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

VALORACIÓN DE ACTIVOS FINANCIEROS VÍA ANÁLISIS DE NOTICIAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

SHADY SHAHWAN CAUSA

PROFESOR GUÍA:
MARCELA VALENZUELA BRAVO

MIEMBROS DE LA COMISIÓN:
ALEJANDRO BERNALES SILVA
PATRICIO VALENZUELA AROS

SANTIAGO DE CHILE
2019

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial
POR: Shady Shahwan Causa
FECHA: 09/12/2019
PROFESOR GUÍA: Marcela Valenzuela Bravo

VALORACIÓN DE ACTIVOS FINANCIEROS VÍA ANÁLISIS DE NOTICIAS

Una de las grandes interrogantes asociadas a las finanzas es como los inversionistas toman decisiones. Es de esperar que los inversionistas tomen decisiones financieras en base a información. De esto nace una rama de las finanzas (Behavioral Finance) y una gran interrogante: como los medios de comunicación influyen en la decisión de los inversionistas. Es por esto por lo que durante el desarrollo de esta investigación se genero una base de datos histórica de las noticias publicadas diariamente por el New York Times desde su puesta en circulación en el año 1851 hasta el 2018.

Para el primer caso fue posible determinar la cantidad de noticias publicadas de manera mensual y calcular un proxy para sobrecarga de información. Con esto se encontró que, a mayor sobrecarga de información, aumenta el volumen transado (sobrecarga de volumen). Luego se concluyó que a mayor volumen mayores son los retornos de mercado.

Por otro lado, fue posible determinar el sentimiento de mercado a partir de las noticias publicadas y ver que existe una relación positiva entre el sentimiento de mercado y los retornos de mercado (S&P 500). Mas aun, el sentimiento de mercado es capaz de predecir los retornos a 1 mes, al igual que la sobrecarga de volumen.

Después se utilizan variables de control para entender si el sentimiento de mercado y el volumen están explicando algún fenómeno ya explicado en la literatura. Como resultado se obtiene que ambas variables son significativas frente a las variables de control tanto en el corte transversal de los datos como en la serie de tiempo de los retornos. Esto nos sugiere que podrían existir 2 factores uno asociado al sentimiento de mercado y el otro al volumen transado.

“La mejor forma de predecir el futuro es inventándolo”

Alan Kay

Agradecimientos

En primer lugar, me gustaría agradecer a mi profesora guía Marcela Valenzuela y al profesor Alejandro Bernales por su confianza. Continuamente me llevaron por esta aventura y sin su dirección, ayuda constante y apoyo este trabajo no habría sido posible.

Gracias a mis padres por ser los principales promotores de mis sueños. Me enseñaron la pasión que tengo por las matemáticas, ciencias y finanzas. Por sus palabras de aliento que no me dejaban caer en los momentos más difíciles. Por haberme apoyado durante estos 6 años de universidad. Este logro es gracias a ustedes.

A mi hermana Catalina por ser una fuente de motivación e inspiración para poder superarme cada día más.

A mis amigos: Andrés, Benjamín, Boris, Cristóbal, José Tomás, Nicolás, Pablo y Pedro Pablo por los buenos ratos, discusiones y constante apoyo tanto en la universidad como fuera de ella.

Finalmente agradezco a todas las personas que estuvieron conmigo en este proceso. Sin ustedes no hubiese sido capaz de terminar una de las etapas mas importantes de mi vida

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

1. Introducción

Sin dudas, la era actual corresponde a una era de notables innovaciones en tecnologías de información. Es fascinante considerar como podrían afectar al futuro de las finanzas y economías. Si bien están sucediendo innovaciones radicales, una cosa queda inamovible: la toma de decisiones de los agentes humanos. Una pregunta que aún no ha sido resuelta es como la información afecta a las personas, y en particular, la influencia de las noticias en la toma de decisiones de los agentes y como esto se ve reflejado en el movimiento de los mercados.

Los diarios, radios, revistas y, en general, la mayoría de los medios, han experimentado crecimiento significativo en las últimas décadas. Una de las razones de esto es la influencia que tienen en los lectores/espectadores. En particular el efecto de los medios en las decisiones de los inversionistas y como se refleja en los movimientos de los mercados financieros en un problema que ha recibido considerable interés de la comunidad académica en el último tiempo (e.g.: Fang y Peress (2009), Engelberg y Parsons (2011), Dougal et al (2012), García (2013), Peress (2014), Baker et al (2016), Manela and Moreira (2017)).

Una posible explicación de este efecto es que los contenidos de las noticias influencian el sentimiento de mercado de los inversionistas. El sentimiento de mercado se refiere a la actitud de los inversionistas frente a un instrumento en específico o al mercado financiero. Se trata del sentimiento y emoción de un mercado, o psicología de las masas, la que se ve reflejada en la cantidad de actividad en los mercados y los movimientos de precio. La reciente literatura argumenta que el sentimiento de mercado se ve más afectado por noticias financieras negativas, haciendo que los inversionistas tomen decisiones basadas en ese sentimiento, anticipando los movimientos del mercado. Tetlock (2007) encuentra que los contenidos de los diarios pueden predecir los movimientos del mercado desde 1984 hasta 1999. Además, muestra que un pesimismo promedio alto – calculado a partir de una columna financiera del diario Wall Street Journal – predice una baja en los precios de las acciones seguido por una reversión a los fundamentales. En un paper similar, Garcia (2013) estudia los efectos del sentimiento de mercado en los precios de activos desde un periodo de 1905 hasta 2005. El índice de sentimiento es extraído de dos columnas de noticias financieras del diario New York Times.

En primer lugar, esta tesis se enfoca en el efecto del sentimiento de mercado en los movimientos de los mercados financieros a través de las noticias relacionadas a economía y finanzas. Luego se examinan que tipos de noticias financieras tienen más efecto sobre los retornos accionarios. Se examinan 5 dimensiones: crédito, commodities, gobierno, mercado y corporaciones.

En segundo lugar, esta tesis se enfoca en examinar el efecto de la sobrecarga de información en el mercado financiero. De acuerdo con la Fundación Gallup/Knight (2018), el 58% de los adultos estadounidenses dicen que es difícil estar informado hoy en día debido a la cantidad de información y varias fuentes de noticias disponibles, mientras que el 38% dice que fácil. La sobrecarga de información podría afectar los sesgos de los inversionistas. Estos sesgos podrían prevenir un análisis claro sobre que está sucediendo en el mercado accionario, llevando a los inversionistas a tomar decisiones apresuradas, considerando información que es irrelevante en la toma de decisión y omitiendo información importante. Esto se traduce en toma de decisiones financieras malas y en precios de activos que divergen del valor fundamental. En periodos de problemas financieros, se espera que los inversionistas se vean aún más afectados por la sobrecarga de información. En particular, se examinará la influencia de exceso de información sobre la volatilidad de los retornos y el volumen transado tanto para periodos estables como inestables. Además, se estudiará el efecto en los retornos de mercado.

Para el desarrollo de este estudio, se extraen todas las noticias publicadas en el diario New York Times para los años 1851 hasta 2018. Para construir un proxy de sentimiento de mercado, se seleccionarán todas las noticias relacionadas a finanzas. Luego se contarán la cantidad de palabras positivas y negativas de cada artículo para calcular tonalidad. En otras palabras, si la proporción de palabras positivas es mayor a la cantidad de palabras negativas para un artículo, entonces se dirá que ese artículo tiene una tonalidad positiva. En resumen, se tendrá la proporción de palabras positivas y negativas para cada artículo relacionado a finanzas en tiempo t. El índice de sentimiento se obtiene calculando el promedio ponderado de la tonalidad sobre la cantidad de palabras por artículo.

Por otro lado, para calcular un proxy de la sobrecarga de información, se contarán todas las noticias financieras, divididas por la cantidad de noticias de una fecha t. Luego se aplicará un filtro a esa proporción para obtener su tendencia histórica, así se puede calcular un estimado de cuanto es el número usual de artículos financieros. Esa tendencia corresponde a un umbral que permite identificar los días con sobrecarga de información. Cuando la proporción de noticias financieras está sobre la tendencia, se dirá que existe exceso de información. Finalmente, el proxy será diferencia entre la proporción de noticias financieras y su tendencia histórica dada que la proporción está sobre la tendencia.

2. Datos

2.1. Fuentes de Información

En primer lugar, se extraerán todas las noticias del New York Times comprendida entre septiembre de 1851 y diciembre de 2018. Se extrae este intervalo porque es cuando se publica la primera noticia del New York Times. Las noticias incluirán los siguientes atributos[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)

[1] https://developer.nytimes.com/docs/articlesearch-product/1/overview

- *Url*: Url del New York Times para acceder al artículo (desde 1851)

Se utiliza como fuente de información el *New York Times*, dado que es el periódico de mayor circulación en Estados Unidos, con casi dos millones de ejemplares emitidos diariamente[2], de manera adicional, es la página web 170 con más visitas a nivel mundial y 301,88 millones de visitas[3]. Además, posee una infraestructura tecnológica acorde con el desafío de extracción de información. El *New York Times* tiene una plataforma que permiten acceder a su información mediante el uso de APIs. El término API es un acrónimo anglosajón (*Application Programming Interface*), las APIs definen las reglas que los programadores deben seguir para poder interactuar con algún software[4]. El *New York Times* cuenta con 11 APIs:

- *Archive API*: permite obtener toda la metadata 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

[2] https://www.worldatlas.com/articles/the-10-most-popular-daily-newspapers-in-the-united-states.html Consultada el 29 de junio de 2019

[3] https://www.similarweb.com/website/nytimes.com Consultada el 29 de junio de 2019

[4] https://www.howtogeek.com/343877/what-is-an-api/

Para efectos de esta investigación, se utiliza *Archive API* ya que es la API que permite descargar todos los artículos del New York Times de manera mensual.

