UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

EFECTO DE LAS NOTICIAS EN EL MERCADO FINANCIERO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MATÍAS IGNACIO MAYER ABETT DE LA TORRE

PROFESORA GUÍA:
MARCELA VALENZUELA BRAVO

MIEMBROS DE LA COMISIÓN:
ALEJANDRO BERNALES SILVA
PATRICIO VALENZUELA AROS

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: MATÍAS MAYER ABETT DE LA TORRE
FECHA: 05/08/2020
PROFESORA GUÍA: SRA. MARCELA VALENZUELA

EFECTO DE LAS NOTICAS EN EL MERCADO FINANCIERO

Desde el inicio de los mercados financieros, una de los principales desafíos ha sido predecir o estimar cual será el comportamiento de los inversionistas frente a la toma de decisiones. El supuesto formulado es que las personas que participan en el mercado, se informan financieramente, ya que lo que ellos buscan es maximizar su beneficio personal. En respuesta al estudio de la "mentalidad" o raciocinio de los inversores se genera la rama de Behavioral Finance, con la que se intenta explicar lo anterior. La gran interrogante asociada a esto es la forma en que los medios de comunicación afectan en la toma de decisiones financieras. Por esta razón, se utilizó una base de datos generada con anterioridad, la que guardaba las noticias publicadas diariamente por el New York Times desde el año 1851 hasta el 2018 y utilizando las categorías de Calomiris and Mamaysky (C-M), se crearon diferentes indicadores como proxis a la cantidad de información en la noticia, basados en los diferentes tópicos, como noticias de crédito, noticias de gobierno, noticias de mercado, entre otras. Por otro parte, se determinó el sentimiento de mercado (market sentiment) en base a los proxis ya estudiados anteriormente en trabajos del ámbito económico. En este caso se utilizaron las noticias ya procesadas del New York times para calcular los índices de sentimiento para luego estudiar la existencia de una relación positiva con las variables de interés.

Se estudió el efecto que tenía la variable de market sentiment sobre dos variables que capturan el comportamiento macroeconómico del mercado. En este caso se propuso la producción industrial estadounidense en un primer estudio y el producto interno bruto real en el segundo, como las variables dependientes. Como regresores se utilizó la ya mencionada sentiment y, después se prosiguió a agregar otras de control como la inflación, por ejemplo, para observar si el efecto cambiaba. El resultado fue entregado sugiere que la variable de interés si tiene un efecto, significativo y positivo sobre la producción, con y sin variables de control, lo que se puede explicar por las teorías de la economía conductual y el efecto del market sentiment sobre los retornos y el sentimiento del inversor. En el caso del PIB, el efecto también es positivo para ambos casos y aquí también entran el juego las relaciones estudiadas en la literatura económica del market sentiment sobre la inversión y la liquidez del mercado.

No importa la lentitud con la que
avances, siempre y cuando
no te detengas.
Confucio

Agradecimientos

Darle las gracias a todas las personas que fueron parte de este proceso, amigos, familiares, cuerpos docentes.

Primero darle gracias a mi profesora guía Marcela Valenzuela, por darme la gran oportunidad y confianza de trabajar con ella. Creo firmemente que este proceso me ayudo mucho a crecer como profesional, a poner en práctica los conocimientos adquiridos a través de mi carrera y desarrollar nuevas habilidades. Gracias por transmitirme su experiencia, su apoyo y por su guía, ya que esto fue clave para desarrollar este trabajo.

Mi familia como pilar fundamental en esta aventura, manteniendo su apoyo constante durante mi formación y siempre tratando de entregarme las herramientas mas adecuadas para poder enfrentarme de la mejor forma a la vida. Creo que una parte importante de este logro va dirigido a ustedes.

A mi pareja por tenerme paciencia y ser un apoyo siempre cuando tuve momentos difíciles. Mis amigos de la vida también fueron parte importante de este proceso, Felipe, Nicolás, Kenjin, Vicente y Tomás, se vendrán nuevos desafíos.

A todos los grandes amigos que hice durante esta increíble etapa, llena de altos y bajos. Fueron un gran soporte en los momentos críticos de la carrera. Los de Gamanchester, los amigos de la gran sección 8 y todos los que fueron apareciendo en otras situaciones.

TABLA DE CONTENIDO

Indice de tablas

Indice de ilustraciones

1.Introducción

Cada día la humanidad es capaz de encontrar grandes avances científicos en todos los ámbitos, inclusive en las finanzas. Los nuevos algoritmos de tradings, nuevos estudios sobre los retornos, entre otras cosas, producen que esta área se vuelva mucho más interesante y competitiva. El problema es que aún no se ha podido llegar a un método infalible para estimar el efecto de las noticias y la forma en que se cuenta esta en los medios a los inversionistas a la hora de tomar decisiones financieras, que afectan los directamente al mercado de activos.

Actualmente, las facilidades para tener la información del mundo han crecido abismalmente, debido a las diversas herramientas tecnológicas que permiten tener acceso a las noticias o premisas de cualquier índole, lo que permite reducir la brecha de la llamada "información asimétrica", es decir, se espera que esta cobertura ayude a generar un mercado financiero más eficiente

Gracias a esto es que se ha generado un mayor interés en cómo afectan los medios de comunicación en la toma de decisiones de los actores en el mercado financiero y de qué manera afecta esta información en los movimientos de estos macroeconómicamente. Esto se ha plasmado a través de una amplia literatura que los flujos de noticias afectan los mercados financieros (ver, por ejemplo, Tetlock, 2007; García, 2013; Calomiris and Mamaysky,2017).

Una teoría que se maneja para explicar este fenómeno es que el contenido de las noticias y la forma en que se plasma el suceso, afecta el sentimiento de mercado de los inversionistas. Este efecto hace alusión a la actitud que los inversionistas toman frente a algún instrumento financiero o al mercado de activos, lo que estaría muy ligado al área que estudia las "finanzas conductuales" y los sesgos que estos generan. Por otra parte, con este sentimiento se hace referencia al estado mental, emoción o actitud que se genera en un mercado, o la llamada psicología de masas que se puede apreciar en la forma en que se comportan los movimientos de un mercado, ya sea en el volumen transado o el precio de los activos financieros dentro de este. En la literatura de los últimos años no se ha logrado definir un patrón claro, ya que si bien algunos han llegado al resultado de que las noticias financieras negativas afectan más este fenómeno, provocando que los inversionistas actúen en base a ese sentimiento, lo que permitiría predecir las fluctuaciones del mercado, otros autores hablan de que depende mucho de la cantidad de información, del lenguaje, de la serie de tiempo seleccionada, llegando en algunos casos a resultados diferentes y que esta variable tendría efecto para la parte positiva. Tetlock (2007) llega al resultado de que la información y la forma en que los

diarios entregan la información entre 1984 hasta 1999, pueden prever las fluctuaciones en el mercado financiero.

El sentimiento financiero es un tópico que en los últimos años ha generado mucho interés en diversos autores como un factor clave para explicar el comportamiento de los inversores, los movimientos en el mercado de activos y el crecimiento económico. Por ejemplo, algunos sugieren que el sentimiento de mercado es una variable fundamental en la recuperación de la economía ya que esta no será convincente hasta que haya una clara mejora en el espíritu de los consumidores. En otros estudios, Junyan Shena, Jianfeng Yub, Shen Zhao (2017) llegan a la conclusión de que es teóricamente factible que el sentimiento pueda estar relacionado con la aversión al riesgo efectiva y, por lo tanto, con el precio del riesgo, por lo que su interpretación basada en el sentimiento es consistente con la historia de aversión al riesgo que varía con el tiempo.

Para probar la relación entre el desempeño del mercado de valores y el crecimiento de la producción, varios estudios encuentran evidencia para apoyar una relación positiva. Levine y Zervos (1998) utilizan un modelo de regresión a través del país para examinar varios países para el período 1976-1993. Muestran que la liquidez del mercado de renta variable está correlacionada de manera positiva y sólida con las tasas de crecimiento económico, de acumulación de capital y de productividad contemporáneas y futuras. El estudio de Thomas C. Chiang (2017), concluye decisivamente que el desempeño del mercado de valores, ya sea medido por los rendimientos del mercado de valores o el riesgo a la baja, presenta una causa unidireccional del mercado de valores al crecimiento de la producción industrial

Con respecto al efecto de las noticias sobre el “sentiment”, los 'choques noticiosos', es decir, los casos en que las representaciones noticiosas de las condiciones económicas se mueven inesperadamente en relación con los datos económicos entrantes, son importantes para explicar las fluctuaciones a corto plazo en el sentimiento del consumidor, lo que representa 1/3 de la variación de su pronóstico error en un horizonte de 6-12 meses. (Martha Starr, 2010). En el caso de la producción industrial, Oh y Waldman (1990) descubrieron que los datos de "anuncios falsos" sobre los indicadores principales que fueron revisados posteriormente, y que pueden interpretarse como choques de expectativas, explican más del 20% de las fluctuaciones en el crecimiento de este indicador con una frecuencia trimestral.

Un importante estudio realizado por Carroll, Fuhrer y Wilcox (1998) encontró que, después de controlar los determinantes fundamentales del gasto, el sentimiento aún

tiene un valor pequeño pero significativo para predecir cambios futuros en el gasto. Por otra parte, Ho y Hung (2009) muestran que la incorporación del sentimiento de los inversores en el modelado de la dinámica de las exposiciones al riesgo mejora el poder explicativo de los modelos de fijación de precios de activos para la rentabilidad de las acciones.

En la literatura financiera, también ha surgido interés con estudiar el efecto del sentimiento de mercado y componentes de la economía conductual en indicadores financieros como el consumo, la inversión, la inflación, la liquidez entre otros. Por ejemplo, Dashan Huang, 2014 propone un nuevo índice de confianza del inversor alineado para predecir el mercado de existencias agregadas, basados en los estudios de Baker y Wurgler (2006) en conjunto con el método financiero de Kelly y Pruitt (2013). Con esta nueva medida, llegan a la conclusión de que el sentimiento de los inversores tiene un poder predictivo mucho mayor para el mercado de valores agregado de lo que se pensaba anteriormente. En otros estudios, Longstaff (2004) propone que, si los cambios en la composición de las carteras de los inversores están relacionados con temores de la liquidez de mercado de valores, se puede observar una fuga hacia la liquidez.

