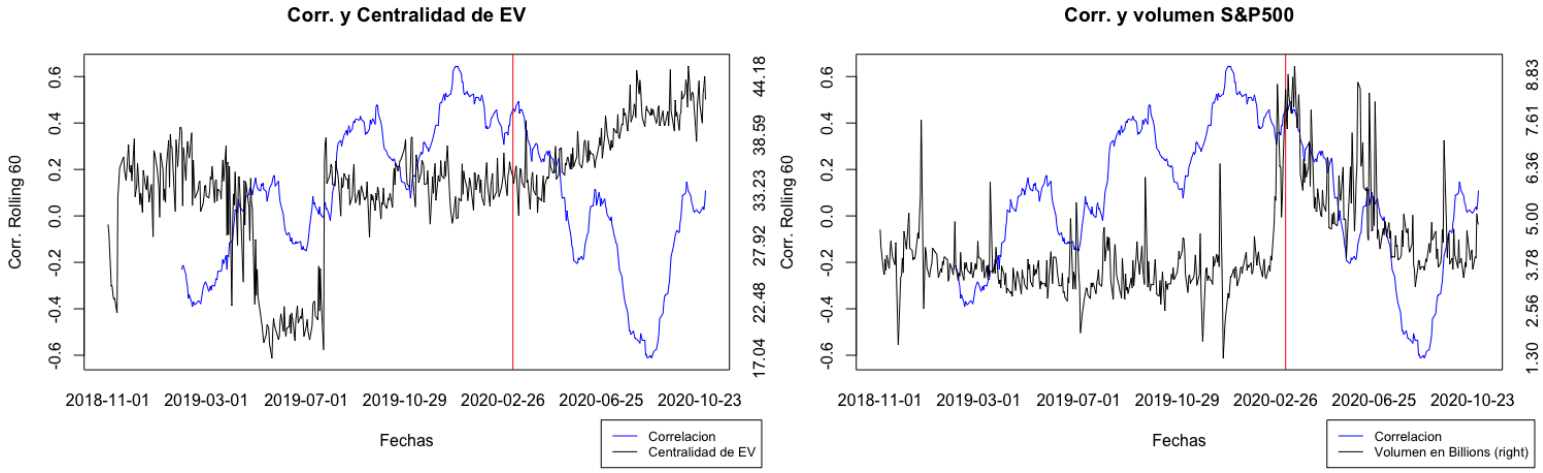


Figura 7.3.3: Correlación móvil de Spearman de la estadística centralidad de EV

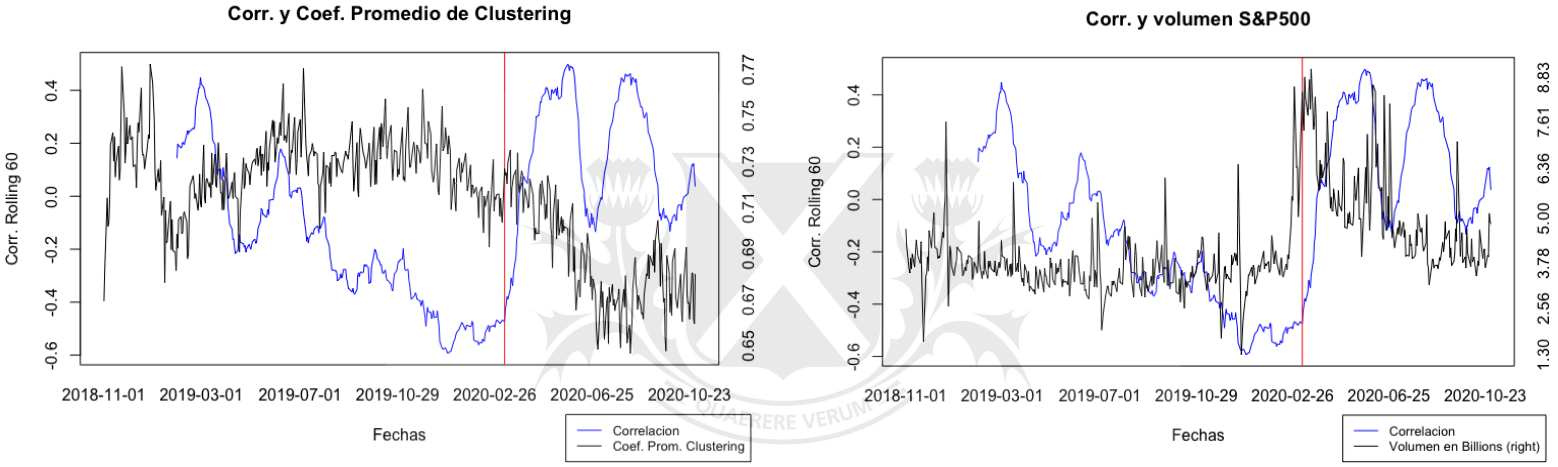


Figura 7.3.4: Correlación móvil de Spearman de la estadística coeficiente promedio de *clustering*