2.2. Descarga de Datos

Para poder descargar los datos se utilizó la *Archive API*. Para poder utilizar esta API, hay que crear una llave. Esta se obtiene en la página web de Desarrolladores del *New York Times*: https://developer.nytimes.com, mediante la creación de un usuario y contraseña. El problema que conlleva esto es que solo se pueden realizar 1.000 llamadas por día (donde 1 llamada corresponde a 10 artículo), que es muy inferior a la cantidad total de artículos a descargar (2.263.925). Para poder solucionar este problema, se crearon 100 llaves de manera manual, utilizando mails temporales (de la página web: https://temp-mail.org/es/). El uso de mail temporales evita el gasto innecesario de tiempo en la creación de un mail. Para poder llevar registro de los usuarios y llaves creadas se creó un archivo csv con 3 columnas: email registrado, contraseña y llave API obtenida.

Una vez creada las llaves se procedió a descargar los datos. Para descargar los artículos se creó un programa en Python que permitía utilizar la *Archive API*. El programa en un inicio consultaba la URL de la API:

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

Donde año y mes corresponde a la fecha que se va a consultar y llave corresponde a la llave API obtenida. La respuesta a este URL es un JSON (*JavaScript Objet Notation*), el cual es un formato utilizado para el intercambio de datos, además de ser liviano y parseable de manera muy sencilla[5]. Ejemplo de una respuesta JSON: “headline”: “Gunpowder Explosion at Norfolk”. El JSON contiene todas las noticias para ese mes de ese año con todos los atributos mencionados en el capítulo 1.1. Una vez que se descargan todas las noticias del mes, el algoritmo Python guarda el JSON en un archivo de texto formato “txt” que será procesado más adelante (se guarda un archivo por mes-año).

[5] https://www.json.org/

Para poder optimizar el código y aprovechar el límite diario de 1.000.000 de artículos (de las 100 llaves de la API generadas), se utilizaron *Threads*. Esto permite que una vez que se hace la consulta a la página web y se espera a que se termine de descargar, el programa inmediatamente pase a la siguiente llave a hacer una segunda consulta y así hasta que la primera descarga termine y comience la segunda iteración. El tiempo de demora total para descargar los archivos utilizando el programa en Python es de aproximadamente 2 días y 1 hora, limitado principalmente por el límite diario de 1.000.000 de artículos.

Para evitar baneos de parte del servidor de *New York Times* por hacer más de 1.000 llamadas por una dirección IP, se procedió a usar Tor. Tor dirige el tráfico web a través de una red mundial voluntaria de más de 7.000 computadores para prevenir el rastreo de tráfico de un computador[6]. Esto se incorporó dentro del programa Python, donde cada vez que se utiliza una llave API, el programa activa la red Tor y renueva la IP entregada para no repetir la dirección IP con la misma llave, haciendo más difícil el rastreo de parte del servidor del *New York Times*.

2.3. Creación de Base de Datos

Dado que los artículos se descargan en formato JSON, estos deben ser parseados antes de poder ser utilizados. Para parsearlos se utilizó otro programa en Python que transforma los JSON a un Dataframe (similar a una tabla de datos en formato csv). Básicamente lo que hace el programa es lo siguiente: para un JSON año-mes lee artículo por artículo el archivo (lo que se simplifica ya que cada artículo está separado dentro de la estructura del JSON) y va codificando en un diccionario si lo que leyó fue un *id, abstract, blog* o cualquier otro atributo asociado al diccionario y lo pone como valor en la llave correspondiente. Una vez que leyó esa línea, adjunta el diccionario a una lista. Cuando termina de leer el archivo de texto que contiene el JSON, se obtiene una lista con diccionarios, donde cada diccionario corresponde al artículo con sus atributos. Esto se hace para cada mes de un año, luego se juntan las listas y se transforman a un Dataframe usando el paquete de pandas. Finalmente, el dataframe obtenido se guarda en un archivo csv, equivalente a una tabla de datos, lo que es muy fácil de procesar.

[6] https://www.torproject.org/

2.4. Limpieza de la Base de Datos

Una vez que se obtienen los datos en csv, se procede a limpiar la base de datos. Un aspecto muy importante es que la base de datos contiene todas las noticias del *New York Times* tanto noticias relacionadas a la Economía y Finanzas como no. Este es un problema ya que podría sesgar los resultados obtenidos. Si bien la BBDD contiene la columna *section_name* que contiene la sección a la que pertenece el artículo, esta solo existe para las noticias de 1980 en adelante. Todas las noticias anteriores a ese año no contienen ese atributo, por lo que es imposible saber si son o no relacionadas a Economía y Finanzas. Para solucionar este problema se utilizaron algoritmos de clasificación de machine learning[7]. La razón de esta solución es porque se necesitaban algoritmos capaces de aprender si una noticia es o no es relacionada a finanzas, además de cumplir con las condiciones básicas asociadas a estas técnicas: tener un gran set de datos (todas las noticias desde 1980 hasta 2018) y aprender de manera supervisada (le estamos diciendo al algoritmo que noticia es relacionada a finanzas y cuales no)[8].

Otro argumento a favor de utilizar algoritmos de machine learning es que cuando se reduce la dimensionalidad, mediante técnicas como PCA o t-SNE, para proyectar las noticias en un plano 2-D se observa que las categorías caen en distintas áreas por lo que podemos esperar que la precisión de los algoritmos sea alta:

[7] https://www.cs.princeton.edu/~schapire/talks/picasso-minicourse.pdf

[8] https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9

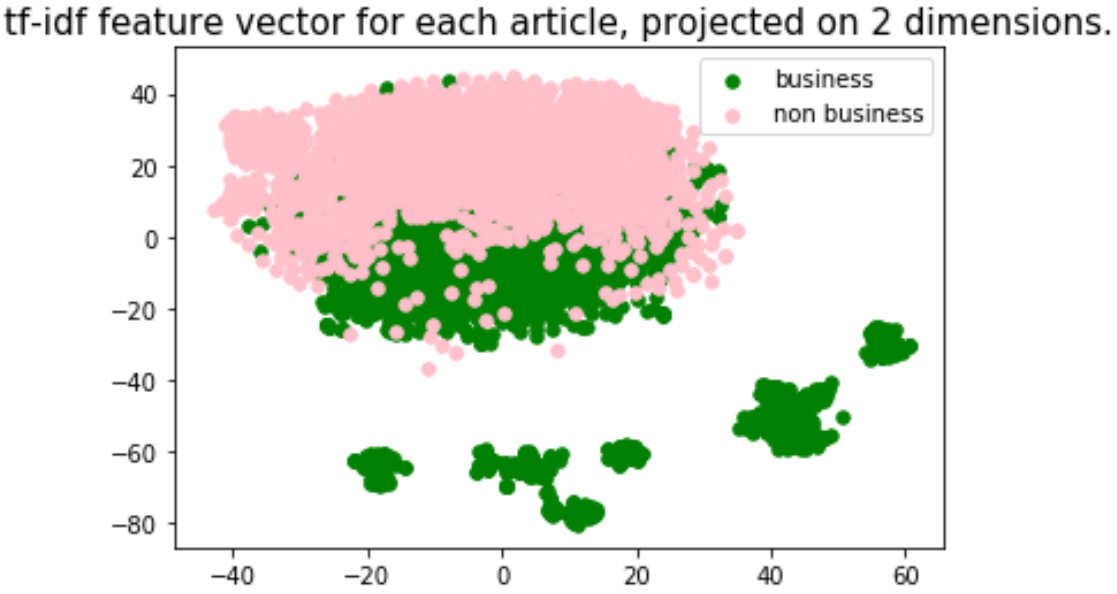

Ilustración 1: Proyección en 2 Dimensiones de t-SNE.

Esto es para mostrar que las noticias relacionadas a finanzas están en un set disjunto de las noticias no relacionadas a finanzas. Esto demuestra que es posible aplicar técnicas de Machine Learning para resolver el problema.

Para decidir sobre cual algoritmo usar (de acuerdo con el teorema de David Wolpert y William Macready cualquier par de algoritmos de optimización son equivalentes cuando su performance es calculada como un promedio sobre todos los posibles problemas)[9], se procedió a probar con 3 tipos de algoritmos: redes neuronales artificiales, regresión logística y descenso de gradiente estocástico. Estos algoritmos vienen programados en Python en paquetes especializados como Tensorflow (paquete de Google que contiene redes neuronales artificiales) y scikit-learn (contiene regresión logística y descenso de gradiente estocástico), lo que facilita el uso de ellos.

Para el uso de redes neuronales, primero hay que entrenarlas. Para esto se separa la data (los artículos de noticias desde 1980 a 2018) en 2 partes: la primera (que se llamará set de entrenamiento) contiene 80% de las noticias (donde 40% son financieras y 40% no son financieras) y la segunda (set de prueba) contiene 20% de las noticias (donde 10% son financieras y 10% no son financieras). Es importante que los datos estén balanceados así se previene que los resultados de la matriz de confusión no estén sesgados. Luego de separar los datos, se crea una red neuronal, usando

[9] Wolpert, D.H., and Macready, W.G. (2005) "Coevolutionary free lunches"

Tensorflow, de dos capas de 500 neuronas la primera capa y 100 neuronas la segunda capa, conectadas entre capas (fig. XX ilustrativa).

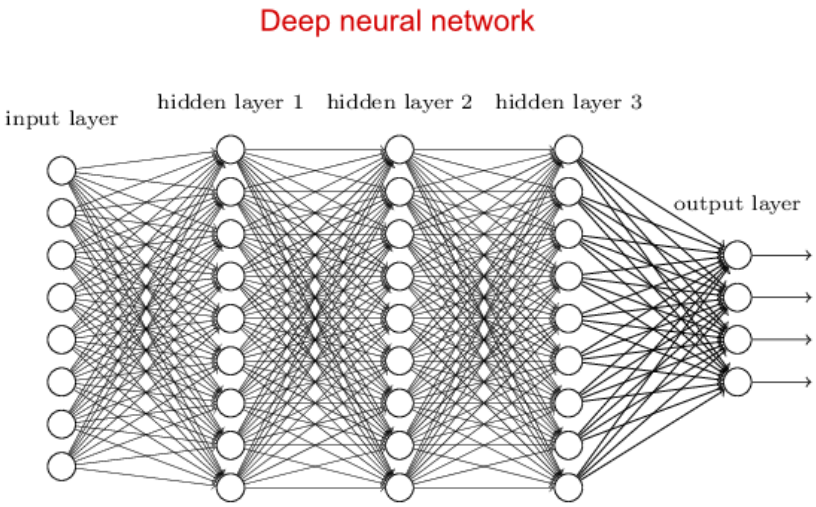

Ilustración 2: Composición de una Red Neuronal.