En el caso de Kaul y Kayacetin (2009), ellos estudian dos medidas de flujos del mercado de valores en el periodo 1988 a 2004. El resultado que obtienen es que ambas tienen un alto poder de predicción sobre las tasas de crecimiento futuro de la producción industrial y el PIB real. Además, Levine (1991), llega a la conclusión de que un mercado bursátil liquido ayuda a impulsar los proyectos de alto rendimiento y de esta manera estimular el crecimiento de las ganancias y la productividad. Naes (2010), a través de su estudio sobre la liquidez, dice que este indicador es un predictor efectivo y robusto del crecimiento del PIB real, el desempleo y el crecimiento de la inversión. Finalmente, los hallazgos de Florackis (2014) respaldan una relación negativa entre la falta de liquidez en el mercado de valores y el crecimiento futuro del PIB. Bhimjee (2016), por otro lado, llega a que un aumento fundamental en el sentimiento del mercado después de la crisis aumenta la demanda de fondos prestables, lo que aumenta el nivel total de inversión bajo un alto monitoreo

El primer modelo desarrollado en este trabajo se centra en la posible relación entre el sentimiento de mercado y la producción industrial, a través de las noticias relacionadas con economía y finanzas. Luego se dividen las noticias en las 5 dimensiones de Calomiris y Mamaysky (2018), para observar si alguna tiene un mayor efecto sobre el crecimiento

de la producción industrial. Estas categorías son crédito, commodities, gobierno, mercado y corporaciones.

En la siguiente sección, se estudia la relación entre la variable de market sentiment y el crecimiento del producto interno bruto de los Estados Unidos, ya que dado lo mencionado anteriormente, podría haber un efecto implícito de la variable de interés sobre el PIB, a través de otras variables como la inversión o consumo.

En el estudio, como la primera etapa consistió en la extracción de todas las noticias publicadas en el New York times desde 1851 a 2018. A través de distintas técnicas de procesamiento de datos, se crea un base de datos con todas las noticias publicadas diariamente, la que es el input clavea la hora del cálculo del “sentimiento de mercado”. Se utilizan solo las noticias relacionadas con el área financiera en base al diccionario de L-M y la variable lo que plasma es la tendencia de la noticia, es decir, si es más positiva o negativa.

2. Datos

2.1.Fuentes de Información

Para este trabajo, los datos utilizados fueron las noticias del New York Times, para el periodo comprendido entre 1851 y 2018. La descarga se realizó desde la API del diario mencionado y cada noticia posee la siguiente estructura[1]:

- *Id*: ID único interno del *New York Times (desde 1851)*
- *Abstract*: Resumen de la noticia (desde 1851)
- *Blog*: Contiene la palabra *blog* si la noticia proviene del *blog* del *New York Times* (siempre vacío)
- *Byline*: Autor de la noticia si está disponible (desde 1851)
- *Document_type*: Contiene la palabra *article* si es un artículo de noticia, *blogspot* si es un blog y *media* si es un video o diapositivas (desde 1851)
- *Headline*: Titular de la noticia (desde 1851)
- *Locations*: Ubicaciones geográficas (como ciudades) mencionadas en la noticia (desde 1851)
- *Subjects*: Temas asociados a las noticias (desde 1851)
- *Lead_Paragraph*: Primer párrafo de la noticia (desde 1851)
- *Multimedia*: Si la noticia contiene imágenes este atributo guarda la url de la imagen (desde 2002)
- *News_desk*: Departamento al que pertenece la noticia (desde 1980)
- *Print_page*: Número de la página donde aparece la noticia si es que apareció en la versión impresa del *New York Times (desde 1851)*
- *Date*: Fecha de publicación del artículo (desde 1851)
- *Section_name*: Sección a la que pertenece el artículo (desde 1980)

- *Slideshow_credits*: Autor, solo si el atributo *Document_type* es *blog* o *media* (siempre vacío)
- *Snippet*: Texto que se muestra cuando se busca el artículo en la página web del New York Times (desde 1851)
- *Source*: Fuente del artículo, New York Times si es del diario (desde 1851)
- *Subsection_name*: Subsección del artículo (desde 1980)
- *Type*: Campo vacío (desde 1851)
- *Word_count*: Cantidad de palabras en el artículo (desde 1851)
- *Url*: Url del New York Times para acceder al artículo (desde 1851)

En este estudio nuevamente los datos obtenidos son del New York Times por dos razones: La primera por la importancia del diario en Estados Unidos (es el periódico con mayores emisiones, llegando a los 2 millones) y la segunda fue para corroborar o refutar un estudio anterior de similares características en la metodología de obtención de datos. Además, el sitio web de la empresa posee una arquitectura de fácil acceso a la hora de querer obtener la información para el estudio (mediante la API de la plataforma). A continuación, se definen las APIs del sitio:

- *Archive API*: permite obtener toda la meta-data para los artículos del *New York Times* en un determinado mes.
- *Article Search API*: se utiliza para buscar artículos del *New York Times.*
- *Books API*: contiene la lista de los libros *Best Sellers* y críticas por libro.
- *Community API*: corresponde a los comentarios hechos por usuarios en la página del *New York Times.*
- *Geo API*: permite extender la lista de ubicaciones de las noticias usando una BBDD propietaria del *New York Times.*
- *Most Popular API*: contiene la lista de los artículo más compartidos, vistos o enviados por mail.
- *Movie Reviews API*: se utiliza para buscar críticas de películas.
- *Semantic API*: permite obtener términos semánticos de los artículos (personas, lugares, organizaciones y ubicaciones).

- *Times Tags API*: corresponde a los términos que controlan los algoritmos de búsqueda de la metada del *New York Times.*

- *Times Wire API*: permite descargar en tiempo real los artículos del *New York Times* a medida que se van publicando.

- *Top Stories API*: se utiliza para descargar los principales artículos de una sección específica.

Debido a que para este estudio se necesitan las noticias publicadas por este periódico, se utilizó la “Archive API”, ya que era esta la que permitía obtener las publicaciones mensuales del diario.

2.2.Descarga de Datos

En la obtención de los datos, en este caso, las noticias del New York Times, se utilizó la ya mencionada Archive Api. Como este proceso ya había sido realizado, lo que se llevó a cabo en este estudio fue una corroboración, es decir, se replicó un método ya realizado y compararon los resultados con los obtenidos anteriormente. A diferencia de la investigación relacionada, en este se desarrollaron los scripts en el programa Rstudio.

El primer paso en esta metodología es crear una llave, con la que se realizaran las peticiones a l API del New York Times. La página https://developes.nytimes.com permite obtener una llave, pero es necesario crear un usuario y contraseña, la que tiene una limitación ya que solo se pueden realizar 1000 solicitudes por día, lo que llevaría a que el tiempo de descarga fuese muy alto, porque cada llamada permite descargar 10 artículos y son 2.263.925 el total. La solución empleada fue crear mails temporales y de esta manera se obtuvieron 100 llaves. De esta forma se optimiza el tiempo empleado para realizar las descargas de la información. Para realizar esta tarea de forma óptima, se creó un archivo .csv, el que contenía email registrado, contraseña y la API genera. De esta forma las consultas a la página se lograron automatizar.

Con el proceso anterior completo, se inició la descarga de los datos desde la plataforma del new york times. Dada la cantidad de archivos, se desarrolló un script en Rstudio, que permitía descargar las noticias utilizando la Archive API de la página web. La labor que lleva a cabo el programa es la realización de una solicitud a la URL de la API:

http://api.nytimes.com/svc/archive/v1/{año}/{mes}.json?api-key={llave}

Los argumentos clave de la de este método son año, mes y llave, ya que son los que varían en cada solicitud que se le realiza a la API. Lo que la URL retorna es un objeto del tipo JSON, que son datos livianos y en un formato fácil de trabajar. Para cada noticia se tendrán las etiquetas mencionadas en el capítulo anterior y luego a través de comandos del programa, serán guardados en un archivo “txt”, de manare mensual (ejemplo; 1851-09).

Un punto importante que se mencionó con anterioridad era la limitante por día que se presentaba por parte de la API a la que se le realizaba la consulta por lo que, a través del script, se implementó una función que permitía continuar la consulta inmediatamente después de que una de las APIs terminara de descargar los datos. El problema que generaba esto era un posible baneo por parte de la plataforma, al estar sobrepasando el límite de consultas que se podían realizar desde la misma IP del computador, por lo que se utilizó el programa Tor, a través del que se podía modificar la IP en cada iteración a través de una red de computadores conectados a éste, por lo que era más difícil que el servidor rastreara la IP en cada instancia.

2.3.Desarrollo del Dataset

Como se describió en el punto anterior, los archivos son obtenidos en formato JSON, por lo que se debió crear un código que permitiera transformar los datos antes de poder procesarlos. El script fue desarrollado en “Rstudio”, el que tomaba estos archivos JSON y los transformaba en un Dataframe, que es una tabla con los nombres de cada atributo en la cabecera de las columnas y tantas filas como noticias haya ese año, las que guardadas luego como archivos .csv.

En términos simples, el programa recibe el archivo JSON, leyéndolo por año y mes, aprovechando que cada cadena tiene una separación dentro del JSON, para luego guardarlo en una tabla cada uno de los atributos que leyó. Al terminar de leer los datos, la tabla recibe todas las llaves y sus respectivos datos asociados. Este proceso se repite para cada mes del año, por lo que luego de esto se crea una función “merge”, para poder unir estas tablas y crear un dataframe más ordenado. Para terminar el proceso, se guarda cada año con las noticias ordenadas por mes y con todos los atributos disponibles en un archivo csv, que es un tipo de archivo simple para seguir procesando más adelante.

2.4.Limpieza del Dataset

Luego de completar la descarga de los archivos desde la API del new york times, el siguiente paso fue revisar y limpiar la base de datos que contenía los archivos csv. Dado que el estudio está centrado en el área financiera, dentro de esta limpieza se eliminaron los datos que no estuviesen relacionados con ésta. Para llevar a cabo este filtro, dentro

de las columnas que tenía cada observación se encontraba presente *section_name*, la que en la mayoría de los casos contenía la información de la categoría a la que esa noticia pertenecía.