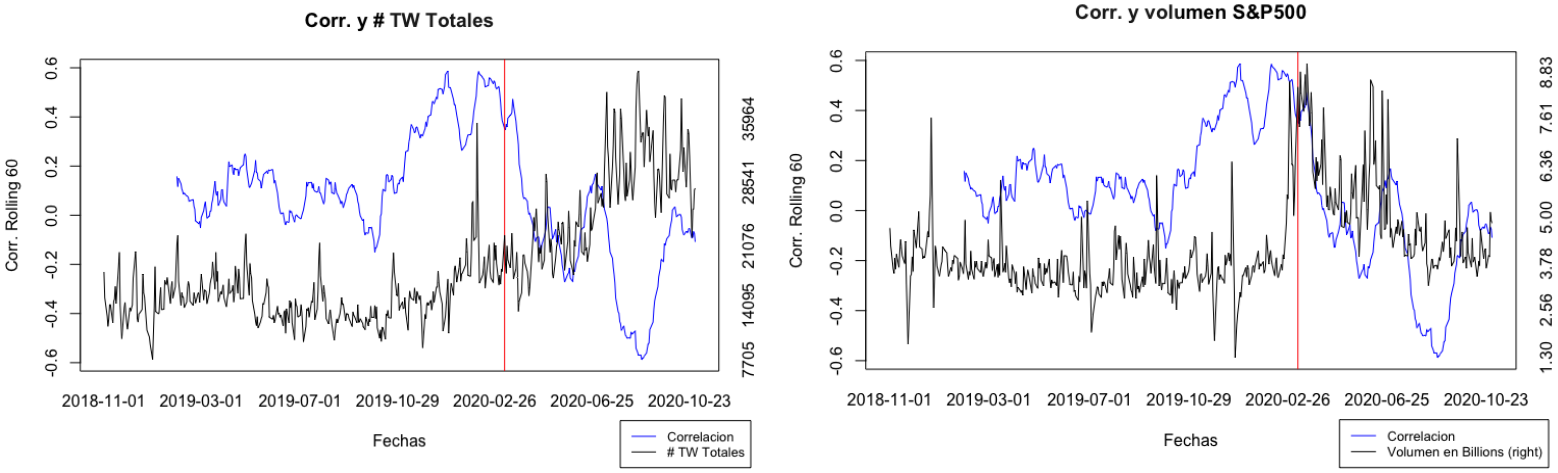


Figura 7.3.5: Correlación móvil de Spearman de la estadística cantidad de *tweets* totales

8. Modelos de predicción

Luego de analizar los datos y sus correlaciones, se observó que, en principio, no habría poder predictivo de las estadísticas de las redes sociales sobre los retornos del índice S&P500. Mientras que, las estadísticas podrían llegar a tener poder predictivo sobre el volumen diario operado en este mismo índice. Es por esto por lo que se decidió realizar diversos modelos predictivos del volumen.

Se implementaron modelos de regresión lineal, modelos autorregresivos integrados de media móvil y redes neuronales.

Cada modelo se implementó en dos particiones de tiempo; La primera, llamada *full sample*, la cual contiene los datos completos de la muestra desde el 01/11/2018 al 01/11/2020. Mientras que, la segunda partición, llamada *PRE-COVID*, contiene los datos de la muestra hasta la fecha en la que se reportaron numerosos casos de una neumonía híper contagiosa en China, es decir desde el 01/11/2018 hasta el 31/12/2019. La partición de los datos entre *train set* y *test set* fue de 80/20 en todos los modelos.

8.1 Multicolinealidad

Al observar la matriz de correlaciones entre todas las variables explicativas de nuestros modelos, se observa que existe una alta correlación entre las mismas.

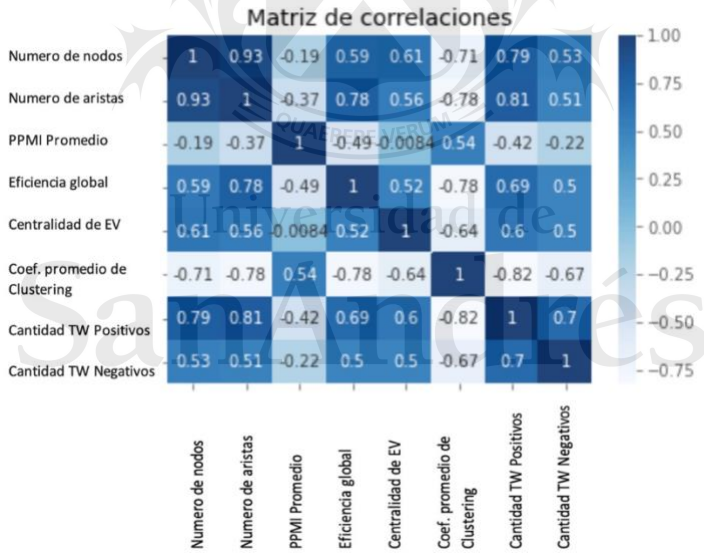


Figura 8.1.1: Matriz de correlaciones entre todas las estadísticas

Para confirmar la existencia de multicolinealidad, se realizó el test de factor de inflación de la varianza (*VIF)*.