Estas contienen una capa de inputs que es donde reciben la información, capas ocultas y una capa final de output. Es importante notar que, en una red neuronal profunda, todas las neuronas están conectadas entre capas.

Se escogen 2 capas y esa cantidad de neuronas por capa ya que maximizan la precisión de la red neuronal. Luego de crear la red neuronal se procede a entrenar la red neuronal utilizando el set de entrenamiento. Una vez que se termina de entrenar, se procede a testear en el set de prueba. Además de utilizar una red neuronal de 2 capas, se probó usando la misma red, pero con un módulo pre-entrenado. Esto quiere decir que a la red se le entregan pesos adicionales y se escoge si entrenar solo la red o entrenar a la red y al módulo nuevamente. Se usaron 3 módulos: nnlm-en-dim128, random-nnlm-en-dim128 y universal-sentence-decoder. Los 3 módulos se probaron en la red neuronal con y sin entramiento de módulos. Los resultados obtenidos fueron los siguientes:

- DNN NNLM Module Training 70%
- DNN NNLM Module 68%
- DNN Random NNLM Training 65%
- DNN Random NNLM 60%
- DNN Universal Sentence Decoder 84.2%

- DNN Universal Sentence Decoder Training 80%

Para usar la regresión logística y descenso de gradiente estocástico, primero hay que codificar las palabras en números. La razón de esto es que, a diferencia de las redes neuronales que automáticamente le asignan un peso a cada palabra, las regresiones logísticas y descenso de gradiente estocástico solo reciben números ya que estos son multiplicados por pesos y entregan un output. Para solucionar este problema se utiliza el algoritmo de TF-IDF. El algoritmo de TF-IDF es una estadística numérica que está diseñada para mostrar que tan importante es una palabra para un documento de una colección de documentos o corpus[10]. El valor de tf-idf crece de manera proporcional al número de veces que una palabra aparece en el documento y disminuye por la cantidad de documentos en la que la palabra aparece. En términos simples, si una palabra es técnica, va a aparecer en un documento muchas veces y en pocos documentos por lo que tiene mucha relevancia y se le asigna un puntaje tf-idf elevado, mientras que una palabra común como un artículo (el o la) aparece en muchos documentos muchas veces por lo que se le asigna un puntaje bajo. La fórmula para calcular el tf-idf es la siguiente:

$$TF - IDF_{i,j} = \frac{n_{i,j}}{\sum_{i \in j} n_{i,j}} \cdot \log\left(\frac{N}{df_i}\right)$$

Donde:

$$n_{i,j} = cantidad\ de\ veces\ que\ aparece\ la\ palabra\ i\ en\ documento\ j$$

$$df_i = número\ de\ documentos\ que\ contienen\ i$$

$$N = número\ total\ de\ documentos$$

Luego que se calcula la matriz de tf-idf, separo los artículos de noticias en 2 muestras: primera muestra (set de entrenamiento) con el 67% de los artículos (33,5% financieros y 33,5% no financieros) y segunda muestra (set de prueba) con el 33% de los artículos restantes (16,5% financieros y 16,5% no financieros). Luego de separar los datos se procede a crear 2 modelos: una regresión logística y un descenso de gradiente

[10] Rajaraman, A.; Ullman, J.D. (2011). "Data Mining"

estocástico usando el paquete de scikit-learn. Ambos se entrenan con el set de entrenamiento y luego se prueban en el set de muestra.

Los resultados de precisión obtenidos son:

- Stochastic Gradient Descent 92.0%
- Logistic Regression 92.5%

Si se comparan todos los modelos, se obtiene que el mejor modelo es la regresión logística, por lo que es el modelo que se escoge para clasificar los artículos. Para poder clasificar los artículos, se guarda la regresión logística con sus pesos (que corresponde al aprendizaje que tuvo con el 66% de artículos) y se le entregan las noticias desde 1851 hasta 1979. Para cada noticia da un output de 0 o 1 donde 0 es que el artículo no es financiero y 1 si lo es. La probabilidad de que cada título esté bien clasificado es de un 92,5%.

2.5. Cálculo del Sentimiento

Una vez que se clasifican todos los artículos en categorías financieras como no financieras, se procede a separar la BBDD en 2 partes: artículos relacionados a economía y finanzas y artículos no relacionados a economía y finanzas. Para efectos de esta investigación, solo se utilizan los artículos relacionados a economía y finanzas. Luego se procede a calcular el sentimiento. Para calcular el sentimiento, se asegura que todos los artículos no tengan faltas de ortografía (si las tienen, a la hora de comparar las palabras el algoritmo no las va a reconocer). Para corregir los artículos, se utiliza el paquete SymSpell implementando en Python. Una vez que se utiliza SymSpell por cada artículo, se procede a calcular el sentimiento por cada artículo. De acuerdo con Loughran-McDonald (2011), para calcular el sentimiento se utiliza la siguiente ecuación:

$$s_j = \frac{POS_j - NEG_j}{a_j}$$

Donde POS_j, NEG_j y a_j son la cantidad de palabras positivas, negativas y totales en el artículo. Para clasificar los artículos, se utiliza la lista de palabras de Loughran-McDonald (2011). Si bien esta lista es una de las más usadas en finanzas, tiene problemas conocidos (Loughran-McDonald (2016)):

- **Problema de la doble negación**: "*…la información negativa generalmente se utiliza acompañada de palabras positivas por lo que no tener esto en cuenta hace que el sentimiento obtenido sea positivo y por lo tanto ambiguo*". Por ejemplo: los resultados no son buenos (bueno es positivo, pero él no hace que la oración tenga una connotación negativa.

- **Problema de los pesos**: todas las palabras no tienen la misma positividad o negatividad (bueno v/s excelente o malo v/s paupérrimo).

- **Problema de 2 categorías**: los diccionarios solo tienen palabras positivas o negativas, pero algunas palabras no quedan clasificadas en ninguna de las 2 categorías.

2.6. Resolviendo la Doble Negación

Para resolver el problema de la doble negación se procede a utilizar el algoritmo de Standford Dependencies. Este algoritmo permite entender la relación gramatical entre palabras en una oración[11]. En otras palabras, con este algoritmo es posible ver a que palabras de la oración está modificando una palabra negativa. Por ejemplo, consideremos la oración en inglés "*I do not hate my enemies*". Es fácil ver que la palabra *not* modifica todo lo que viene después. El algoritmo de Standford Dependencies descompone la oración de la siguiente manera:

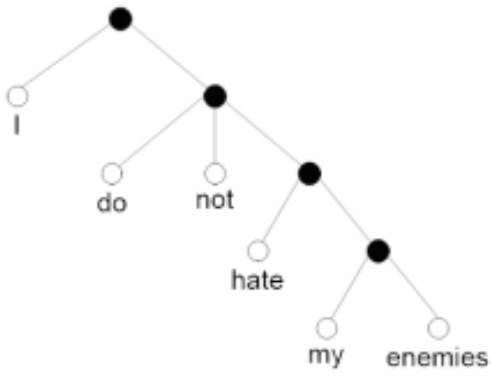

Ilustración 3: Funcionamiento del Algoritmo de Standford Dependencies.

Además de identificar sujeto y predicado, entre otras cosas, es importante notar que este algoritmo permite identificar las palabras que están siendo modificadas por las palabras negativas.

Entonces, como *not* antecede la siguiente rama, es posible entender cuáles palabras están siendo modificadas por *not* (cambiando de un sentimiento negativo a positivo o positivo a negativo). Como Standford Dependencies permite entender la relación, es intuitivo cómo funcionará el algoritmo para calcular el sentimiento:

1. Calcular sentimiento de una oración por palabra y almacenarlo en un vector (donde cada valor corresponde al sentimiento de esa palabra de la oración)

[11] https://nlp.stanford.edu/software/stanford-dependencies.html

2. Chequear si existen palabras modificadoras de sentimiento[12]. En caso de que no existan saltar al paso 5

3. Utilizar el algoritmo de Standford Dependencies para entender la relación entre modificadores y palabras de la oración

4. Multiplicar por -1 todas las palabras que dependan de las palabras modificadoras de sentimiento

5. Sumar todos los valores del vector y dividirlo por el largo del vector (emulando la fórmula de Loughran-McDonald (2011))

2.7. Resolviendo el Problema de los Pesos

Para resolver el problema de los pesos se recurre a la definición de IDF. IDF significa *Inverse Document Frequency*, lo que se traduce a frecuencia inversa de documento, o sea es la medida numérica que expresa cuán común es un término en una colección de documentos. Esta se calcula mediante la siguiente fórmula:

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

Donde N es el número total de documentos en la colección $|\{d \in D : t \in d\}|$ es el número de documentos donde el término T aparece. Básicamente IDF premia con un mayor puntaje a palabras menos comunes y castiga a palabras más comunes (entre documentos). Intuitivamente tiene sentido ya que, si siempre se ocupara la palabra "bueno", está deja de ser tan positiva como la palabra "excelente" que es menos común.