A partir de los datos procesados, es decir, luego de leer los JSON, transformarlos a un formato que se pueda trabajar ("parsearlos") y generar los archivos csv, se procedió a "limpiar" los datos, ya que estos podían tener problemas de formato o no ser útiles en el análisis que se deseaba realizar. Las noticias descargadas desde la plataforma del New York Times, contenían todas las categorías existentes, es decir, no solamente la sección de economía y finanzas, por lo que el resultado entregado por el análisis tendría posibles problemas de sesgo o no ser conclusivo, ya que mezclaría noticias de todo que no tienen efecto tan directo sobre la problemática que se buscaba abordar. Dentro del dataframe generado por cada año de noticias, en cada fila se encontraba un variable llamada *section_name*, la que permitía distinguir a que sección pertenecía cada artículo. El obstáculo que se presentó fue que solo se tenía este atributo desde 1981 hasta el 2018, el resto de los datos no presentaba alguna identificación la que permitiese categorizarlos en alguno de los tópicos que se tenían. Esto fue un primer obstáculo, ya que para este estudio solo se utilizarían las noticias relacionadas al área económica, por lo que fue fundamental desarrollar una herramienta para solucionar esta problemática. Dado que este trabajo se basa en uno anterior, se emplearon técnicas de machine learning ya testeadas, las que servían para resolver este inconveniente.

Se tenían tres algoritmos distintos para tratar el problema de la clasificación. Es posible utilizar esta clase de técnicas, ya que la idea principal era desarrollar una herramienta que pudiese aprender sobre las noticias ya etiquetadas, para luego poder clasificar al resto. Un aspecto clave que permitió el uso de estas técnicas en este estudio fue que al reducir la dimensionalidad de los datos (en este caso clasificando por noticias de negocios y con la técnica de PCA) sobre un plano de dos dimensiones, se pudo apreciar que las categorías eran más o menos disjuntas, es decir, se representaban en áreas distintas, por lo que se esperaba que tuviese un buen rendimiento al clasificar

Dado que se tenía un set con una cantidad adecuada de información (desde 1981 a 2019), fue posible emplear estos algoritmos. Dado que esto fue realizado con anterioridad en un experimento similar, se utilizaron las conclusiones de este, para evitar tener que realizar testeos de algoritmos y ver cual tenía una mejor performance. En base a los resultados obtenidos, luego de entrenar, testear y validar los métodos propuestos (Redes Neuronales, Método del Gradiente y Regresión Logística), el algoritmo que mejor performance tuvo fue la Regresión Logística, con una precisión de 92.5%. Con estos resultados, se reutilizaron los archivos "pickle" (un output entregado por el algoritmo en el programa Python) con los que se procedió a clasificar las noticias. Estos archivos permiten reutilizar los parámetros optimizados que entrego el modelo de regresión logística, para realizar las clasificaciones.

Finalmente, el algoritmo entrega un set de datos con una columna nueva, la que con un 1 indica se la noticia es del ámbito financiero y con un 0 si no, con una probabilidad de acierto del 92.5%. Es importante mencionar que durante el trabajo se realizó un análisis minucioso de las metodologías utilizadas para llegar a este resultado, lo que sirvió como una corroboración del estudio anterior, ya que se llegaron a las mismas conclusiones en el uso de los algoritmos de clasificación. El resultado final luego de aplicar esta metodología son dos bases de datos diferentes, cada una con todas las noticias publicadas en el intervalo de tiempo mencionado, pero una contiene las noticias financieras y la segunda el resto de noticias.

2.5.Cálculo de Sentimiento o "Sentiment" financiero

Luego de tener procesados los datos, filtrando y clasificando según el estudio que se quiere realizar, se procede a crear los códigos en Rstudio que permitirán el cálculo de las variables de interés. Para eso se desarrolla un script con el que se lee y procesa las noticias diarias de la base de datos con los artículos financieros, que ya están etiquetadas con las categorías de Calomiris-Mamaysky. La variable a calcular tiene la siguiente forma:

$$sen_j = \frac{pos\_w_j + neg\_w_j}{w\_t_j}$$

Donde,
pos_w_j: Cantidad de palabras positivas en el texto o artículo j
neg_w_j: Cantidad de palabras negativas en el texto o artículo j
w_t_j: Cantidad de palabras en el texto o artículo j.

Esta variable es la denominada como sentimiento (sentiment en inglés) o sentimiento financiero, que fue el apronte definido en este estudio para medir el "estado de ánimo" con el que se publican las noticas, el que se postula que puede tener un efecto en los inversionistas al momento de tomar decisiones financieras.

Para llegar a este indicador, se realizó el siguiente trabajo:

a) Se definió un atributo llamado total_text, en la que se juntaba el título de la noticia, el primer párrafo, el abstract o resumen de la noticia y el "snippet", ya que eran los indicadores que entregaban información de la noticia para calcular el sentiment, utilizando las palabras de cada una.

b) Como se tenían todas las noticias diarias, primero se debía procesar cada una de las noticias (en este caso sobre la variable total_text), para calcular la variable de interés de cada una y luego sumarlas para tener el indicador de manera diaria. Para eso se desarrolló un "loop" y se utilizaron diferentes paquetes para procesamiento de palabras, con los que se buscaba eliminar los caracteres que no servían, corregir errores ortográficos y dejar los párrafos de la forma más simple posible para su procesamiento, es decir, aplicar técnicas de procesamiento de texto, para obtener una lista de las palabras dentro de la variable total_text y luego analizarlas.

c) Luego de eso, se definieron distintas funciones para clasificar las palabras, según el diccionario de Loughran-McDonald (2011). Este proceso fue realizado para cada artículo, es decir, para cada fila del dataset diario. La idea principal fue que por cada palabra de la lista de L-M que apareciera en el artículo, se guardaba en una de las variables descritas (la que contenía las positivas o las negativas), dependiendo de su tonalidad. Finalmente se contaban todas las palabras presentes en la noticia para poder así obtener el indicador de sentimiento financiero descrito anteriormente.

A pesar de ser la lista más utilizada por autores que estudian estos fenómenos, en un artículo de los mismos autores, pero más reciente (Loughran-McDonald (2016)), exponen las problemáticas y limitaciones de este cálculo. Se pueden resumir en estas tres aristas, que son las que presentan un problema para realizar el análisis de sentiment:

1) Doble negación: Este fenómeno hace alusión al uso de una negación antes de una palabra de connotación negativa o positiva, cambiando el sentido de la palabra. Un ejemplo sería "el resultado de la operación no fue malo". El algoritmo original solo tomaría en cuenta la palabra "malo", por lo que el sentimiento de la frase se catalogaría como negativo, siendo que en la realidad no fue así. Lo mismo pasaría para "el gobierno no lo está haciendo bien", ya que tiene connotación negativa, pero solo tomaría en cuenta la palabra bien, por lo que tendría sentimiento de mercado positivo.

   La forma en que se resuelve esta problemática es modificando un paquete[x] del software Rstudio, que calcula el sentiment de una oración. En un estudio similar, el lenguaje de programación utilizado fue distinto (Python donde se utilizaron paquetes de este programa y su sintaxis), por lo que se decidió generar un método que permitiera calcular el sentiment de cada observación, solucionando esta problemática y poder reafirmar que la forma anterior estaba correcta, pero en un software distinto y de esta manera comparar los resultados. Para lograr esto se descargó el script del paquete ya mencionado y se realizaron cambios dentro de este ya que, si bien la función lograba calcular el sentiment y tratar la barrera de la doble negación, también incluía otros métodos que no servían para el estudio realizado. Lo que este algoritmo realizaba era un análisis de las oraciones, es

decir, hasta donde se encontraba con un punto, asumía que podía haber variaciones en el "sentido" de las frases. Para esto tenía como input, una lista con todos los términos que le cambian el sentido a una frase u otra palabra dentro del artículo que se analizaba. Esta variable con los términos, el articulo y el diccionario de L-M eran los valores que recibía la función dentro del paquete obtenido, lo que permitía tener como resultado las palabras positivas, negativas y el total, modificadas de ser necesario si estaba presente alguno de las palabras que le cambiaba el sentiment.

2) Pesos de las palabras: como una primera aproximación, es una buena idea definir valores -1, 0 y 1 para palabras negativas, positivas y neutras respectivamente, pero siendo más objetivos, las palabras no tienen la misma intencionalidad, es decir, hay algunas que pueden ser catalogadas como "más negativas" o "más positivas". Un ejemplo sería bueno vs excelente.

   Para tratar esto, se utilizó la metodología IDF (inverse document frecuency), que es una buena aproximación a un método que se encargue de este problema, ya que, según la estructura del párrafo o noticia en este caso, permite medir la importancia de la palabra dentro de la forma del texto. Lo que hace este algoritmo es expresar numéricamente cuan común es un término en un numero de documentos estudiados. La fórmula es la siguiente:

$$idf(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right)$$

   En donde N está asociada al número total de textos o documentos escritos en la colección que se está estudiando (en este caso las noticias), el denominador es el número de documentos donde el término T se encuentra presente. Lo que finalmente hace esta fórmula es darle más peso a las palabras que tienen menores apariciones en los documentos, es decir, premia las palabras menos comunes, entre la colección de textos. Para motivos del estudio, hace bastante sentido considerar esto, ya que no es lo mismo decir pésimo que malo en un artículo (los inversores lo consideran diferente). Además, con este algoritmo se incluye la importancia por la frecuencia de la palabra, ya que hay términos que son más "comunes" que otros en la redacción de los artículos (por ejemplo, terrible versus malo al narrar un hecho).

   Es importante considerar que, a través del tiempo, la forma en que se redactan los artículos cambia, ya que el lenguaje no es estático, por lo que es una buena práctica dividir el diccionario en distintos periodos. Para esto se generar tres bases distintas que abarcan los siguientes intervalos; La primera va desde 1851 a 1900, la segunda de 1901 a 1951 y la tercera de 1952 a 2018. Con esta nueva división, se procedió a utilizar un algoritmo el que limpiaba los artículos, transformaba las

palabras con contracciones, es decir, se aplicaban técnicas de “text mining”, para dejar las noticias en un formato procesable. Finalmente se cuenta aplica el algoritmo idf descrito anteriormente utilizando las palabras encontradas en los artículos que estaban dentro del diccionario de L-M, para obtener la medida de sentiment corregida.