Test Variance Inflation Factor (VIF)	
Estadistica	VIF
Numero de nodos	746
Numero de aristas	294
PPMI promedio	948
Eficiencia global	2211
Centralidad de EV	90
Coef. promedio de clustering	2866
# tweets positivos	27
# tweets negativos	21
# tweets totales	25

Figura 8.1.2: Valores del Test VIF

Como se puede ver, se obtuvieron como resultado valores de *VIF* altos, lo que confirma la presencia de multicolinealidad. Ante la presencia de multicolinealidad, en los modelos de regresión lineal y en los modelos autorregresivos integrados de media móvil, optamos por dos alternativas:

1) Se predice la variable dependiente, con una sola variable independiente. Por lo que se genera un modelo por cada variable independiente.
2) Para poder incorporar todas las variables en un mismo modelo y evitar el problema de multicolinealidad, se aplica un análisis de componentes principales (*PCA).*

8.2 PCA

Al aplicar esta técnica, se procesa un extenso conjunto de datos y a su vez se reduce la dimensionalidad con una pérdida mínima de información. Esta técnica, además permite describir un conjunto de datos en términos de nuevas variables no correlacionadas y así solucionar el problema de la multicolinealidad entre las variables explicativas de nuestro modelo.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Numero de nodos	0,37	0,40	-0,20	0,37	0,31	-0,41	0,40	0,33
Numero de aristas	0,24	0,02	0,81	-0,21	0,45	0,13	0,01	0,15
PPMI promedio	-0,43	-0,48	-0,08	-0,17	0,11	-0,14	0,19	0,70
Eficiencia global	0,36	0,21	0,09	-0,36	-0,70	0,14	0,19	0,38
Centralidad de EV	-0,28	0,09	0,40	0,75	-0,33	0,13	-0,11	0,23
Clustering promedio	-0,17	-0,20	0,21	0,03	-0,15	-0,03	0,83	-0,41
# tweets positivos	0,06	0,04	-0,29	0,12	0,25	0,87	0,24	0,13
# tweets negativos	-0,62	0,72	0,04	-0,29	0,08	0,01	0,07	0,02

Figura 8.2.1: Aporte de cada variable a cada componente en el modelo PCA

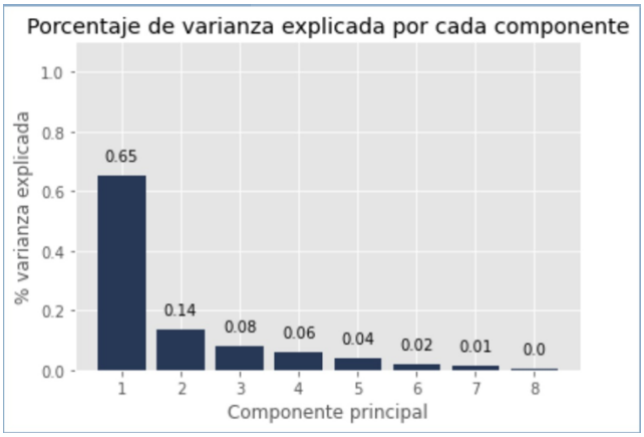


Figura 8.2.2: Porcentaje de varianza explicada por cada componente

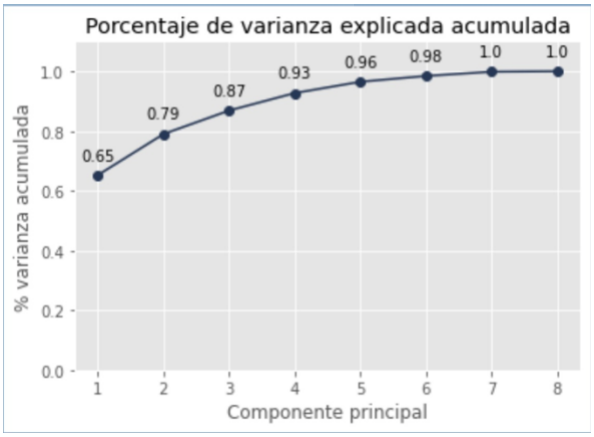


Figura 8.2.3: Porcentaje de varianza acumulada explicada por cada componente

Se optó por utilizar hasta el componente número tres, captando así el 87% de la varianza acumulada en los datos. Al ver la matriz de correlaciones de los componentes, se concluye que están perfectamente no correlacionados, por lo que el problema de la multicolinealidad fue solucionado.