Por otro lado, es importante controlar por el efecto de cambio de uso de palabras en la prensa. Por lo tanto, se separa la BBDD en 3 partes: desde 1851 hasta 1900, desde 1901 hasta 1951 y desde 1951 hasta 2019. Luego se procede a calcular esta medida, utilizando un programa en Python. El algoritmo que utiliza para cada BBDD es el siguiente:

[12] Revisar anexo 1

1. A todos los artículos se le saca la puntuación y caracteres especiales

2. Se transforman las contracciones a las palabras correctas

3. Luego se crea un diccionario que contiene todas las palabras de los artículos

4. Cada vez que se encuentra la palabra en un artículo, se suma 1 a su valor en el diccionario

5. Después de terminar los artículos se cuenta con un diccionario que tiene como llaves todas las palabras encontradas en los artículos y como valores la cantidad de documentos donde aparece

6. Finalmente se calcula IDF solo utilizando las palabras del diccionario de Loughran-McDonald y usando la fórmula descrita anteriormente

2.8. Resolviendo el Problema de 2 Categorías

Para resolver este problema se utiliza VADER. VADER es un acrónimo anglosajón que significa *Valence Aware Dictionary and sEntiment Reasoner*. Básicamente es un algoritmo que combina un diccionario de palabras con puntuación de sentimiento, es decir existe una jerarquía entre palabras positivas y entre palabras negativas. Además, incluye manejo de doble negación y reconocimiento de signos de puntuación y acrónimos[13]. Para utilizar este algoritmo, se usa un paquete para Python llamado vaderSentiment. Este paquete recibe como input un texto y otorga como resultado un vector con 4 números:

- Puntaje positivo: porción del texto que es positivo

- Puntaje neutral: porción del texto que es neutral

- Puntaje negativo: porción del texto que es negativo

- Puntaje compuesto: puntaje compuesto estandarizado entre -1 y 1 donde -1 es lo más negativo y 1 es lo más positivo

[13] http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

Para capturar el efecto de cada solución propuesta, se cuentan con 4 métricas de sentimiento:

- Loughran-McDonald (2011)
- Doble Negación
- Diccionario con Pesos
- VADER

2.9. Calculando Tópicos de Calomiris – Mamaysky

Dado que se quiere buscar la existencia de una relación entre el sentimiento del mercado y crisis crediticios, se propone encontrar un proxy del sentimiento de mercado crediticio. Para esto, Calomiris-Mamaysky (2018), proponen que el sentimiento de mercado se puede separar en 5 tópicos: mercado, gobierno, corporativo, commodities y crédito. Cada tópico se define como una lista de palabras, relacionadas con economía y finanzas, que no se solapa. Para separar en estos 5 tópicos, se utiliza el siguiente índice:

$$f_{\tau,j} = \frac{e_{\tau,j}}{e_j}$$

Donde $e_{\tau,j}$ corresponde a la cantidad de palabras económicas que son del tema τ y e_j corresponde a la cantidad de palabras económicas que hay en el artículo. Por lo tanto, está fracción se calcula para cada uno de los 5 tópicos. Para calcular esté índice, se utiliza un programa en Python que chequea si cada palabra del artículo está presente en alguno de los tópicos. Luego esto lo guarda en una lista, suma por cada tópico y lo divide por el total de palabras económicas. Finalmente, para extrapolar el sentimiento de mercado a un sentimiento por tópico se utiliza la siguiente ecuación:

$$s_{\tau,j} = f_{\tau,j} \cdot s_j$$

Donde s_j es el sentimiento para el artículo j. Dado que hay más de un artículo por día, hay que agregar el sentimiento de todos los artículos. Para agregar la data a un nivel diario se utiliza la siguiente ecuación:

$$s_\tau = \sum_j \frac{a_j}{a} \cdot s_{\tau,j}$$

Donde a_j es el número total de palabras mencionadas en el artículo j y a es el número total de palabras mencionados en todos los artículos para un determinado día.

2.10. Calculando Information Overload

Dado que se quiere buscar la existencia de una relación entre la sobrecarga de información y los retornos del mercado accionario, se propone encontrar un proxy de la sobrecarga de información. Para efectos de esta tesis se calcula lo siguiente:

$$Information\ Overload = IO_t = \frac{monthly\_news_{y,m}}{yearly\_news_{y-1}}$$

Donde $y = año$, $m = mes$, por lo que el ratio se calcula como el número de noticias publicadas en un mes dividido por el número de noticias publicadas el año pasado. Esto indica la proporción de noticias relacionadas a finanzas que se publicaron este mes en comparación al año anterior.

Por otro lado, de manera análoga, se incorpora un proxy que explique el volumen transado. Para esto se calcula lo siguiente:

$$Volumen\ Overload = V_t = \frac{monthly\_volume_{y,m}}{yearly\_volume_{y-1}}$$

Donde $y = año$, $m = mes$, por lo que el ratio se calcula como la cantidad total de acciones transadas (acciones totales del día por precio de cierre) en el mercado estadounidense en un mes dividido por el número de acciones transadas en el mercado estadounidense el año anterior. Esto indica la proporción del volumen de acciones que se transaron en un mes en comparación al año anterior.

3. Relación Medios y Retornos

3.1. Análisis Descriptivo

Para comenzar el análisis descriptivo de la data se realizó el conteo del número de noticias publicadas anualmente, el cual se puede ver en la ilustración 4. Estas noticias fueron extraídas por medio de la API de New York Times, previamente explicado en sección 1. Datos. De este gráfico se desprende como la cantidad de noticias presenta una tendencia al aumento con el paso del tiempo. Esto intuitivamente hace mucho sentido ya que con el paso del tiempo ha aumentado la velocidad de comunicación tanto entre países como dentro del país por lo que se recibe información que antes no se recibía, lo que permite que se publiquen noticias más variadas que en la antigüedad. Por otro lado, se observan peaks claros en el gráfico.

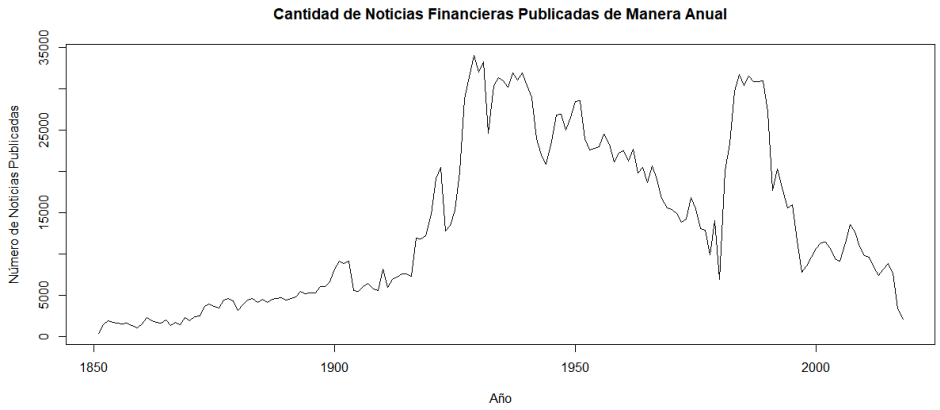

Ilustración 4: Cantidad de Noticias Financieras Publicadas de Manera Anual.

De la misma forma se estudia el comportamiento del ratio “Information Overload”. En la ilustración 5, se observa como el ratio sigue una tendencia en el tiempo alrededor de un valor, notando claros peaks y valles en que se escapan de la tendencia. Además, es importante notar que la serie no es muy ruidosa. Esto es esperable ya que el diario cuenta con secciones específicas ligadas a Economía y Finanzas publicadas de manera diaria.

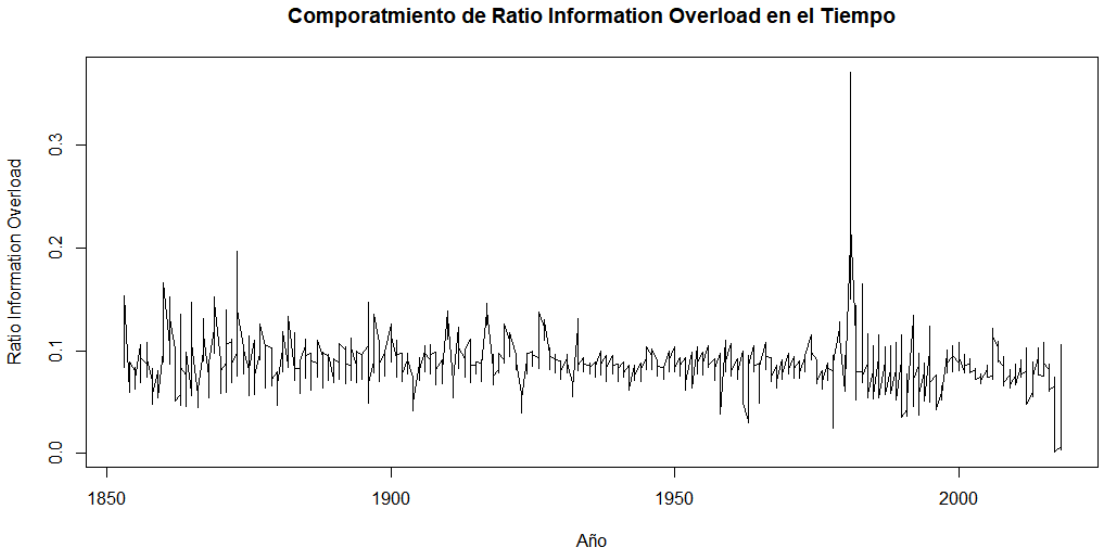

Ilustración 5: Comportamiento de Ratio “Information Overload” en el Tiempo.

Por último, se estudia el comportamiento del ratio “Volume Overload”. En la ilustración 6, se observa como el ratio sigue una tendencia en el tiempo alrededor de un valor. A diferencia del ratio “Information Overload”, este ratio es un poco más volátil y es más sucio. Intuitivamente tiene sentido ya que se esperaría que el volumen transado distribuya con kurtosis mayor a 3, así se presentaría el fenómeno de “fat tails”, lo que modelaría correctamente eventos de muy altos o bajos retornos (e.g.: earnings surprise).