3) Doble categoría: Solo se clasifican algunas palabras, en este caso, las que están dentro de los diccionarios utilizados en el estudio, lo que se provoca que una cantidad importante de palabras quede sin ninguna influencia en el cálculo de esta variable (no perteneces a alguna de las dos categorías definidas).

   Para no dejar de lado esta información e incluir de alguna manera en el cálculo de la variable estas palabras neutras, se incluirá el método de VADER, con el que se soluciona esta dificultad. Este algoritmo de VADER *(valence aware dictionary and sentiment reasoner*), lo que lleva a cabo es una combinación entre un diccionario con las palabras que tienen asignación de puntaje en base al sentimiento, clasificándolas en positivas y negativas, pero con una jerarquía dentro de cada grupo. Un punto a favor de este algoritmo es que ya incluye una fórmula para lidiar con el problema de la doble negación, reconociendo además los signos de puntuación dentro de los textos y los acrónimos presentes. La forma en que se emplea este método es a través de un paquete que contenía este algoritmo, por lo que solo se debió incluir en el script, donde recibe la variable con la noticia y retorna el sentiment corrigiendo las problemáticas descritas.

2.6.Implementando Tópicos de Calomiris – Mamaysky (C-M)

El estudio tenía la intención de encontrar una relación entre el sentimiento de mercado y los retornos accionarios. Para esto había dos conceptos claros, la forma de calcular los retornos y la hipótesis de la relación (de lo que se habla más adelante). En el primer concepto, una idea fue utilizar los tópicos de Calomiris-Mamaysky (2018), ya que ellos en su trabajo postulan que el sentimiento de mercado se puede separar en 5 tópicos: mercado, gobierno, corporativo, crédito y commodities. Para identificar cada una de estas categorías, Calomiris-Mamaysky (C-M) desarrollan una lista con palabras, la que contenía conceptos del área econo/financiera. A través de ésta y el uso de la programación (procesamiento de texto), se desarrolló un código que permitió calcular el índice de C-M, para posteriormente clasificar cada una de las noticias. Se debe mencionar que cada una de estas categorías tenía palabras únicas asociadas, es decir, ninguna en la lista podía estar asociada a dos tópicos diferentes. El índice se calcula de la siguiente manera:

$$n_{\tau,j} = \frac{c_{\tau,j}}{c_j}$$

Donde $c_{\tau,j}$ es la cantidad de palabras asociadas al área econo/financiera, relacionadas con el tópico τ y c_j es el número de palabras dentro de la lista definida por CM, en la noticia. Este proceso es implementado para cada tópico, a través de un script desarrollado, el que verifica a que categoría pertenecen las palabras dentro de cada noticia, obteniendo el total para cada una de estas y luego las divide por la suma de las palabras económicas. Esto es clave ya que para obtener una medida de "sentiment" por cada categoría, se definió de la siguiente forma:

$$sen_{\tau,j} = n_{\tau,j} \cdot sen_j$$

Donde sen_j es el "sentiment" para el artículo j, y esto se tiene para cada uno de los artículos que se emiten de forma diaria. En el caso de este estudio, se estudiará la agregación mensual de esta métrica, por lo que será necesario utilizar el siguiente esquema:

$$s_\tau = \sum_j \frac{c_j}{c} \cdot s_{\tau,j}$$

Con c_j como la cantidad de apariciones de palabras de una de las categorías de C-M en los artículos y con c como el número total de palabras asociadas a alguna de las categorías.

2.7.Cálculo de InfoOverload

El estudio se centra en las noticias económicas, por lo que luego de obtener solo las que estén dentro de esta categoría, se estudia la tendencia de la cantidad de noticias a través del tiempo. Para esto se define una variable que intente capturar este efecto y se calcula de siguiente forma:

$$Number\_News = numNews_t = \frac{monthly\_news_{y,m}}{yearly\_news_{y-1}}$$

Donde $y = año$, $m = mes$. Finalmente, el indicador se obtiene dividiendo la cantidad de noticias en cada mes, por el número de artículos publicados en el año anterior, lo que permite observar cómo cambia la proporción de noticias por mes respecto al año anterior.

Con las metodologías anteriores ya definidas, se procedió a aplicarlas a dos casos distintos. El primero para la muestra total de noticias financieras y el segundo para los cinco subgrupos definidos por las categorías de Calomiris y Mamaysky.

2.8.Otras variables

Es importante controlar el efecto que se quiere medir, por otras variables econométricas que afecten a la problemática estudiada ya que, a pesar de obtener alguna significancia en el estudio, esta podría ser explicada con otra que ya es conocida en este ámbito, por lo que es una buena práctica incluirlas en los modelos, para evitar conclusiones erróneas. Para este trabajo, se utilizarán las siguientes:

- Inflación: Para su desarrollo se utiliza el índice de precios al consumidor (CPI), el que obtenido desde “Bureau of Labor Statistics”. Este se puede calcular de varias formas, pero en este trabajo se obtiene de la diferencia logarítmica de dos meses consecutivos del CPI. Se incluye dentro del estudio ya que, autores como Famma y Schwert (1977) y Fama (1981) realizan trabajos con este indicador y demuestran que este afecta directamente a los retornos de mercado

- Tasa libre de riesgo: Este indicador se calcula como el valor de un bono del tesoro de 1 mes (obtenido de la página de la Reserva Federal de St. Louis) menos el promedio de los 12 meses anteriores (es la tasa de interés).

- Producto interno bruto real (rgdp): Se incluye como variable independiente en el estudio, ya que es un indicador del crecimiento económico de un país y se calcula como C + I + G + (X – I), con C como el consumo, I como la inversión, G como el gasto público, X como exportaciones e I como importaciones.

- Varianza: se incluye la volatilidad de los retornos, calculadas mediante el método de GARCH, ya que es un factor que tiene directa implicancia sobre los movimientos del mercado (Guo, 2006).

3. Metodología Econométrica

En este trabajo, la metodología econométrica utilizada fue la de datos de panel en conjunto con una regresión lineal, a través del que se estudió la relación entre el market sentiment y la producción industrial en primera instancia, para luego realizar el mismo método, pero con el PIB real como variable dependiente. Dentro de esta metodología, se agregaron una serie de variables de control para encontrar el efecto real del indicador de interés y se utilizaron técnicas de time series, para agregar “lags” en cada experimento.

4. Análisis Preliminar

4.1. Análisis exploratorio variables interés

Es fundamental tener una idea general de cómo se comportan los datos antes de realizar cualquier tipo de análisis en profundidad. De esta manera se puede tener un enfoque más claro de que es lo que se puede hacer y si vale la pena seguir adelante con el estudio dados los datos que se tienen.

El primer apronte para esto es generar un gráfico que permita observar el comportamiento de la cantidad de noticias por año, para observar el comportamiento de esta variable y si existe alguna tendencia o un periodo anormal. El resultado se puede ver en la ilustración 1.

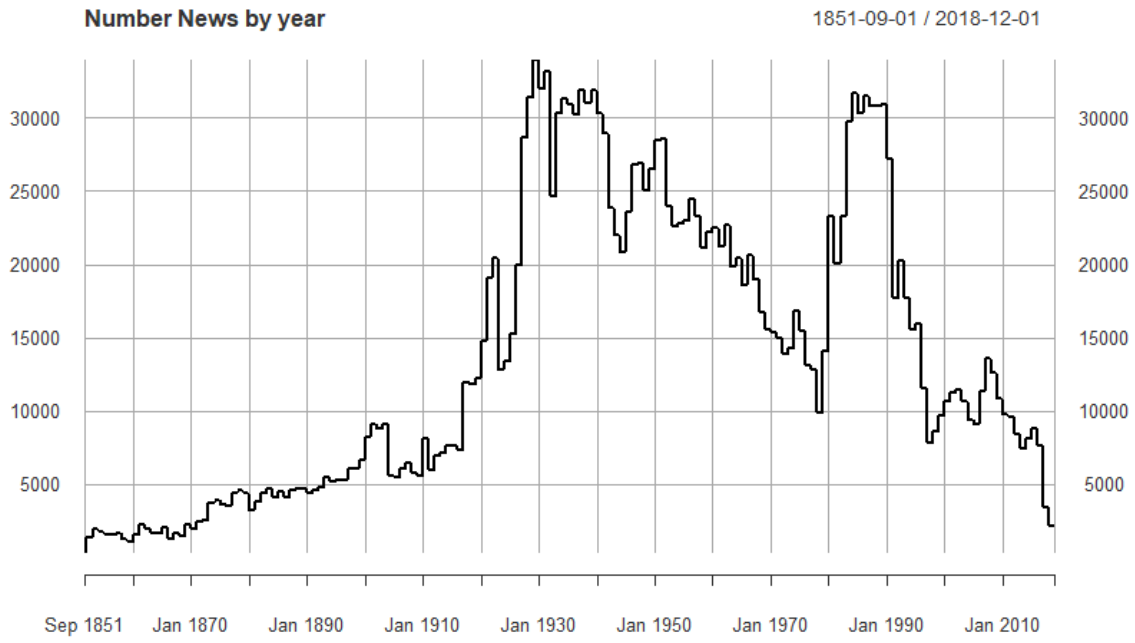

Ilustración 1: Cantidad de Noticias de Negocios Publicadas de Manera Anual.

En la sección número 1, se explica que esta información se obtuvo desde la página del New York times y luego fue procesada para llegar a este resultado. Se puede observar una tendencia al crecimiento de la cantidad de noticias publicadas, exceptuando periodos que tienen que ver con algún tipo de crisis, hasta aproximadamente los años 90. Esto hace sentido ya que, con los avances tecnológicos e innovaciones en el área de comunicación, permitieron tener un mundo más interconectado y con mayor acceso a este medio, por lo que se tiene una mayor facilidad a la hora de comunicar información

entre distintos países o dentro del mismo país, por lo que las noticias son más variadas y se tiene un mayor acceso. Por otra parte, la tendencia a la baja tiene que ver con que la gente hoy en día tiene otro tipo de opciones para informarse y de más fácil acceso.