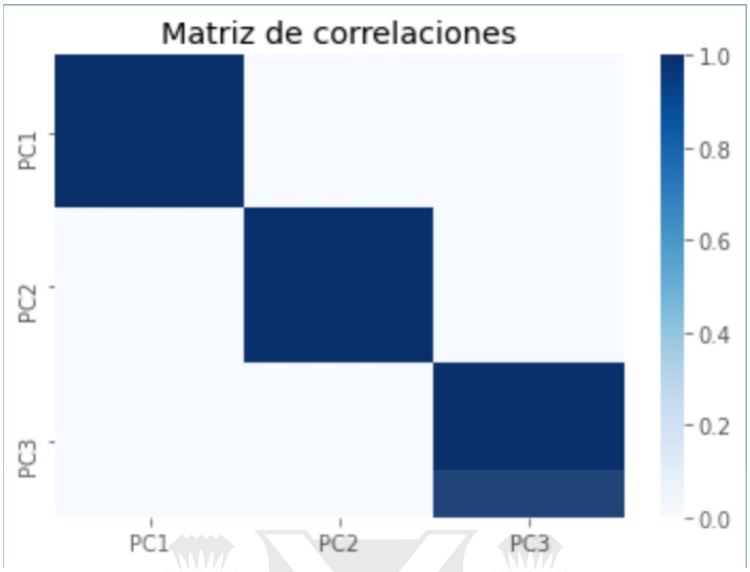


Figura 8.2.4: Matriz de correlaciones entre los 3 componentes seleccionados

9. Regresiones Lineales

En primer lugar, se implementaron regresiones lineales simples *full sample* entre cada variable explicativa con un lag (las estadísticas de las redes sociales y los tres componentes (*PCA)*) y la variable independiente (logaritmo natural del volumen diario operado en el índice *S&P500*):

$$\log(Volumen_t) = \alpha_0 + \beta_1 \mathrm{X}_{t-1} + \in$$

Reg FULL SAMPLE				
Estadistica	**Lags**	**MAPE**	**R2**	
Numero de nodos	1	27,18%	7,34%	***
Numero de aristas	1	37,65%	9,14%	***
PPMI promedio	1	11,88%	-0,25%	
Eficiencia global	1	21,60%	14,30%	***
Centralidad de EV	1	14,49%	10,42%	***
Coef. promedio de clustering	1	21,79%	24,51%	***
# tweets positivos	1	69,00%	15,75%	***
# tweets negativos	1	24,25%	26,28%	***
# tweets totales	1	64,69%	27,86%	***
PCA (3 componentes)	1	70,22%	33,38%	***,*,***

Figura 9.1: Resultados de regresiones lineales *full sample* (sin *dummy*) con 1 lag cada variable

Luego se implementaron las mismas regresiones lineales simples, pero para la partición de datos *PRE-COVID*.

Reg PRE COVID				
Estadistica	**Lags**	**MAPE**	**R2**	
Numero de nodos	1	13,79%	5,62%	***
Numero de aristas	1	14,65%	2,00%	*
PPMI promedio	1	17,00%	4,55%	***
Eficiencia global	1	16,84%	-0,36%	
Centralidad de EV	1	17,44%	5,66%	***
Coef. promedio de clustering	1	15,12%	2,78%	**
# tweets positivos	1	16,08%	0,04%	
# tweets negativos	1	15,74%	3,28%	**
# tweets totales	1	15,58%	3,16%	**
PCA (3 componentes)	1	15,22%	9,39%	**,***,

Figura 9.2: Resultados de regresiones lineales *PRE-COVID* con 1 lag cada variable

Luego, se le agregó a estas regresiones lineales simples *full sample* una variable *dummy* que toma valores de uno para fechas posteriores al día del caso 100 de *COVID* en Estados Unidos y valores de 0 para el complemento:

$$\log(Volumen_t) = \alpha_0 + \beta_1 \mathrm{X}_{\mathrm{t}-1} + \beta_2 Dummy_{1,0} + \in$$

Reg FULL SAMPLE (con dummy)				
Estadistica	**Lags**	**MAPE**	**R2**	
SOLO DUMMY		43,92%	50,07%	***
Numero de nodos	1	65,64%	52,57%	***,***
Numero de aristas	1	66,06%	51,86%	***,***
PPMI promedio	1	40,23%	51,10%	**,***
Eficiencia global	1	50,38%	51,17%	**,***
Centralidad de EV	1	49,87%	52,14%	***,***
Coef. promedio de clustering	1	57,53%	52,07%	***,***
# tweets positivos	1	61,43%	50,79%	**,***
# tweets negativos	1	51,68%	55,21%	***,***
# tweets totales	1	64,97%	52,74%	***,***
PCA (3 componentes)	1	72,93%	56,45%	***,***,*,***