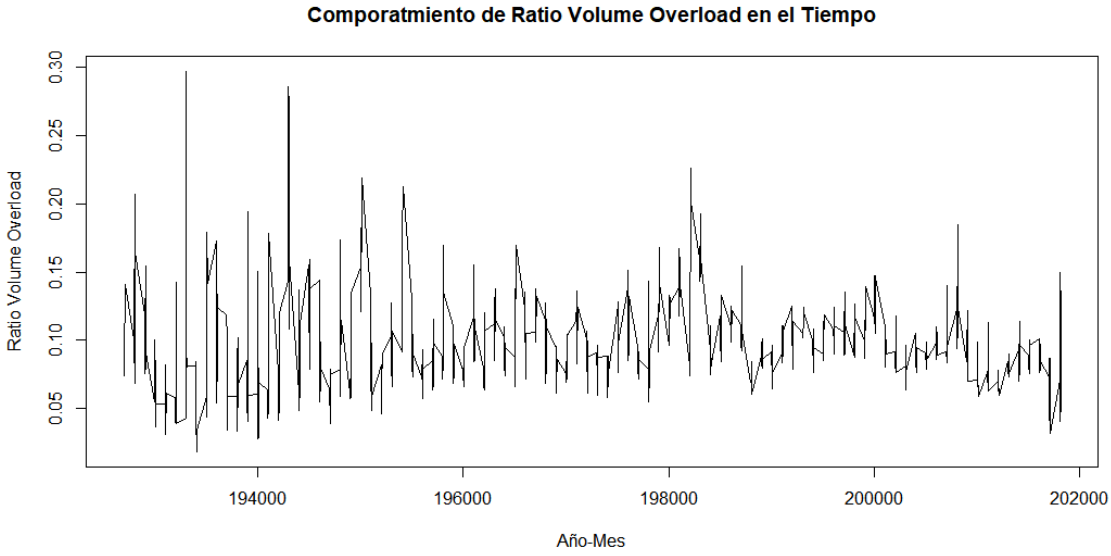

Ilustración 6: Comportamiento de Ratio “Volumen Overload” en el Tiempo.

Variables Medios	Media	Mediana	Cuantil 25%	Cuantil 75%	Desviación Estándar
Panel A: Toda la data					
Positivo	0.81%	0.78%	0.53%	1.04%	0.54%
Negativo	2.40%	2.29%	1.77%	2.91%	1.24%
Pesimismo	-1.59%	-1.47%	-2.15%	-0.89%	1.37%
Panel B: Recesiones					
Positivo	0.82%	0.78%	0.50%	1.09%	0.57%
Negativo	2.59%	2.46%	1.88%	3.14%	1.32%
Pesimismo	-1.76%	-1.63%	-2.40%	-0.97%	1.47%
Panel C: Expansiones					
Positivo	0.81%	0.77%	0.54%	1.03%	0.53%
Negativo	2.32%	2.23%	1.72%	2.81%	1.19%
Pesimismo	-1.51%	-1.41%	-2.05%	-0.86%	1.31%

Tabla 1: Estadística Descriptiva para Variables Medios durante Recesiones y Expansiones Utilizando Diccionario Loughran McDonald

Variables Medios	Media	Mediana	Cuantil 25%	Cuantil 75%	Desviación Estándar
Panel A: Toda la data					
Positivo	0.84%	0.81%	0.56%	1.08%	0.55%
Negativo	2.37%	2.26%	1.74%	2.88%	1.23%
Pesimismo	-1.53%	-1.41%	-2.07%	-0.83%	1.36%
Panel B: Recesiones					
Positivo	0.85%	0.81%	0.53%	1.12%	0.58%
Negativo	2.56%	2.43%	1.85%	3.10%	1.32%
Pesimismo	-1.71%	-1.57%	-2.33%	-0.92%	1.47%
Panel C: Expansiones					
Positivo	0.84%	0.81%	0.58%	1.06%	0.54%
Negativo	2.29%	2.20%	1.70%	2.77%	1.19%
Pesimismo	-1.45%	-1.35%	-1.97%	-0.80%	1.30%

Tabla 2: Estadística Descriptiva para Variables Medios durante Recesiones y Expansiones Utilizando Diccioanrio de Loughran McDonald y Negation Handling

Las tablas 1 y 2 muestran estadística descriptiva para variables de medios que se utilizan en esta memoria. Estas medidas se construyen de la siguiente manera:

- Positivo: cantidad de palabras positivas por día dividido en el total de palabras de ese día
- Negativo: cantidad de palabras negativas por día dividido en el total de palabras de ese día
- Pesimismo: diferencia entre positivo y negativo calculado de manera diaria

El panel A presenta las estadísticas calculadas para la muestra completa que considera 60,064 días. Por otro lado, los paneles B y C dividen la muestra por ciclos económicos. El Panel B considera todos los días que hubo recesión entre 1851 y 2018 (18,259), mientras que el Panel C considera todos los días donde hubo expansión (41,805).

Al observas los paneles B y C de ambas tablas, se observa que la medida positiva es casi similar tanto en expansiones como recesiones, mientras que la medida negativa es mucho mayor en recesiones que en expansiones.

La diferencia entre las tablas 1 y 2 es casi marginal. Esto se debe a que la cantidad de palabras que deben ser cambiadas de positivas a negativas y viceversa por doble negación es muy pequeña al ser comparada con el total (0.06%).

3.2. Desarrollo de Hipótesis

2.2.1. Hipótesis 1: Relación Information Overload, Volumen y Retornos

En el desarrollo de esta hipótesis se busca poder caracterizar la existencia de alguna relación entre los retornos del mercado estadounidense (caracterizado por el S&P500) y los ratios “Information Overload” y “Volume Overload”. El objetivo de esta hipótesis es entender el efecto de la sobrecarga de información en el comportamiento de los inversionistas. Se espera que la cobertura de los medios reduzca la asimetría de información entre el inversionista promedio y participantes informados del mercado, por lo tanto, aumentando la eficiencia del mercado. A través de distintas disciplinas se ha encontrado que la calidad de la decisión está correlacionada de manera positiva con la cantidad de información que reciben. Si bien es bueno el aumento de información, existe un límite relacionado a la cantidad de información que puede procesar un inversionista, donde sobre ese umbral la información deja de ser útil. En otras palabras, si un inversionista recibe mucha información, entonces la precisión de su decisión disminuye rápidamente.

La base teórica de teórica de esta idea viene de psicólogos y científicos cognitivos (Miller (1956), Schroder et al. (1967) y Simon y Newell (1971)). Por ejemplo, en el modelo de Schroder et al (1967), el rendimiento de un individuo mejora mientras más información es recibida. Sin embargo, cuando la información sobrepasa la capacidad de procesar la información, la información adicional confunde al individuo y, por ende, la calidad de la toma de decisión. Es por esto por lo que se espera que, en periodos de sobrecarga de información, inversionistas van a actuar en el mercado considerando información irrelevante o desechando información que es importante. Por lo tanto, se espera que cuando la información sea excesiva, los inversionistas intercambien más,

pero que esas decisiones, en vez de proveer información eficiente en los precios, produzcan cambios drásticos aumentando la volatilidad.

Para validar esta hipótesis en tiempo diario, se utiliza la siguiente regresión:

$$r_{m_t} = \beta_0 + \beta_1 \cdot V_t + \beta_2 \cdot IO_t$$

Donde V_t es "Volume Overload" e IO_t es "Information Overload" y se utiliza un tiempo diario. Los resultados obtenidos son los siguientes:

	Daily Returns	
	β	$t-stat$
V_t	0.18	4.77
IO_t	0.13	2.42
$Intercept$	-0.00	-3.98
$\boldsymbol{R^2_{adj}}$	**0.001**	

Tabla 3: Resultados Regresión Variable Dependiente "Market Daily Returns" y Variables Independientes "Volume Overload" e "Information Overload"

Se observa que los retornos de mercado diarios están correlacionados con Volume e Information Overload. Además, se ve que, a mayor cantidad de información, mayores son los retornos de mercado (comportamiento análogo para Volume Overload). El problema de esta variable es que no permite separar sobrecarga de información de bajo carga de información. Para poder hacer esto, se crean 2 nuevas variables: IO_t^+ (sobrecarga de información) e IO_t^- (bajo carga de información). Estas variables se calculan de la siguiente manera:

$$IO_t^+ = \begin{cases} N_t - \tau_t & si\ N_t \geq \tau_t \\ 0 & si\ no \end{cases}$$

$$IO_t^- = \begin{cases} -N_t + \tau_t & si\ N_t \leq \tau_t \\ 0 & si\ no \end{cases}$$

Para poder calcular la tendencia se utiliza un filtro de un lado de Hodrick y Prescott (1997)[14]. Luego se procede a calcular la siguiente regresión:

$$r_{m_t} = \beta_0 + \beta_1 \cdot V_t + \beta_2 \cdot IO_t^+ + \beta_3 \cdot IO_t^-$$

	Daily Returns	
	β	$t-stat$
V_t	0.19	5.28
IO_t^+	0.00	0.48
IO_t^-	0.01	4.09
$Intercept$	-0.00	-2.43
$\boldsymbol{R^2_{adj}}$	**0.002**	

Tabla 4: Resultados Regresión Variable Dependiente "Market Daily Returns" y Variables Independientes "Volume Overload", "Information Overload" e "Information Underload"

Se observa que, de manera diaria, al separar la sobrecarga de la bajo carga de información, la bajo carga de información es la variable que explica los retornos diarios ya que la sobrecarga de información pierde significancia. Este fenómeno está relacionado con el fallo del proceso de "Price Discovery" por falta de información.