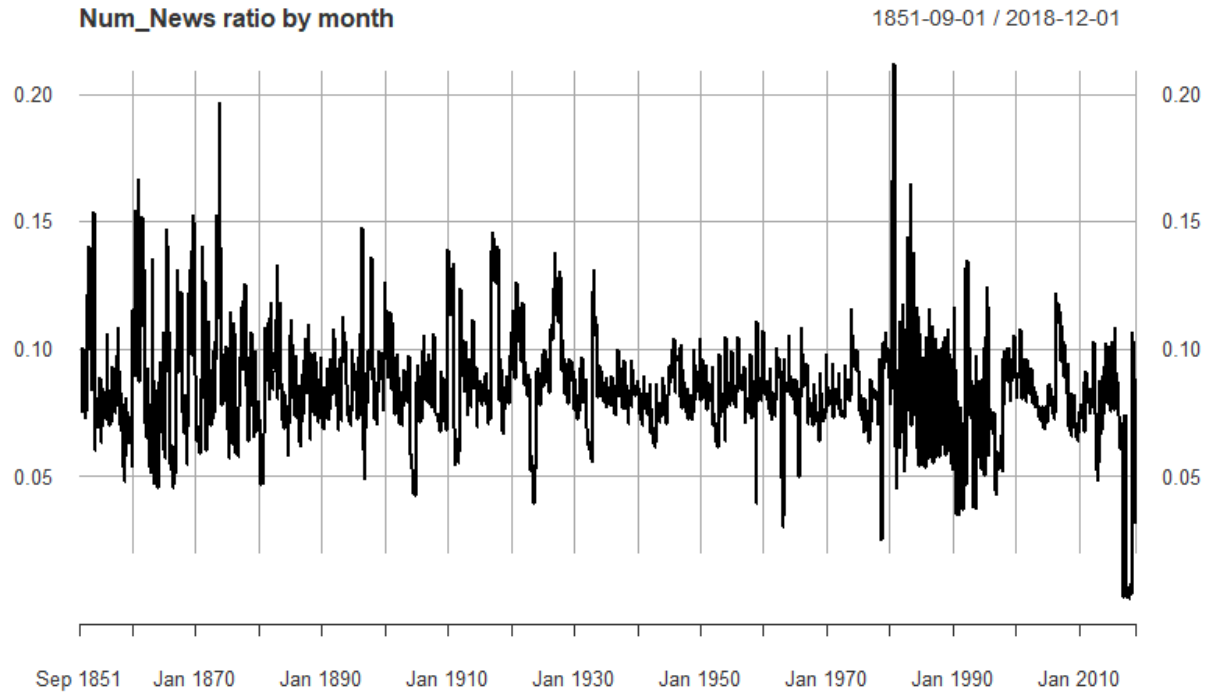

Ilustración 2: Evolución de Ratio "NumberNews" en el Tiempo.

Otro punto importante a tener en cuenta cómo ha evolucionado la tendencia de las noticias a través del tiempo. En este caso se desarrolla un gráfico del comportamiento de la métrica "Num_News ratio", como se plantea en la sección anterior. Se puede notar en la ilustración 2 como el "ratio" tiene una tendencia clara, la que se plasma como una oscilación en la ilustración, que se mueve tras ciertos valores, con algunos años en los que se presentan ciertos valles y peaks que no se explican a simple vista con la tendencia. Se puede agregar que la serie no es muy ruidosa, lo que tiene sentido debido a que las noticias deberían ser en esperanza constantes dado que estas se presentan diariamente. A partir de esta grafica se puede desprender que la cantidad de información ha crecido a través de los años en conjunto con las herramientas para informarse como se describió en la sección anterior, por lo que los agentes financieros tienen un acceso variado y rápido a la información.

El foco principal de esta investigación se encuentra en la variable "sentiment financiero", por lo que es importante ver su comportamiento. Específicamente, se decidió enfocarse en sentiment_lm_nh (que fue descrita más arriba) y en la misma variable, pero filtrada por la categoría de "market" de C-M (market sentiment_lm_nh). A continuación, se

muestra la evolución de la variable sentiment_lm_nh a través del tiempo, destacando los periodos donde se presentaron crisis, obtenidos desde la página de [1]nber[1], para poder observar detenidamente el comportamiento en estos periodos y si hay diferencias con el resto.

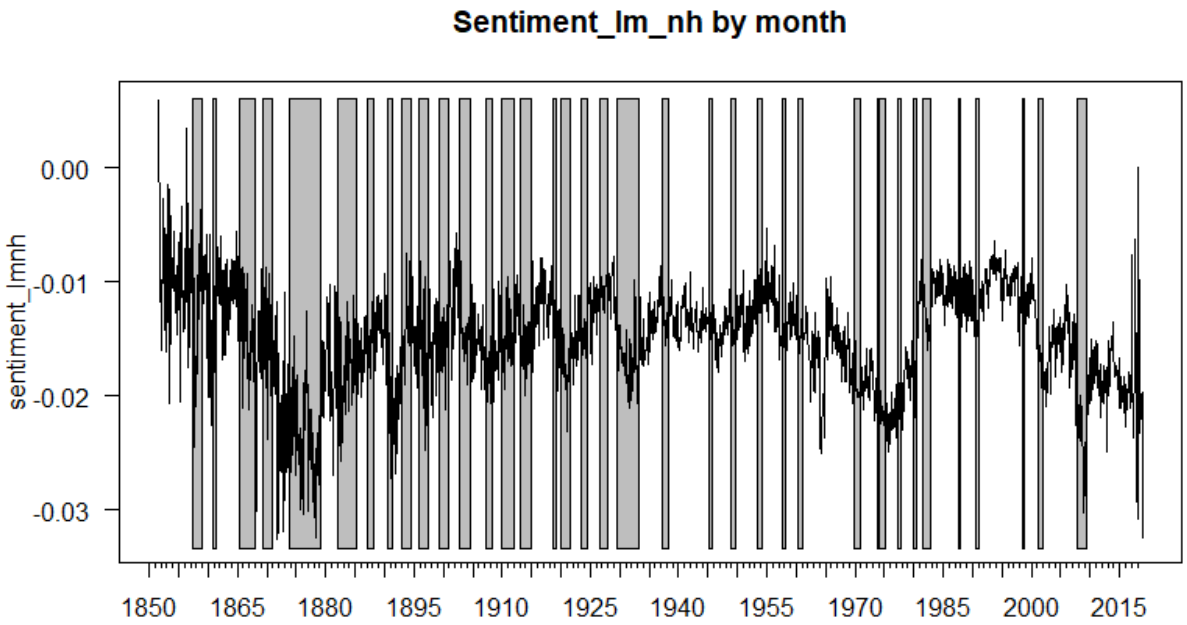

Ilustración 3: Sentimiento financiero con el diccionario de L-M y N-H, evolución mensual, destacando crisis.

Se puede observar que los valores parten positivos, tienden a disminuir hasta 1880 para luego en 1900 retomar un nivel del que luego no se presentan variaciones muy abruptas (entre -0.01 a -0.025), exceptuando los periodos mencionados anteriomente (crisis 1970, 2008) y el tramo de los datos mas recientes (2015 en adelante), donde el compotamiento que tiene presenta variaciones muy abruptas, pero sin alcanzar niveles positivos. Que los valores se mantengan siempre en valores negativos da una pista clara de como los medios tienden a exacerbar más los eventos negativos a través del tiempo, por lo que se puede concluir que historicamente la sociedad esta marcada por un pesimismo. Además entrega la primera pista sobre validez de la hipótesis que se desarrollará en la próxima sección, que tendrá como base que esta medida puede afectar los retornos y el rendimiento industrial del país, en este caso el sentimiento financiero más negativo coincide con las mayores crisis de la historia, como se puede ver en el gráfico.

[1] http://www.nber.org/cycles/cyclesmain.html

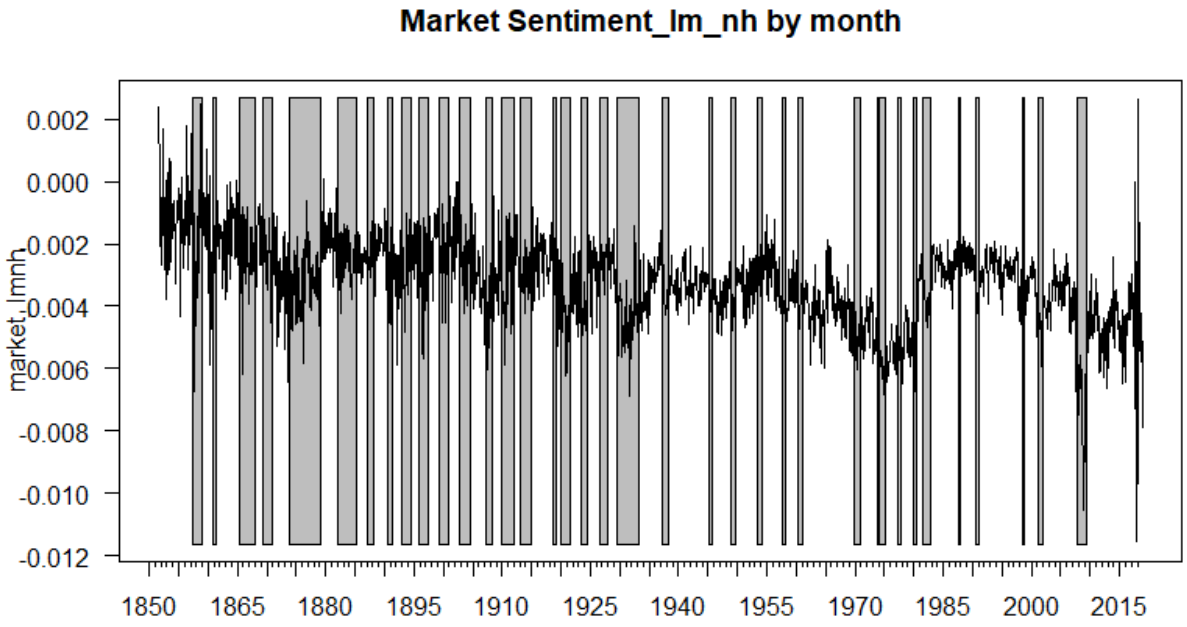

Ilustración 4: Sentimiento financiero con el diccionario de L-M y N-H, evolución mensual, destacando crisis, solo la categoría market de C-M.

En esta figura, se puede apreciar como la variable responde a lo esperado, es decir, para las crisis el sentimiento financiero disminuye (áreas destacadas en gris), como en los periodos de guerra, en 1980, 2000 y 2008. Para el resto de los periodos, disminuye la presencia de ruido, tiende a mantenerse estable la medida y con tendencia negativa.