Figura 9.3: Resultados de regresiones lineales *full sample* (con *dummy*) con 1 lag cada variable

A continuación, se realizaron los mismos modelos de regresiones lineales simples pero en vez de utilizar un lag arbitrario en cada variable explicativa, se utilizó el *lag* óptimo de cada variable explicativa. Para la elección de los *lags* óptimos de cada variable explicativa se utilizó el modelo de regresión progresiva, *forward stepwise*. A continuación, se muestran los resultados del siguiente modelo tanto para el modelo *full sample* como el *PRE-COVID*:

$$\log(Volumen_t) = \alpha_0 + \beta_1 \mathrm{X}_{\mathrm{t-lag\,óptimo}} + \in$$

Reg FULL SAMPLE				
Estadistica	**Lags**	**MAPE**	**R2**	
Numero de nodos	1,5	21,86%	3,67%	**,
Numero de aristas	1	37,65%	9,14%	***
PPMI promedio	5	11,46%	0,35%	
Eficiencia global	1	21,60%	14,30%	***
Centralidad de EV	2	13,99%	8,25%	***
Coef. promedio de clustering	4	22,86%	7,84%	***
# tweets positivos	1	69,00%	15,75%	***
# tweets negativos	1	24,25%	26,28%	***
# tweets totales	1	64,69%	27,86%	***
PCA (3 componentes)	1	70,22%	33,38%	***,*,***

Figura 9.4: Resultados de regresiones lineales *full sample* (sin *dummy*) con lag óptimo de cada variable

Reg PRE COVID				
Estadistica	**Lags**	**MAPE**	**R2**	
Numero de nodos	1,5	14,930%	2,72%	*,
Numero de aristas	1	14,65%	2,17%	*
PPMI promedio	5	17,210%	2,20%	*
Eficiencia global	1	16,84%	-0,36%	
Centralidad de EV	2	17,710%	5,28%	***
Coef. promedio de clustering	4	16,540%	-0,44%	
# tweets positivos	1	16,08%	0,04%	
# tweets negativos	1	15,74%	3,28%	**
# tweets totales	1	15,58%	3,16%	**
PCA (3 componentes)	1	15,22%	9,39%	**,***,

Figura 9.5: Resultados de regresiones lineales *PRE-COVID* con lag óptimo de cada variable

A continuación, se presentan los resultados del modelo con *lags* óptimos y la variable dummy:

$$\log(Volumen_t) = \alpha_0 + \beta_1 \mathrm{X}_{\,\mathrm{t-lag\ óptimo}} + \beta_2 Dummy_{1,0} + \in$$

Reg FULL SAMPLE (con dummy)				
Estadistica	**Lags**	**MAPE**	**R2**	
SOLO DUMMY		43,92%	50,07%	***
Numero de nodos	1,5	54,85%	50,77%	*,,***
Numero de aristas	1	66,06%	51,86%	***,***
PPMI promedio	5	42,54%	50,32%	,***
Eficiencia global	1	50,38%	51,17%	**,***
Centralidad de EV	2	48,32%	51,44%	**,***
Coef. promedio de clustering	4	48,08%	51,29%	**,***
# tweets positivos	1	61,43%	50,79%	**,***
# tweets negativos	1	51,68%	55,21%	***,***
# tweets totales	1	64,97%	52,74%	***,***
PCA (3 componentes)	1	72,93%	56,45%	***,***,*,***

Figura 9.6: Resultados de regresiones lineales *full sample (*con *dummy)* con lag óptimo de cada variable

Se observa que para la partición *full sample*, los *MAPEs* son en general altos tanto en el modelo con 1 *lag* arbitrario cómo en el modelo con *lags* óptimos. Los *lags* óptimos contribuyen a mejorar el *MAPE* en algunos casos.

También se puede ver que tanto las regresiones de la estadística Centralidad de EV, como la de coeficiente promedio de *clustering*, tienen como resultado *MAPEs* aceptables junto con *R2* relativamente altos. Permitiendo inferir parcialmente, que ambas podrían tener poder de predicción sobre el logaritmo natural del volumen.

En la partición *PRE-COVID* todos los *MAPEs* mejoran a niveles aceptables con respecto de la partición *full sample*. Esto era esperable ya que el COVID generó una mayor volatilidad en los mercados e incertidumbre. Es lógico que en el periodo *PRE-COVID* los errores sean menores al querer intentar predecir los valores futuros del volumen operado.