Por otro lado, se procede a estudiar el fenómeno a tiempo mensual, utilizando la siguiente regresión:

$$V_t = \beta_0 + \beta_1 \cdot IO_t$$

Donde V_t es "Volume Overload" e IO_t es "Information Overload" y se utiliza un tiempo mensual. Los resultados obtenidos son los siguientes:

	Volume Overload	
	β	$t-stat$
IO_t	0.12	2.42

[14] Para más información revisar anexo 2 y 3

$Intercept$	0.09	20.76
$\boldsymbol{R^2_{adj}}$	**0.004**	

Tabla 5: Resultados Regresión Variable Dependiente "Volume Overload" y Variable Independiente "Information Overload"

Se observa que "Information Overload" y "Volume Overload" son variables correlacionadas. Además, se ve que, a mayor cantidad de información, mayor es el volumen transado en comparación al año pasado. El problema de esta variable es que no permite separar sobrecarga de información de bajo carga de información. Para poder hacer esto, se utilizan las variables: IO_t^+ e IO_t^-.

Luego se procede a calcular la siguiente regresión:

$$V_t = \beta_0 + \beta_1 \cdot IO_t^+ + \beta_2 \cdot IO_t^-$$

	Volume Overload	
	β	$t-stat$
IO_t^+	0.21	3.00
IO_t^-	-0.01	-0.08
$Intercept$	0.09	65.03
$\boldsymbol{R^2_{adj}}$	**0.006**	

Tabla 6: Resultados Regresión Variable Dependiente "Volume Overload" y Variables Independientes "Information Overload" e "Information Underload"

De la tabla 6 se observa que la sobrecarga de información es la variable que explica el aumento de volumen, mientras que la bajo carga de información no afecta en nada a la variable de volumen. Para poder entender el mecanismo de acción sobre los precios se procede a calcular la siguiente regresión:

$$r_{m_t} = \beta_0 + \beta_1 \cdot V_t$$

	Monthly Returns	
	β	$t-stat$
V_t	0.27	6.05
$Intercept$	-0.02	-3.56

R^2_{adj}	**0.03**

Tabla 7: Resultados Regresión Variable Dependiente “Market Monthly Returns” y Variable Independiente “Volume Overload”

De la tabla 7 se observa que la sobrecarga de volumen está directamente relacionada con los retornos mensuales. Esto tiene sentido ya que es consistente con la anomalía de momentum (Asness et al. (2013)).

2.2.2. Hipótesis 2: Relación Sentimiento de Mercado, Volumen y Retornos

En el desarrollo de esta hipótesis se busca poder caracterizar la existencia de alguna relación entre los retornos del mercado estadounidense (caracterizado por el S&P500) y los ratios “Market Sentiment” y “Volume Overload”. El objetivo de esta hipótesis es entender el efecto del sentimiento de mercado sobre los retornos. Se espera que, mientras mayor sea el sentimiento de mercado en valor absoluto ($|Market\ Sentiment| \approx 1$) mayor sea el exceso de retorno. Esta hipótesis se basa en el paper de García (2013), donde el sentimiento de mercado (calculado a partir de dos columnas del New York Times) es capaz de predecir los retornos diarios, particularmente durante recesiones.

Para la validación de esta hipótesis se decidió utilizar un modelo de regresión lineal, el cual se presenta en la ecuación siguiente:

$$r_{m,t} = \beta_0 + \beta_1 \cdot M_t + \beta_2 \cdot V_t$$

Donde $M_t = Market\ Sentiment$ y $V_t = Volume\ Overload$ ambas en tiempo t. De manera adicional, la agregación utilizada en los datos es de manera mensual. Esto viene de la hipótesis que los datos diarios son muy ruidosos vs los agregados de manera mensual y anual (García 2013). Para poder estudiar como afecta la manera en que se calcula el sentimiento de mercado, se utilizan 4 métodos:

1. Loughran-McDonald (LM)
2. Loughran-McDonald Negate Handling (LM NH)
3. Loughran-McDonald Negate Handling idf (LM NH IDF)

4. Vader

Como resultado se obtiene lo siguiente:

	LM		**LM NH**		**LM NH IDF**		**Vader**	
	β	$t-stat$	β	$t-stat$	β	$t-stat$	β	$t-stat$
V_t	0.225	5.04	0.223	5	0.24	5.43	0.23	5.30
M_t	1.544	4.06	1.614	4.12	3.50	2.79	0.08	3.76
$Intercept$	0.012	1.45	0.012	1.47	0.00	0.28	-0.03	-5.12
$\boldsymbol{R^2_{adj}}$	**0.04**		**0.05**		**0.04**		**0.04**	

Tabla 8: Resultados Regresiones Variable Dependiente "Market Monthly Returns" y Variables Independientes "Volume Overload", "Market Sentiment (4 métodos)"

De la tabla 8 se observa que ambas variables, "Market Sentiment" y "Volume Overload" son significativas y capaces de explicar el exceso de retorno. Por otro lado, se observa que el mejor R ajustado es utilizando el diccionario de Loughran -McDonald corregido con el manejo de negación. Esto es interesante ya que indica que para explicar los retornos en el mediano plazo (a nivel mensual) los agentes no hacen diferencias entre palabras positivas y entre palabras negativas. Otra cosa que llama la atención es que el diccionario de Loughran-McDonald es un diccionario que fue construido a partir de los artículos 10-k que publican todas las empresas que son transadas en Estados Unidos, en otras palabras, es un diccionario que es puramente financiero. Al utilizar un diccionario que no es construido de documentos financieros (método Vader) se obtienen resultados comparables con el diccionario especializado. Esto es importante ya que indica que las noticias no necesariamente utilizan palabras técnicas de finanzas, están escritas para que un inversionista común las pueda leer.

Si bien se obtiene que ambas variables son negativas, se procede a utilizar variables de control para ver si no están explicando algo que otras variables conocidas en finanzas explican. Las variables de control que se utilizan son:

- Ratios de valuación: precio/dividendo (P_t/D_t) y precio/ganancias (P_t/E_t) han sido reconocidos como predictores de los retornos de mercado (Campbell y Shiller (1988), Fama y French (1988), Campbell y Yogo (2006) entre otros)

- Consumo sobre riqueza: Lettau y Ludvigson (2001) muestran que fluctuaciones sobre el ratio consumo sobre riqueza (CAY_t) son fuertes predictores de exceso de retorno del mercado. Esta variable ha sido incluida en diferentes estudios: Guo

(2006), Bollerslev et al. (2009) y Pollet y Wilson (2010). Esto se obtiene de la página web de Lettau[15].

- Default spread, term spread y tasa libre de riesgo: la variación en el retorno del mercado de acciones también es explicada por variables que se utilizan en los retornos de bonos (Keim y Stambaugh (1986), Campbell (1987), Fama y French (1989)). Siguiendo estos estudios se incluyen el default spread (DS_t), term spread (TS_t) y tasa relativa libre de riesgo ($RREL_t$). DS_t es la diferencia entre los bonos corporativos BAA y AAA de la Reserva Federal de St. Louis, TS_t es la diferencia entre el bono del tesoro de 10 años y 3 meses de la Reserva Federal de St. Louis y $RREL_t$ se construye como el bono del tesoro de 1 mes menos su promedio móvil de los últimos 12 meses.

- Inflación: Lintner (1975), Fama y Schwert (1977) y Fama (1981) muestran que la inflación tiene poder predictivo sobre el premio de mercado. Se utiliza el Índice de Precios al Consumidor (CPI), obtenido de "Bureau of Labor Statistics" y se calcula la inflación la diferencia del logaritmo de dos meses consecutivos.

- Liquidez: Amihud (2002) provee evidencia que la liquidez del mercado tiene suficiente poder predictivo sobre los retornos de mercado. Por lo que se incluye la medida de iliquidez de mercado de Amihud, LIQ_t, que es calculada como el ratio del retorno absoluto del mercado sobre su volumen

- Varianza y correlación: siguiendo a Guo (2006) se incluye la volatilidad de los retornos (calculadas mediante un GARCH) y la correlación de los 25 portafolios de Fama y French (1992) construidos por tamaño y valor libro.

Para poder estudiar el efecto de las variables de control se utiliza la siguiente ecuación:

$$r_{m_{t+h}} = V_t + M_t + \ controls + \epsilon_t$$

Al utilizar las variables de control contra exceso de retorno se obtiene lo siguiente:

Horizon	**LM**	**LM NH**	**LM NH IDF**

[15] http://faculty.haas.berkeley.edu/lettau/data.html

(h=0)	I	II	III	I	II	III	I	II	III
$\log\left(\frac{P_t}{D_t}\right)$	-0.00			-0.00			-0.00		
	(-0.41)			(-0.58)			(-0.76)		
$\log\left(\frac{P_t}{E_t}\right)$	0.00			0.00			0.00		
	(0.46)			(0.52)			(0.48)		
CAY_t		-0.29			-0.31			-0.35	
		(-1.25)			(-1.35)			(-1.40)	
DS_t		-0.00			-0.00			-0.01	
		(-0.49)			(-0.40)			(-0.90)	
TS_t		-0.00			-0.00			-0.00	
		(-0.26)			(-0.31)			(-0.32)	
$RREL_t$		0.00			0.00			0.00	
		(0.26)			(0.16)			(0.33)	
$INFL_t$		0.00			0.00			0.00	
		(0.16)			(0.14)			(0.16)	
LIQ_t			-209.22			-209.64			-205.58
			(-2.61)			(-2.62)			(-2.56)
IV_t			0.26			0.26			0.25
			(2.77)			(2.81)			(2.74)
IC_t			-0.09			-0.09			-0.09
			(-7.49)			(-7.48)			(-7.92)
V_t	0.23	-0.60	0.24	0.23	-0.60	0.24	0.25	-0.62	0.25
	(4.96)	(-3.56)	(5.18)	(4.98)	(-3.58)	(5.13)	(5.46)	(-3.65)	(5.42)
M_t	1.55	1.78	0.58	1.61	1.97	0.70	3.46	5.29	1.08
	(3.95)	(1.54)	(1.46)	(4.04)	(1.69)	(1.70)	(2.72)	(1.34)	(0.84)
$Intercept$	0.01	0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.05
	(0.87)	(3.29)	(4.77)	(1.00)	(3.39)	(4.87)	(0.52)	(3.11)	(4.53)
$\boldsymbol{R^2_{adj}}$	**0.04**	**0.09**	**0.09**	**0.04**	**0.10**	**0.09**	**0.04**	**0.09**	**0.09**

Tabla 9: Resultados Regresiones Variable Dependiente “Market Monthly Returns” y Variable Independiente “Volume Overload”, “Market Sentiment (3 métodos)” y Variables de Control.