Para ver si existe alguna diferencia, se realizó una división en tres periodos, para observar de manera más clara cómo se comporta este indicador. Los intervalos de tiempo se quedaron conformados de la siguiente forma:

1) Periodo previo a las guerras mundiales; 1851 a 1910
2) Periodo de guerras y reestructuración mundial: 1911 a 1960
3) Periodo post guerras; 1961 a 2018

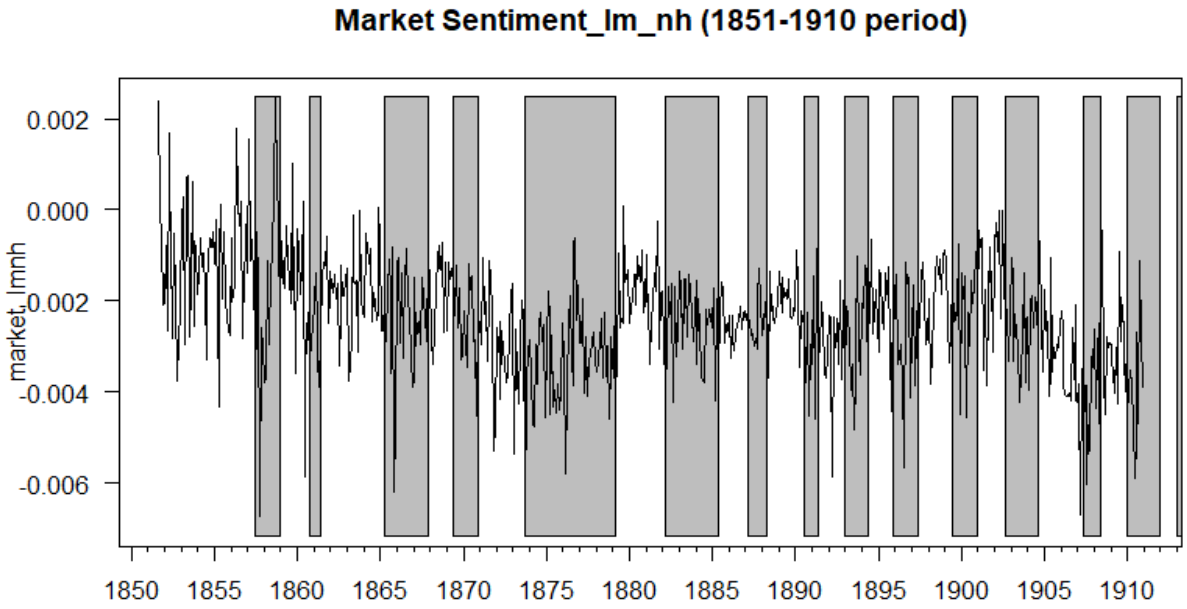

Ilustración 5: Sentimiento financiero con el diccionario de L-M y N-H, evolución mensual, destacando crisis, solo la categoría market de C-M entre 1851 y 1910.

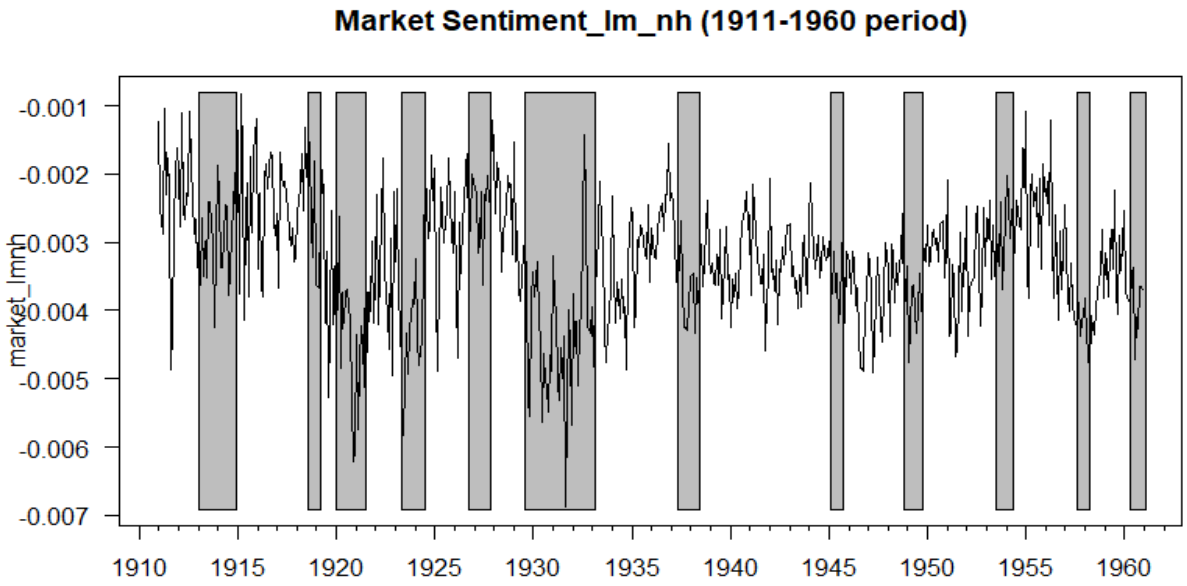

Ilustración 6: Sentimiento financiero con el diccionario de L-M y N-H, evolución mensual, destacando crisis, solo la categoría market de C-M entre 1911 y 1960.

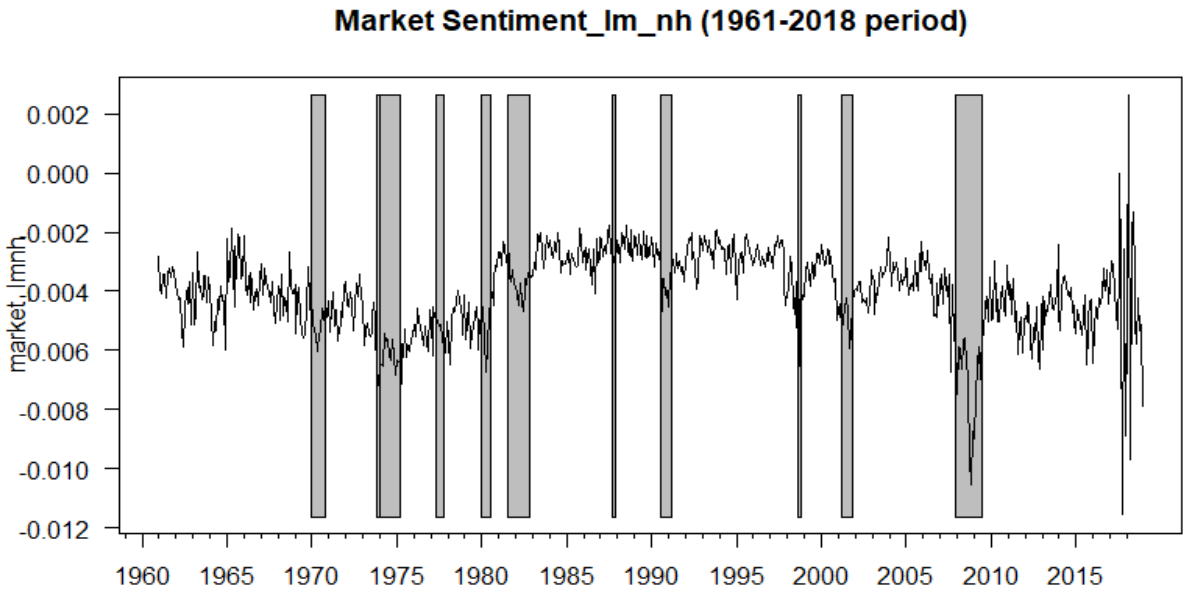

Ilustración 7: Sentimiento financiero con el diccionario de L-M y N-H, evolución mensual, destacando crisis, solo la categoría market de C-M entre 1961 y 2018.

Al generar estos tres gráficos, el efecto de las crisis representadas por los rectángulos grises se captura de mejor manera. Además, la serie se hace menos ruidosa observando de mejor forma como en los periodos de crisis la variable tiene el comportamiento esperado.

5. Resultados

5.1. Relación de sentimiento financiero y el producto interno bruto

El primer modelo planteado en este estudio tiene como foco estudiar la relación entre la variable de market_sentiment y el crecimiento del producto interno bruto de Estados unidos (gdp_growth).

La hipótesis propuesta en este primer experimento, se basa en las ideas propuestas anteriormente sobre la relación entre la liquidez del mercado, el sentimiento de mercado y la inversión en el país, ya que la relación entre estas tres, tienen un efecto en el crecimiento económico. Como primer foco, el trabajo de Danso (2019), que propone un modelo con el que encuentra que el sentimiento de mercado impacta la inversión de las empresas cuando estas incorporan los niveles de ganancia futura en las decisiones que toman, lo que se traduce como una relación sentimiento-inversión significativa y positiva en todos sus modelos, que se mantiene al tratar problemas de endogeneidad, plasmado un impacto significativo en periodos previos y posteriores a las crisis. Por esto, esta razón es que se tiene la conjetura de que el efecto del sentiment es positivo sobre el crecimiento del producto interno bruto.

$$Y_{t+h} = \beta_0 + MS_t + \gamma_h + Y_t + \epsilon_t$$

Donde,

Y: es el crecimiento del producto interno bruto (pib_growth), siendo la variable de interés del estudio. Además, aparecen como regresor para controlar por los lags (time series),
β: es el intercepto de la regresión,
γ: es el conjunto de variables de control en la regresión, siendo las siguientes:
$dstir$: es el delta de la tasa de interés,
$stdev$: es la volatilidad,
$infr$: es la tasa de inflación de los Estados Unidos,
$rgdp$: es el PIB real de Estados Unidos,
MS: es el sentimiento de mercado, siendo esta la variable de interés de la que se intenta medir el posible efecto sobre el crecimiento del producto interno bruto (market_sentiment),
ϵ_t: el error asociado al modelo.