De los modelos de regresión lineal *full sample*, el mejor modelo en términos de *MAPE* y significatividad fue el siguiente, obteniendo un *MAPE* de 13,99% y una significativad de 8,25%.

$$\log(Volumen_t) = \alpha_0 + \beta_1 \text{Centralidad de EV}_{t-2} + \in$$

Mientras que, de los modelos de regresión lineal *PRE-COVID*, el mejor modelo en términos de *MAPE* y significatividad fue el siguiente, con valores de 13,79% y 5,62%, respectivamente.

$$\log(Volumen_t) = \alpha_0 + \beta_1 \text{Número de nodos}_{t-1} + \in$$

10. Modelo Autorregresivo Integrado de Media Movil (ARIMA)

Para los modelos ARIMA también se decidió realizar el modelo con dos particiones temporales de los datos. La primera llamada *full sample* donde se utilizaron todos los datos disponibles, y la segunda *PRE-COVID* donde se utilizó hasta el día 31/12/2019.

Dentro de los modelos *full sample* se estudiaron dos alternativas. La primera es el modelo $ARIMA(1,1,1)$ para cada estadística, donde se corrieron 11 modelos distintos sobre el *train set.* Un modelo por cada estadística como variable regresora y uno donde únicamente se dejó el efecto del volumen. La variable a predecir en todos los modelos fue el $log(Volumen)$. Las estadísticas se incluyeron en todos los modelos con 1 lag, para de esta manera utilizar información de t_{-1} para predecir el volumen del momento t. La segunda alternativa es similar a la anterior, con la única diferencia que en cada uno de los 11 modelos se decidió incluir como variable regresora adicional una *dummy*, que toma valor de 0 hasta el día 6/3/2020 y luego toma valor 1 hasta el final de la muestra en el día 1/11/2020.

Una vez que se entrenaron los modelos se realizó la predicción sobre el *test set* y se calcularon los *MAPEs* aplicando la función exponencial a los valores observados y predichos para obtener el MAPE en nivel y poder compararlos contra los modelos de regresión lineal y de red neuronal. Los resultados se pueden observar a continuación:

FULL SAMPLE (SIN DUMMY) - Modelo(1,1,1)	
Estadistica	**MAPE**
SOLO VOLUMEN	51,45%
Numero de nodos	60,64%
Numero de aristas	59,12%
PPMI promedio	51,36%
Eficiencia global	52,29%
Centralidad de EV	53,35%
Coef. promedio de clustering	51,47%
# tweets positivos	55,00%
# tweets negativos	53,70%
# tweets totales	58,71%
PCA (3 componentes)	59,21%

Figura 10.1: *MAPEs* del modelo ARIMA(1,1,1) *full sample* (sin variable *dummy*)

FULL SAMPLE (CON DUMMY) - Modelo(1,1,1)	
Estadistica	**MAPE**
SOLO VOLUMEN	47,40%
Numero de nodos	55,91%
Numero de aristas	55,14%
PPMI promedio	47,38%
Eficiencia global	48,79%
Centralidad de EV	49,34%
Coef. promedio de clustering	47,61%
# tweets positivos	50,56%
# tweets negativos	49,01%
# tweets totales	53,99%
PCA (3 componentes)	55,46%

Figura 10.2: *MAPEs* del modelo ARIMA(1,1,1) *full sample* (con variable *dummy*)

Para el modelo PRE-COVID se realizó el mismo procedimiento que en la primera alternativa explicada anteriormente. Se generaron 11 modelos distintos donde cada uno contenía como variable regresora a cada estadística en t_{-1} (con 1 lag). Los resultados fueron los siguientes:

PRE COVID - Modelo(1,1,1)	
Estadistica	**MAPE**
SOLO VOLUMEN	15,38%
Numero de nodos	15,08%
Numero de aristas	15,19%
PPMI promedio	15,79%
Eficiencia global	15,26%
Centralidad de EV	15,42%
Coef. promedio de clustering	15,83%
# tweets positivos	15,44%
# tweets negativos	15,41%
# tweets totales	15,33%
PCA (3 componentes)	15,75%

Figura 10.3: *MAPEs* del modelo ARIMA(1,1,1) *PRE-COVID*

Como se puede observar en los *MAPEs* obtenidos hay una clara diferencia en los errores del modelo *full sample* y el *PRE-COVID*. Mientras que en el primero los *MAPEs* obtenidos varían entre 47% y 60%, en el caso de los modelos *PRE-COVID* varían entre 15% y 16%. Esto era esperable ya que,

como se comentó anteriormente, el COVID generó una mayor volatilidad en los mercados, un mayor ruido, y por ende un mayor error al querer intentar predecir los valores futuros del volumen operado.