	Vader		
	I	II	III
$\log\left(\frac{P_t}{D_t}\right)$	-0.01		
	(-1.94)		
$\log\left(\frac{P_t}{E_t}\right)$	0.01		
	(1.03)		
CAY_t		-0.36	
		(-1.52)	
DS_t		-0.01	
		(-0.87)	
TS_t		0.00	
		(0.20)	
$RREL_t$		-0.00	
		(-0.12)	
$INFL_t$		0.00	
		(0.34)	
LIQ_t			-209.25

			(-2.62)
VR_t			0.28
			(2.98)
IC_t			-0.09
			(-8.08)
V_t	0.25	-0.64	0.23
	(5.57)	(-3.78)	(5.08)
M_t	0.09	0.09	0.07
	(4.15)	(1.92)	(3.14)
$Intercept$	-0.01	0.05	0.03
	(-1.12)	(1.96)	(2.89)
$\boldsymbol{R^2_{adj}}$	**0.04**	**0.10**	**0.10**

Tabla 10: Resultados Regresiones Variable Dependiente "Market Monthly Returns" y Variable Independiente "Volume Overload", "Market Sentiment (método Vader)" y Variables de Control.

De las tablas 9 y 10 se observa que el efecto del sentimiento de mercado, calculado mediante los métodos de Loughran McDonald y Loughran McDonald NH IDF es explicado por las otras variables de control, mientras que Loughran McDonald NH y Vader son significativos siempre además de entregar mejores R^2_{adj}. Esto significa que lo anterior es correcto: los inversionistas no dimensionan entre palabras positivas y entre palabras negativas y en el uso del diccionario de Loughran McDonald hay que corregir por doble negación.

Ahora, se procede a estudiar el efecto predictivo que tienen tanto las variables de volumen y sentimiento de mercado como las variables de control sobre los retornos de mercado. Para hacer esto, se utiliza la ecuación anterior, pero con $h = 1$ y sin las variables de control dado que se desea estudiar el efecto predictivo a un mes en el futuro. Al realizar esta regresión se obtienen los siguientes resultados:

	LM		**LM NH**		**LM NH IDF**		**Vader**	
	β	$t-stat$	β	$t-stat$	β	$t-stat$	β	$t-stat$
V_t	0.38	8.75	0.38	8.72	2.80	2.29	0.40	9.12
M_t	1.25	3.36	1.30	3.38	0.40	9.11	0.05	2.42
$Intercept$	-0.01	-0.95	-0.01	-0.97	-0.02	-2.02	-0.04	-6.98
$\boldsymbol{R^2_{adj}}$	**0.09**		**0.09**		**0.08**		**0.08**	

Tabla 11: Resultados Regresiones Variable Dependiente "Market Monthly Returns" de 1 mes en el futuro y Variables Independientes "Volume Overload", "Market Sentiment (4 métodos)"

De la tabla 11 se observa que ambas variables, “Market Sentiment” y “Volume Overload” son significativas y capaces de explicar el exceso de retorno futuro.[16] Además, su comportamiento de signo sigue siendo el mismo a 1 mes que en el corte transversal de datos. Como las variables son significativas, se procede a agregar las variables de control. Al agregarlas se obtiene lo siguiente:

Horizon	**LM**			**LM NH**			**LM NH IDF**		
(h=1)	**I**	**II**	**III**	**I**	**II**	**III**	**I**	**II**	**III**
$\log\left(\frac{P_t}{D_t}\right)$	-0.00			-0.00			-0.00		
	(-0.15)			(-0.31)			(-0.47)		
$\log\left(\frac{P_t}{E_t}\right)$	0.01			0.01			0.01		
	(1.47)			(1.52)			(1.48)		
CAY_t		-0.45			-0.46			-0.50	
		(-1.98)			(-2.01)			(-2.01)	
DS_t		-0.01			-0.01			-0.01	
		(-1.22)			(-1.18)			(-1.50)	
TS_t		0.00			0.00			0.00	
		(0.96)			(0.93)			(0.92)	
$RREL_t$		0.00			0.00			0.00	
		(0.36)			(0.31)			(0.36)	
$INFL_t$		0.03			0.03			0.03	
		(2.89)			(2.89)			(2.86)	
LIQ_t			-402.19			-402.39			-396.22
			(-5.07)			(-5.07)			(-4.99)
IV_t			0.33			0.33			0.33
			(3.61)			(3.65)			(3.56)
IC_t			-0.05			-0.05			-0.05
			(-4.11)			(-4.15)			(-4.63)
V_t	0.38	-0.31	0.34	0.38	-0.31	0.34	0.40	-0.33	0.35
	(8.40)	(-1.87)	(7.36)	(8.44)	(-1.88)	(7.35)	(8.90)	(-1.95)	(7.75)
M_t	1.39	1.03	0.95	1.41	1.10	1.01	3.13	3.46	1.76
	(3.64)	(0.90)	(2.39)	(3.63)	(0.95)	(2.48)	(2.52)	(0.89)	(1.37)
$Intercept$	-0.02	0.05	0.02	-0.02	0.05	0.02	-0.03	0.06	0.01
	(-1.91)	(1.75)	(1.73)	(-1.80)	(-1.78)	(1.75)	(-2.20)	(1.72)	(1.33)
$\boldsymbol{R^2_{adj}}$	**0.09**	**0.14**	**0.11**	**0.09**	**0.14**	**0.11**	**0.08**	**0.14**	**0.11**

Tabla 12: Resultados Regresiones Variable Dependiente “Market Monthly Returns” de 1 mes en el futuro y Variable Independiente “Volume Overload”, “Market Sentiment (3 métodos)” y Variables de Control.

	Vader		
	I	**II**	**III**

[16] Para ver qué sucede en caso de aumentar la ventana de tiempo de 1 mes a 4 meses, revisar anexo 4 y 5

$\log\left(\frac{P_t}{D_t}\right)$	-0.01 (-1.24)		
$\log\left(\frac{P_t}{E_t}\right)$	0.01 (1.74)		
CAY_t		-0.45 (-1.94)	
DS_t		-0.01 (-1.61)	
TS_t		0.00 (1.08)	
$RREL_t$		0.00 (0.33)	
$INFL_t$		0.03 (3.04)	
LIQ_t			-0.04 (-5.04)
VR_t			0.33 (3.66)
IC_t			-0.05 (-4.89)
V_t	0.40 (9.17)	-0.32 (-1.92)	0.34 (7.63)
M_t	0.06 (2.65)	0.03 (0.71)	0.05 (2.20)
$Intercept$	-0.05 (-3.69)	0.03 (1.11)	-0.00 (-0.10)
$\boldsymbol{R^2_{adj}}$	**0.09**	**0.14**	**0.11**

Tabla 13: Resultados Regresiones Variable Dependiente “Market Monthly Returns” de 1 Mes en el Futuro y Variable Independiente “Volume Overload”, “Market Sentiment (Método Vader)” y Variables de Control.

De las tablas 12 y 13 se observa el poder predictivo a 1 mes del sentimiento de mercado y volumen, además de mantener las mismas propiedades que en la corte transversal de los datos: los métodos de Loughran McDonald y Loughran McDonald NH IDF son explicados por las otras variables de control, mientras que Loughran McDonald NH y Vader son significativos siempre además de entregar mejores R^2_{adj}.

4. Conclusiones

Durante el desarrollo de esta investigación se trabajó con las noticias asociadas al New York Times desde su puesta en circulación en septiembre de 1851 hasta el año 2018. Del total de las noticas del New York times se filtraron todas las noticias financieras mediante el uso de técnicas de “machine learning”. De la misma forma se utilizaron los retornos asociados al S&P 500 entre los años 1927 y 2018. Todo esto con el fin de explicar de el posible efecto que podrían tener los medios de comunicación sobre las decisiones de los agentes en la economía.

Hipótesis 1: Relación Information Overload, Volumen y Retornos

Respecto a los efectos encontrados, fue posible encontrar evidencia de la existencia de una relación entre sobrecarga de información (cantidad de noticias financieras publicadas por mes en el New York Times) y volumen transado (sobrecarga de volumen). De la misma forma, se observó que existe una relación positiva entre volumen transado y los retornos.

En base a los efectos encontrados es posible hablar de la variable de Momentum (Asness (2013)), ya que se concluye que el volumen tiene un comportamiento similar al momentum: a mayor volumen mayores son los retornos y a menor volumen menores son los retornos.

En cuanto a posibles mejoras que se podrían realizar al modelo estudiado, parece interesante estudiar el fenómeno de manera diaria y ver si afecta tanto el volumen transado como los retornos. Por otro lado, sería interesante estudiar si la sobrecarga de información tiene alguna consecuencia sobre las crisis, para poder entender si la cantidad de información tiene alguna influencia en situaciones extremas. Finalmente se propone separar el volumen transado en dos variables: V_t^+ y V_t^- (mediante la misma metodología con la que se separa la sobrecarga de información).

Hipótesis 2: Relación Sentimiento de Mercado, Volumen y Retornos

Respecto a los efectos encontrados, fue posible encontrar evidencia de la existencia de una relación entre el sentimiento de mercado (obtenido de la cantidad de palabras positivas y negativas sobre el total de palabras de la noticia) y los retornos del S&P 500. Algo que interesante es que cuando se calcula el sentimiento de mercado se utilizan 4 métodos que resuelven problemáticas propuestas por Loughran McDonald (2016). Si bien resuelven las problemáticas propuestas las 2 diferencias significativas se dan cuando: se resuelve el problema de la doble negación y cuando se utiliza un "diccionario" no financiero pero que si resuelve los problemas propuestos por Loughran McDonald (2016) (este "diccionario" corresponde al algoritmo vaderSentiment).