Además, se controla por “lags” (), que van desde 1 a 12 periodos hacia atrás, es decir, se busca capturar para cuantos meses tiene poder predictivo la variable de sentimiento de merado. Por otra parte, las variables están estandarizadas y expresadas como porcentaje. Los resultados obtenidos se pueden observar en la tabla a continuación:

	gdpgrow_market_lm			gdpgrow_market_lm_nh		
Horizon (h)	1	2	3	1	2	3
sentiment	0.35**	0.87***	-0.16	0.38**	0.90***	-0.16
	(0.150)	(0.200)	(0.143)	(0.155)	(0.197)	(0.147)
deltastir	0.07***	0.01	0.02	0.07***	0.01	0.02
	(0.021)	(0.025)	(0.020)	(0.021)	(0.025)	(0.020)
stdev	-0.08***	-0.02	-0.05**	-0.08***	-0.02	-0.05**
	(0.027)	(0.040)	(0.025)	(0.027)	(0.039)	(0.025)
infrate	-0.01	-0.02	-0.07	-0.01	-0.03	-0.07
	(0.073)	(0.070)	(0.069)	(0.073)	(0.070)	(0.069)
rgdp	-0.03	-0.04*	-0.02	-0.03	-0.04*	-0.02
	(0.021)	(0.021)	(0.018)	(0.021)	(0.021)	(0.018)
gdp_grow	-0.18***	-0.18***	0.57***	-0.18***	-0.18***	0.57***
	(0.020)	(0.020)	(0.049)	(0.020)	(0.020)	(0.049)
Constant	0.00***	0.01***	0.00	0.00***	0.01***	0.00
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	851	848	845	851	848	845

Standard errors in parentheses

***** p<0.01, ** p<0.05, * p<0.1**

Tabla 1: Resultados Regresiones Variable Dependiente “Producto Interno Bruto” con “lags” = h y Variable Independiente “Market Sentiment” (2 formas) y 4 Variables de control. La regresión esta ajustada por el método de Newey-West para el p-value.

Para hacer un análisis más robusto, se replica el modelo anterior, pero con un crecimiento logarítmico en la variable dependiente, ya que lo esperado es que, al cambiar la forma de cálculo de la variable dependiente, no afecte significativamente a los resultados del modelo. Éste queda definido de la siguiente manera:

	log_gdpgrow_market_lm			log_gdpg_market_lm_nh		
Horizon (h)	1	2	3	1	2	
sentiment	0.35**	0.87***	-0.16	0.39**	0.90***	-0.16
	(0.149)	(0.199)	(0.142)	(0.154)	(0.196)	(0.145)
deltastir	0.07***	0.01	0.02	0.07***	0.01	0.02
	(0.020)	(0.025)	(0.020)	(0.020)	(0.025)	(0.020)
stdev	-0.08***	-0.02	-0.05**	-0.08***	-0.02	-0.05**
	(0.027)	(0.039)	(0.025)	(0.027)	(0.039)	(0.025)
infrate	-0.01	-0.02	-0.07	-0.01	-0.03	-0.07
	(0.073)	(0.069)	(0.068)	(0.072)	(0.069)	(0.068)
rgdp	-0.03	-0.03*	-0.02	-0.03	-0.04*	-0.02
	(0.021)	(0.021)	(0.018)	(0.021)	(0.021)	(0.018)
log_gdp_grow	-0.18***	-0.18***	0.57***	-0.18***	-0.18***	0.56***
	(0.021)	(0.020)	(0.050)	(0.021)	(0.020)	(0.050)
Constant	0.00***	0.01***	0.00	0.00***	0.01***	0.00
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	851	848	845	851	848	845

Standard errors in parentheses

***** p<0.01, ** p<0.05, * p<0.1**

Tabla 2: Resultados Regresiones Variable Dependiente "Producto Interno Bruto" con "lags" = h y Variable Independiente "Market Sentiment" (2 formas) y 3 Variables de control. La regresión esta ajustada por el método de Newey-West para el p-value.

Observando ambos modelos desarrollados, el resultado significativo se presenta para el modelo con 1 y 2 lags en el control, es decir, el poder predictivo se obtiene para uno y dos meses para la variable de market_sentiment. Para el efecto que tiene sobre la variable de interés, se encontró que aun después de agregar las variables de control descritas es positivo y significativo, además de ser consistente ya que la ésta se mantiene para el modelo logarítmico y el normal. Para los demás lags, no se encontraron efectos significativos y al tener como regresor solamente a la variable de market sentiment, si se presentaban efectos significativos, pero era importante que este se mantuviera al controlar por variables económicas que también se conociera que afectaban a las variables dependientes. Es importante mencionar que los regresores están estandarizados, para conocer el valor real del coeficiente asociado a cada uno. Finalmente, las diferencias al entre el diccionario LM (Loughran McDonald) y el LM-NH (Loughran McDonald corrigiendo la doble negación), no presenta diferencias significativas en el efecto sobre la variable dependiente.

5.2. Relación de sentimiento financiero y producción industrial

El segundo modelo estudiado en este trabajo es el que intenta capturar la relación entre la variable de market_sentiment y el crecimiento de la producción industrial de Estados Unidos (industrial_production_growth). El objetivo de este trabajo es indagar en el posible efecto que pueda tener el sentimiento financiero en la producción del mercado de Norteamérica. Antes declarar la hipótesis, importante definir primero que se entiende como producción industrial. El Índice de Producción Industrial (INDPRO) es un indicador económico que mide la producción real para todas las instalaciones ubicadas en los Estados Unidos que fabrican, explotan, y servicios de electricidad y gas[2].

En base a esto, la hipótesis que plantea es que a mayor "Market_sentiment", este tiene un efecto positivo en el crecimiento de la producción industrial. Esta hipótesis se basa en la relación implícita que existe entre el sentimiento de mercado, el sentimiento del inversor y los retornos de los activos.

Primero, se tiene el estudio de Dashag Huang (2014), donde usando como base los estudios de Baker y Wugler (2006) y el de Kelly y Pruitt (2013), genera una medida de sentimiento del inversor que inversores tiene un poder predictivo mucho mayor para el mercado de valores agregado de lo que se pensaba anteriormente. Además, se desempeña mucho mejor que la mayoría de las variables macroeconómicas comúnmente utilizadas, y su previsibilidad es estadística y económicamente significativa. Gurley y Shaw (1955) enfatizan los flujos de mecanismos de fondos y argumentan que el mercado de valores desempeña un papel vital en la canalización de los ahorros hacia las inversiones, asignando fondos de los prestamistas a los prestatarios para la producción. Por otra parte, los rendimientos de las acciones reflejan información sobre las expectativas corporativas de los flujos de efectivo futuros y las tasas de descuento, que proporcionan una señal efectiva para medir la producción industrial y las expectativas de los inversores (Fama, 1990; Schwert, 1990). A esto, si se le agrega la influencia que tienen las noticias sobre el "sentiment" de los inversores (Martha Starr, 2008) y que Tetlock (2007) llega al resultado de que la información y la forma en que los diarios entre 1984 hasta 1999, pueden prever las fluctuaciones en el mercado financiero. Por otra parte, en un estudio con similitudes al anterior, García (2013), se lleva a cabo un análisis del efecto que tiene el sentimiento de mercado sobre el valor de los activos, entre los años 1905 y 2005. Para calcular la variable sentimiento, se utilizaron dos columnas financieras del New York Times. Por estos antecedentes, se puede concluir que el efecto del "Market_sentiment" debiese ser positivo.

[2] https://es.investing.com/economic-calendar/industrial-production-161

Para esto se desarrolla una regresión lineal la que busca medir la existencia de algún efecto de la variable de interés, controlando por otros factores macroeconómicos. El modelo queda definido de la siguiente forma:

$$Y_{t+h} = \beta_0 + MS_t + \gamma_h + Y_t + \epsilon_t$$

Donde,

Y: es el crecimiento de la producción industrial(ind_growth), siendo la variable de interés del estudio. Además, aparecen como regresor para controlar por los lags (time series),
β: es el intercepto de la regresión,
γ: es el conjunto de variables de control en la regresión, siendo las siguientes:
$dstir$: es el delta de la tasa de interés,
$stdev$: es la volatilidad,
$infr$: es la tasa de inflación de los Estados Unidos,
$rgdp$: es el PIB real de Estados Unidos,
MS: es el sentimiento de mercado, siendo esta la variable de interés de la que se intenta medir el posible efecto sobre el crecimiento del producto interno bruto (market_sentiment),
ϵ_t: el error asociado al modelo.

Además, se controla por "lags" (), que van desde 1 a 12 periodos hacia atrás y se comparan los resultados suavizando la variable dependiente al aplicarle logaritmo. Los resultados obtenidos se pueden observar en la tabla a continuación:

	indgrowth_market_lm			**indgrowth_market_lm_nh**		
Horizon (h)	1	2	3	1	2	3
sentiment	0.61***	0.66***	0.14	0.63***	0.66***	0.15
	-0.212	(0.240)	(0.244)	(0.219)	(0.249)	(0.252)
deltastir	0.07**	0.09***	0.06*	0.07**	0.09***	0.06*
	(0.030)	(0.028)	(0.032)	(0.030)	(0.028)	(0.032)
stdev	-0.12***	-0.18***	-0.19***	-0.12***	-0.18***	-0.19***
	(0.044)	(0.052)	(0.048)	(0.044)	(0.052)	(0.048)
infrate	0.02	-0.11	-0.20	0.02	-0.12	-0.20
	(0.098)	(0.109)	(0.138)	(0.098)	(0.109)	(0.138)
rgdp	-0.02	-0.02	-0.04	-0.02	-0.02	-0.04
	(0.034)	(0.040)	(0.044)	(0.034)	(0.041)	(0.044)
indgrowth	0.33***	0.19***	0.18***	0.33***	0.19***	0.18***
	(0.053)	(0.057)	(0.054)	(0.053)	(0.057)	(0.054)
Constant	0.00***	0.00***	0.00**	0.00***	0.00***	0.00**
	-0.001	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	851	848	845	851	848	845

Standard errors in parentheses
***** p<0.01, ** p<0.05, * p<0.1**

Tabla 3: Resultados Regresiones Variable Dependiente “Industrial Production” con “lags” = h y Variable Independiente “Market Sentiment” (2 formas) y 3 Variables de control. La regresión esta ajustada por el método de Newey-West para el p-value.