Al igual que lo realizado en las regresiones lineales se corrieron nuevamente estos mismos modelos, pero en lugar de utilizar un *lag* para cada variable de forma arbitraria, se utilizaron los *lags* óptimos obtenidos por el método descripto anteriormente. Los resultados se pueden ver a continuación:

FULL SAMPLE (SIN DUMMY) - Modelo(1,1,1) - Lags optimos		
Estadistica	**Lags**	**MAPE**
SOLO VOLUMEN		51,45%
Numero de nodos	1,5	66,82%
Numero de aristas	1	59,12%
PPMI promedio	5	51,56%
Eficiencia global	1	52,29%
Centralidad de EV	2	51,42%
Coef. promedio de clustering	4	47,59%
# tweets positivos	1	55,00%
# tweets negativos	1	53,70%
# tweets totales	1	58,71%
PCA (3 componentes)	1	59,21%

Figura 10.4: *MAPEs* del modelo ARIMA(1,1,1) *full sample* (sin variable *dummy*) y con *lags* óptimos.

FULL SAMPLE (CON DUMMY) - Modelo(1,1,1) - Lags optimos		
Estadistica	**Lags**	**MAPE**
SOLO VOLUMEN		47,40%
Numero de nodos	1,5	62,12%
Numero de aristas	1	55,14%
PPMI promedio	5	47,30%
Eficiencia global	1	48,79%
Centralidad de EV	2	47,58%
Coef. promedio de clustering	4	43,93%
# tweets positivos	1	50,56%
# tweets negativos	1	49,01%
# tweets totales	1	53,99%
PCA (3 componentes)	1	55,46%

Figura 10.5: *MAPEs* del modelo ARIMA(1,1,1) *full sample* (con variable dummy) y con *lags* óptimos.

PRE COVID - Modelo(1,1,1) - Lags optimos		
Estadistica	**Lags**	**MAPE**
SOLO VOLUMEN		15,38%
Numero de nodos	1,5	15,07%
Numero de aristas	1	15,19%
PPMI promedio	5	15,52%
Eficiencia global	1	15,26%
Centralidad de EV	2	15,61%
Coef. promedio de clustering	4	15,57%
# tweets positivos	1	15,44%
# tweets negativos	1	15,41%
# tweets totales	1	15,33%
PCA (3 componentes)	1	15,75%

Figura 10.6: *MAPEs* del modelo ARIMA(1,1,1) *PRE-COVID* y con *lags* óptimos.

Se puede observar que los modelos *full sample* tienen una mejora en términos del *MAPE* al utilizar los *lags* óptimos. Mientras que en el *full sample* (sin *dummy*) antes se obtenía un *MAPE* del 51,36% en el mejor modelo, cuando se utilizan los *lags* óptimos el mejor modelo (con 4 *lags* de la estadística coeficiente de *clustering* promedio) consigue reducir el *MAPE* a 47,59%.

De la misma forma vemos que en el mejor modelo *full sample* (con *dummy*) antes se obtenía un *MAPE* de 47,38%, mientras que en el modelo con los *lags* óptimos (donde nuevamente el mejor es el que utiliza la estadística coeficiente de *clustering* promedio) el *MAPE* se reduce a 43,93%.

En cambio, en el modelo *PRE-COVID* no se observa una mejora significativa en los nuevos modelos.

11. Redes Neuronales

11.1 Estructura del modelo

Luego de implementar los modelos de regresión lineal y modelos autorregresivos integrados de media móvil en todas las combinaciones ya descriptas anteriormente, se implementaron modelos de redes neuronales con el objetivo de predecir el volumen diario operado del índice S&P500. A través de estos modelos de redes neuronales, se buscó captar la no linealidad de los datos que, dada la naturaleza de los modelos de predicción implementados anteriormente, no fue captada hasta el momento.

La variable dependiente del modelo es el volumen diario del índice S&P500. Las variables explicativas son las siguientes:

- Número de nodos en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Número de aristas en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- PPMI promedio en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Grado promedio (*Average degree)* en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Eficiencia local en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Eficiencia global en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Centralidad de vector propio (*Eigenvector centrality)* en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- *Betweenness centrality* en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Coeficiente de agrupamiento promedio (*Average Clustering coefficient)* en $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Volumen $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Cantidad de tweets positivos $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Cantidad de tweets negativos $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$
- Cantidad de tweets totales $t_{-1};\ t_{-2}; t_{-3},; t_{-4};\ t_{-5}$

Al analizar la estructura de los datos, se optó por implementar una red neuronal recurrente con las siguientes características:

- *Long short-term memory*,
- Función de activación *Rectified Linear Unit (ReLU)*
- Algoritmo de optimización *Adam*
- Función de pérdida *Mean Squared Error*
- 100 *EPOCHS*
- Tamaño de *batch* 32

Al igual que en los modelos anteriores, se implementó la red en dos particiones de tiempo. La primera llamada *full sample*, que contiene los datos completos de la muestra desde el 01/11/2018 al 01/11/2020. Y la segunda llamada *PRE-COVID*, que contiene los datos de la muestra desde el 01/11/2018 hasta el 31/12/2019.