Un resultado importante es que, si bien el sentimiento de mercado y volumen explican los retornos de mercado, además logran ser significativos aun cuando son incluidos en la misma regresión de las variables de control. Esto es muy importante ya que se concluye que tanto el sentimiento de mercado como la sobrecarga de volumen son factores que están explicando un riesgo que toman los inversionistas no explicado por los factores conocidos.

Como posibles mejoras a este método se propone, además de separar el volumen transado en dos variables, calcular 25 portafolios según la metodología de Fama y French (1992) utilizando como variables el sentimiento de mercado (positivo, neutro y negativo) y la sobrecarga de volumen (alta y baja). Por otro lado, se propone utilizar las 5 categorías de Calomiris y Mamaysky (2018) para ver como se descompone el sentimiento de mercado y entender como cada componente afecta el mercado. Por otro lado, sería interesante estudiar si existe algún mecanismo en tiempos de recesión, como por ejemplo entre la categoría de crédito, el sentimiento de mercado y los retornos. Finalmente se propone estudiar el fenómeno en otros mercados, para entender si el comportamiento es fundamental a los mecanismos financieros o es específico del mercado de Estados Unidos.

Bibliografía

1. Fang, L. and J. Peress (2009). Media Coverage and the Cross-section of Stock Returns. The Journal of Finance 64 (5), 2023-2052.
2. Engelberg, J. and C. Parsons (2011). The Causal Impact of Media in Financial Markets. The Journal of Finance 66 (1), 67-97.
3. Dougal, C., J. Engelberg, D. García, C. Parsons (2012). Journalists and the Stock Market. The Review of Financial Studies 25 (3), 639-679
4. García, D. (2013). Sentiment during Recessions. The Journal of Finance 68 (3), 1267-1300.
5. Peress, J. (2014). The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes. The Journal of Finance 69 (5), 2007–2043.
6. Baker, S., N. Bloom, S. Davis (2016). Measuring Economic Policy Uncertainty. The Quarterly Journal of Economics 131 (4), 1593–1636.
7. Manela, A. and A. Moreira (2017). News implied volatility and disaster concerns. Journal of Financial Economics 123 (1), 137-162.
8. Tetlock, P. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. The Journal of Finance 62 (3), 1139–1168.
9. Knight Foundation (2018). American Views: Trust, Media and Democracy. https://knightfoundation.org/reports/american-views-trust-media-and-democracy
10. 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.
11. Loughran, T. and B. McDonald (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research 54 (4), 1187-1230.
12. 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.
13. Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. Psychological Review 63 (2), 81-97.
14. Schroder, H., M. Driver and S. Streufert (1967). Human Information Processing: Individuals and Groups Functioning in Complex Social Situations.
15. Simon, H. and Newell, A. (1971). Human Problem Solving: The State of the Theory in 1970. American Psychologist 26 (2), 145-159.
16. Hodrick, R. and E. Prescott (1997). Postwar U.S. Business Cycles: An Empirical Investigation. Journal of Money, Credit and Banking 29 (1), 1-16.
17. Asness, C., T. Moskowitz and L. Pedersen (2013). Value and Momentum Everywhere. The Journal of Finance 68 (3), 929-985.
18. Campbell, J. and R. Shiller (1988). Stock prices, earnings, and expected dividends. Journal of Finance 43, 661–676.
19. Fama, E. and K. French (1988). Dividend yields and expected stock returns. Journal of Financial Economics 22, 3–25.
20. Campbell, J. and M. Yogo (2006). Efficient tests of stock return predictability. Journal of Financial Economics 81, 27–60.

21. Lettau, M. and S. Ludvigson (2001). Consumption, aggregate wealth and expected stock returns. Journal of Finance 56, 815–849.
22. Guo, H. (2006). On the out-of-sample predictability of stock market returns. Journal of Business 79, 645–670.
23. Bollerslev, T., G. Tauchen, and H. Zhou (2009). Expected stock returns and variance risk premia. Review of Financial Studies 22, 4463–4492.
24. Pollet, J. and M. Wilson (2010). Average correlation and stock market returns. Journal of Financial Economics 96, 364–380.
25. Keim, D. B. and R. F. Stambaugh (1986). Predicting returns in the stock and bond markets. Journal of Financial Economics 17, 357–390.
26. Campbell, J. (1987). Stock returns and the term structure. Journal of Financial Economics 18, 373–399.
27. Fama, E. and K. French (1989). Business conditions and expected returns on stocks and bonds. Journal of Financial Economics 25, 23–49
28. Lintner, J. (1975). Inflation and security returns. Journal of Finance 30, 259–280
29. Fama, E. and G. Schwert (1977). Asset returns and inflation. Journal of Financial Economics 5, 115–146.
30. Fama, E. (1981). Stock returns, real activity, inflation, and money. American Economic Review 71, 545–565.
31. Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5, 31–56
32. Fama, E., & French, K. (1992). The Cross-Section of Expected Stock Returns. The Journal of Finance 47 (2), 427–465.

Anexos

1. Listado Palabras Modificadoras de Sentimiento

- No
- Nor
- Not
- None
- No one
- Nobody
- Nothing
- Neither
- Nowhere
- Never
- Hardly
- Scarcely
- Barely
- Doesn’t
- Isn’t
- Shouldn’t
- Wouldn’t
- Couldn’t
- Won’t
- Can’t
- Don’t

2. Cálculo Filtro Hodrick Prescott

Para calcular el filtro de Hodrick Prescott se asume una serie de tiempo que tiene tendencia, una componente cíclica y una componente de error tal que:

$$y_t = \tau_t + c_t + \epsilon_t$$

Por lo tanto, dado un λ adecuado, existe una componente tendencial que resuelve:

$$\min_{\tau} \left(\sum_{t=1}^{T} (y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1} [(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2 \right)$$

Se recomienda un valor de $\lambda = 1600$ para datos trimestrales y $\lambda = 129{,}600$ para datos mensuales.

3. Gráficos Tendencia y Ciclo Filtro Hodrick Prescott

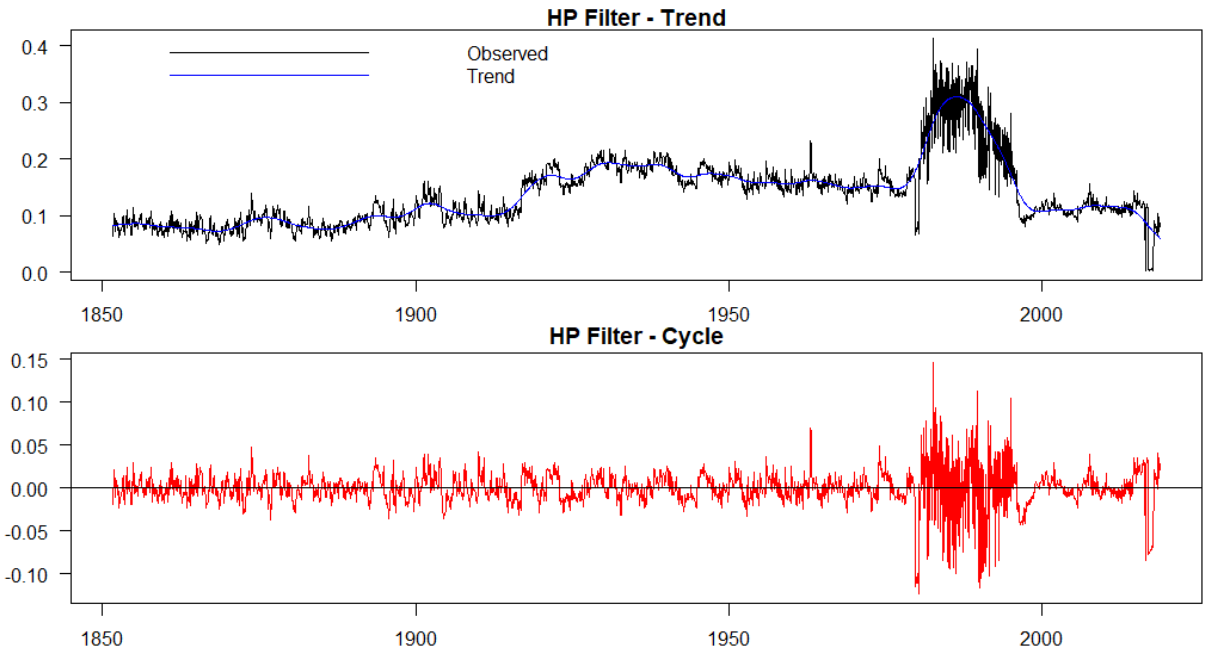

Ilustración A1 y A2: Tendencia y ciclo del filtro HP para la variable “Information Overload”

4. Predicción de Retornos con Sentimiento de Mercado (LM NH)

	Retorno Mensual de Mercado			
M_{t-1}	2.06 (5.33)			
M_{t-2}		1.58 (4.06)		
M_{t-3}			1.24 (3.18)	
M_{t-4}				0.90 (2.31)
$intercept$	0.04 (6.73)	0.03 (5.50)	0.03 (4.65)	0.02 (3.81)
$\boldsymbol{R^2_{adj}}$	**0.02**	**0.01**	**0.01**	**0.004**

Tabla A1: Resultado regresión retorno mensual de mercado vs sentimiento de mercado mediante método LM NH.

5. Predicción de Retornos con Sentimiento de Mercado (Vader)

	Retorno Mensual de Mercado			
M_{t-1}	0.08 (3.99)			
M_{t-2}		0.065 (3.13)		
M_{t-3}			0.04 (2.12)	
M_{t-4}				0.02 (1.10)
$intercept$	-0.01 (-1.58)	-0.00 (-0.78)	0.00 (0.17)	0.00 (1.13)
$\boldsymbol{R^2_{adj}}$	**0.01**	**0.01**	**0.003**	**0.0002**

Tabla A2: Resultado regresión retorno mensual de mercado vs sentimiento de mercado mediante método Vader.