Para hacer un análisis más robusto, se repite lo realizado en el primer estudio, replicando el modelo anterior, pero con un crecimiento logarítmico en la variable dependiente:

	logindgrowth_market_lm			logindgrowth_market_lm_nh		
Horizon (h)	1	2	3	1	2	3
sentiment	0.61***	0.66***	0.14	0.63***	0.66***	0.15
	(0.212)	(0.240)	(0.245)	(0.219)	(0.248)	(0.253)
deltastir	0.06**	0.09***	0.06*	0.07**	0.09***	0.06*
	(0.030)	(0.029)	(0.032)	(0.030)	(0.029)	(0.032)
stdev	-0.13***	-0.18***	-0.19***	-0.12***	-0.19***	-0.19***
	(0.044)	(0.052)	(0.048)	(0.044)	(0.053)	(0.048)
infrate	0.02	-0.11	-0.20	0.02	-0.12	-0.20
	(0.098)	(0.109)	(0.139)	(0.098)	(0.109)	(0.139)
rgdp	-0.01	-0.02	-0.04	-0.02	-0.02	-0.04
	(0.033)	(0.040)	(0.043)	(0.034)	(0.040)	(0.043)
logindgrowth	0.33***	0.19***	0.18***	0.33***	0.19***	0.18***
	(0.053)	(0.057)	(0.054)	(0.053)	(0.057)	(0.054)
Constant	0.00***	0.00***	0.00**	0.00***	0.00***	0.00**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	851	848	845	851	848	845

Standard errors in parentheses

***** p<0.01, ** p<0.05, * p<0.1**

Tabla 4: Resultados Regresiones Variable Dependiente “Log Industrial Production” con “lags” = h y Variable Independiente “Market Sentiment” (2 formas) y 3 Variables de control. La regresión esta ajustada por el método de Newey-West para el p-value.

Como se observa, para ambos modelos el resultado significativo se presenta para el modelo con 1 y 2 lags en el control y el efecto encontrado agregando las variables de control descritas es positivo, además de ser consistente ya que la significancia se mantiene para el modelo logarítmico y el normal. Para los demás lags, no se encontraron efectos significativos y al tener como regresor solamente a la variable de market sentiment, si se presentaban efectos significativos, pero era importante que este se mantuviera al controlar por variables económicas que también se conociera que afectaban a las variables dependientes. Es importante mencionar que los regresores están estandarizados, para conocer el valor real del coeficiente asociado a cada uno. Finalmente, las diferencias al entre el diccionario LM y el LM-NH (corrigiendo la doble negación), no presenta diferencias significativas en el efecto sobre la variable dependiente.

6. Conclusiones

A través de este trabajo, se realizó una serie de procesos a través de los para desarrollar una base de datos que tenga las noticias desde 1851 a 2018 del New York Times para estudiar el efecto del sentimiento de mercado sobre distintas variables de interés.

Como primer punto, el tiempo que toma procesar datos, sobre todo texto es grande, por lo que lo ideal es utilizar algoritmos que optimicen estos procesos. Es importante en esta línea el sustento que prestan los algoritmos de machine learning y las tecinas de minería de datos para que estos proyectos disminuyan su tiempo de ejecución. Como posible mejora sería bueno actualizar los mecanismos que se utilizaron para clasificar las noticias de manera de perder la menor información posible, ya que cada año el New York Times mejora la información en sus bases de datos.

Hipótesis 1: Modelo sentimiento de mercado y Producto Interno Bruto

En este estudio, se encuentra evidencia del efecto positivo que tiene el sentimiento financiero sobre el crecimiento PIB (rgdp en inglés o producto interno bruto). Para los cuatro proxis de la variable de sentiment, solo dos tuvieron resultados significativos y consistentes, en este caso la variable que utilizaba el diccionario de Loughran McDonald (2011) y la que incluía este mismo, pero solucionando el problema de la doble negación. Un punto importante es que dado la cobertura que posee la base de datos (1851 a 2018), que los resultados se mantengan significativos aun controlando por otras variables es aún más relevante. Este efecto desaparece para mas de dos periodos en adelante. Además, este es consiste para la variable sin clasificar por los tópicos de C-M y agregando la importancia de la palabra dentro del artículo (IDF).

Una hipótesis importante es que este resultado tiene que ver con el efecto del sentimiento de mercado sobre distintas variables que afectan al PIB, como son la inversión, el gasto y el consumo de los actores en el mercado financiero. Por lo tanto, la variable de interés explica cierto porcentaje de la varianza del crecimiento del PIB.

En las posibles mejoras, sería interesante ver si el efecto se mantiene al utilizar la variable dependiente de forma mensual en vez de forma cuatrimestral. A esto se le puede sumar ver si el efecto sigue siendo relevante al incluir variables como inversión, consumo, que son otras que en la literatura se ha demostrado verse afectadas por el sentimiento y al PIB. Por último, sería una buena idea testear si este efecto se mantiene en otros mercados que no son el estadounidense.

Hipótesis 2: Modelo sentimiento de mercado y producción industrial

Dados los resultados que se entregó el modelo, se llegó a la conclusión de que existe una relación positiva entre el sentimiento de mercado (proxy de la “tonalidad” de la noticia) y el crecimiento de la producción industrial. Se baraja que ésta se debe a el efecto que tiene la variable dependiente viene de la mano con la relación que tiene el sentiment con los mercados de acciones y los activos, ya que estos influyen directamente a la producción industrial de un país. Un punto muy interesante es que, aun incluyendo variables de control, la variable de interés sigue teniendo poder explicativo sobre el crecimiento de la producción industrial, lo que da una señal de que incluirla en modelos de estimación o análisis de esta, puede llevar a mejores resultados

Para profundizar más en este tipo de estudios y como pasos que se pueden mejorar, separar la variable de sentiment por la tendencia histórica para definir un high y low, lo que podría capturar de mejor manera el efecto de esta variable sobre los indicadores estudiados. Otro punto interesante sería investigar el efecto diario de esta variable, para ver si el efecto se mantiene o desaparece. Además, incluir un estudio más profundo de cómo se compartan este modelo en las épocas de contracción económica también puede ser una arista interesante. Finalmente, dada la cobertura que tiene esta base de datos (desde 1851 a 2018), puede ser un buen input para muchos estudios que busquen ver el efecto de estos proxis de sentimiento de mercado versus otras variables financieras interesantes, como la liquidez.

Bibliografía

1. Baker, Malcom. & Wurgler, Jeffrey. (2006). Investor Sentiment and the Cross-Section of Stock Returns. The Journal of Finance, 61(4), 1645-1680.
2. Bhimjee, D. C., Ramos, S. B., & Dias, J. G. (2016). Banking industry performance in the wake of the global financial crisis. International Review of Financial Analysis, 48, 376-387.

3. Calomiris, C. and H. Mamaysky (2018). How News and Its Context Drive Risk and Returns around the World. Columbia Business School Research Paper No. 17-40.

4. Chiang, T. C., & Chen, X. (2017). Stock Market Activities and Industrial Production Growth: Evidence from 20 International Markets. Advances in Pacific Basin Business, Economics and Finance, 39-75.

5. Danso, A., Lartey, T., Amankwah-Amoah, J., Adomako, S., Lu, Q., & Uddin, M. (2019). Market sentiment and firm investment decision-making. International Review of Financial Analysis, 66, 101369.
6. F.A. Longstaff, Financial claustrophobia: Asset pricing in illiquid markets, Tech. report, National Bureau of Economic Research, 2004.
7. Fama, E. (1981). Stock returns, real activity, inflation, and money. American Economic Review 71, 545–565.
8. Fama, E. and G. Schwert (1977). Asset returns and inflation. Journal of Financial Economics 5, 115–146.

9. Fama, Eugene F. (1990). Stock Returns, Expected Returns, and Real Activity. The Journal of Finance, 45(4), 1089-1108.
10. Fan, C. S., & Wong, P. (1998). Does consumer sentiment forecast household spending? Economics Letters, 58(1), 77-84.
11. Florackis, C., Giorgioni, G., Kostakis, A., & Milas, C. (2014). On stock market illiquidity and real-time GDP growth. Journal of International Money and Finance, 44, 210-229.
12. García, D. (2013). Sentiment during Recessions. The Journal of Finance 68 (3), 1267-1300.
13. Guo, H. (2006). On the out-of-sample predictability of stock market returns. Journal of Business 79, 645–670.
14. Gurley, John G. AND Shaw, Edward S. "Financial Aspects of Economic Development," Amer. Econ. Rev., Sept. 1955, 45(4), pp. 515- 38.

15. Ho, C., & Hung, C.-H. (2009). Investor sentiment as conditioning information in asset pricing. Journal of Banking & Finance, 33(5), 892-903.
16. Huang, D., Jiang, F., Tu, J., & Zhou, G. (2014). Investor Sentiment Aligned: A Powerful Predictor of Stock Returns. Review of Financial Studies, 28(3), 791-837.
17. Kayacetin, Nuri Volkan and Kaul, Aditya, Forecasting Economic Fundamentals and Stock Returns with Equity Market Order Flows (February 9, 2009).
18. Kelly, B., and S. Pruitt. 2013. Market expectations in the cross-section of present values. Journal of Finance68: 1721–1756.
19. Levine, R. and Zervos, S. (1998). Stock Markets, Banks, and Economic Growth. The American Economic Review Vol. 88, No. 3 (Jun., 1998), pp. 537-558
20. Loughran, T. and B. McDonald (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance 66 (1), 35-65.
21. Loughran, T. and B. McDonald (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research 54 (4), 1187-1230.
22. NAES, R., SKJELTORP, J. A., & ØDEGAARD, B. A. (2011). Stock Market Liquidity and the Business Cycle. The Journal of Finance, 66(1), 139-176.
23. Oh, S., & Waldman, M. (1990). The Macroeconomic Effects of False Announcements. The Quarterly Journal of Economics, 105(4), 1017.
24. Ross Levine. Stock markets, growth, and tax policy. Journal of Finance,46:1445–1465,1991.
25. Schwert, G. William, 1990, Stock returns and real activity: A century of evidence. Journal of Finance 45, 1237-1257.
26. Shen, J., Yu, J., & Zhao, S. (2017). Investor sentiment and economic forces. Journal of Monetary Economics, 86, 1-21.
27. Star, M. (2010). CONSUMPTION, SENTIMENT, AND ECONOMIC NEWS. Economic Inquiry, 50(4), 1097-1111.
28. Tetlock, P. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance 62 (3), 1139–1168.