También se decidió intentar predecir el volumen tanto en nivel como en logaritmo. Por este motivo, se utilizaron los siguientes cuatro modelos finales:

- Red neuronal *Full Sample* con la variable volumen a nivel.
- Red neuronal *Full Sample* con el logaritmo natural del volumen.
- Red neuronal *PRE-COVID* con la variable volumen a nivel.
- Red neuronal *PRE-COVID* con el logaritmo natural del volumen.

11.2 Resultados del modelo

El *MAPE* de cada modelo se obtuvo a través de la siguiente metodología. Primero se corrió cada modelo un total de cincuenta veces, por lo que se obtuvo una serie de cincuenta *MAPEs*. A la serie de *MAPEs* posteriormente se le calculó el promedio, estimando así el *MAPE* para cada modelo.

Modelo	MAPE
Full Sample - Volumen en nivel	16,12%
Full Sample - log(Volumen)	15,63%
Pre Covid - Volumen en nivel	15,10%
Pre Covid - log(Volumen)	11,98%

Se observa que los modelos tienen un *MAPE* aceptable y, en el caso del periodo *PRE-COVID*, un modelo mejor en cuanto a reducción del *MAPE* al compararlo con los modelos de regresión lineal y autorregresivos integrados de media móvil descriptos anteriormente.

Además, en promedio, el *MAPE* de los modelos *PRE-COVID* son menores que los de los modelos *Full Sample*. Como ya se viene explicando en los modelos anteriores, esto era esperable en todos los casos.

12. Conclusión

Al observar los modelos realizados a lo largo de este trabajo, se puede afirmar que para la partición *PRE-COVID*, el mejor modelo para predecir el volumen con las variables explicativas seleccionadas en términos de *MAPE* es la red neuronal recurrente implementada, con un *MAPE* de 11,98%.

Mientras que, para la partición *full sample*, el mejor modelo en términos de *MAPE* es la siguiente regresión lineal, con un *MAPE* de 13,99% y una significatividad de 8,25%:

$$\log(Volumen_t) = \alpha_0 + \beta_1 \text{Centralidad de EV}_{t-2} + \in$$

A pesar de estos resultados, en términos generales, la red neuronal fue el modelo que mejor respondió ante las dos particiones. En principio, esto era lo esperado debido a la naturaleza no lineal de los datos.

Luego de realizar este trabajo exploratorio, podríamos afirmar que las redes sociales diarias de menciones conjuntas de acciones del S&P500 en *tweets* contienen información que permite predecir el volumen diario operado en el S&P500. A su vez, de estos *tweets* que constituyen las redes sociales, se obtiene información relevante como el *sentiment*, que también contribuye a explicar el volumen diario operado.

Si bien este fue un trabajo exploratorio, brinda indicios de que el contenido informativo de los *tweets* es relevante para los mercados financieros. En futuros trabajos se podría profundizar en este análisis exploratorio, ya sea implementando nuevos modelos predictivos o analizando los datos a nivel nodo, viendo sí es posible utilizar la información de estas redes sociales para explicar y/o predecir el movimiento conjunto de los activos que la componen.

13. Bibliografía

- Agrawal, Shreyash, Azar, Pablo D, Lo, Andrew W, & Singh, Taranjit. (2018). *Momentum, Mean-Reversion, and Social Media: Evidence from StockTwits and Twitter*. Journal of Portfolio Management, 44(7), 85-95.
- Fama, E. (1970). *Efficient Capital Markets: A Review of Theory and Empirical Work*. The Journal of Finance, 25(2), 383-417. doi:10.2307/2325486
- Gálvez, Ramiro H., Gravano, Agustín (2017). *Assessing the usefulness of online message board mining in automatic stock prediction systems*. Journal of Computational Science, Volume 19.
- Gross, Christian, & Siklos, Pierre. (2018). 78 / July 2018. *Analyzing credit risk transmission to the non-financial sector in Europe: a network approach*. European Systemic Risk Board
- Pagolu, Venkata Sasank, Challa, Kamal Nayan Reddy, Panda, Ganapati, & Majhi, Babita. (2016). *Sentiment Analysis of Twitter Data for Predicting Stock Market Movements*. International conference on Signal Processing, Communication, Power and Embedded System (SCOPES).
- Rosenberg, B., Reid, K. and Lanstein, R. (1985). *Persuasive Evidence of Market Inefficiency*. Journal of Portfolio Management, 11, 9-17.
- Ruiz et al., V. Hristidis, C. Castillo, A. Gionis, A. Jaimes (2012). *Correlating financial time series with micro-blogging activity*. Proceeding of ACM international conference on web search and data mining (WSDM), Seattle, WA.
- Sharma, Charu, & Habib, Amber. (2019). *Mutual information-based stock networks and portfolio selection for intraday traders using high frequency data: An Indian market case study*. PloS One, 14(8), E0221910.
- Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., Li, P.: *User-level sentiment analysis incorporating social networks*. In: Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD (2011)