

## CAPÍTULO VI

# Predicciones financieras basadas en análisis de sentimiento de textos y minería de opiniones

Argimiro Arratia\*

En este capítulo se describe la mecánica básica para construir un modelo de predicción que utiliza indicadores de sentimiento derivados de datos textuales. Enfocamos nuestro objetivo de predicciones en series de tiempo financieras y presentamos un conjunto de hechos empíricos que describen las propiedades estadísticas de los indicadores de sentimiento, con particular atención en aquellos indicadores extraídos de noticias sobre mercados financieros y cuya categorización de sentimientos se basa en diccionarios. El objetivo general es proporcionar pautas para los profesionales en el mundo de las finanzas para la adecuada construcción e interpretación de su propia información numérica dependiente del tiempo y que representa la percepción del público hacia las empresas, los precios de las acciones y los mercados financieros en general.

*Palabras clave:* análisis de sentimiento, emociones financieras, diccionario financiero, modelos de pronóstico.

---

\* El autor agradece la ayuda recibida por Gustavo Avalos, Ariel Duarte y Martí Renedo en la elaboración de resultados presentados en la tabla 2 y la figura 4, y en la recopilación de software para el análisis de sentimiento. A Alejandra Cabaña por sus comentarios y correcciones de los aspectos estadísticos de este artículo.

## 1. INTRODUCCIÓN

Hoy en día existe una gran cantidad de compañías tecnológicas financieras (*fintech*) que ofrecen índices de sentimientos sobre los mercados, contruidos a partir de opiniones publicadas en periódicos electrónicos, redes sociales, y otras fuentes de noticias en Internet. Además ahora existen incontables fuentes de grandes cantidades de datos, que están en continuo crecimiento, y que nos permiten desarrollar estrategias basadas en datos, en particular textos que recogen opiniones del público, en lugar de simplemente replicar y actuar sobre modelos econométricos clásicos.

Estos datos temporales que cuantifican las emociones del público general hacia compañías que cotizan en bolsa, personalidades y eventos del mundo de las finanzas, son utilizados por muchos inversores como información social en sus estrategias de compra y venta en los mercados financieros. La extensa investigación en finanzas conductuales ha mostrado evidencia del hecho que los inversores reaccionan a las noticias, y que, por lo general, muestran una mayor propensión a hacer un movimiento de inversión basado en malas noticias en lugar de buenas noticias (véase por ejemplo Chan, 2003; Brown y Cliff, 2004; Baker y Wurgler, 2007; Kumar y Lee, 2006; Tetlock, 2007). Esta actitud la atribuyen algunos autores a un rasgo de la psicología humana (Baumeister *et al.*, 2001; Rozin y Royzman, 2001), y otros a una estrategia específica de inversión en el mercado de valores (Engelberg, Reed y Ringgenberg, 2012).

El análisis de sentimiento de textos o minería de opinión trata con la categorización de las opiniones expresadas en un texto. En el contexto financiero el propósito es capturar las emociones de los inversionistas hacia el mercado financiero expresadas en redes sociales o medios informativos, y cuantificar estas emociones como variables numéricas que potencialmente serán predictoras del comportamiento de los mercados de valores.

Este objetivo es contrario al mantra de la teoría clásica del mercado, que sostiene que toda información se incorpora a los precios de las acciones tan pronto como se hace pública. Sin embargo, cada vez existen más indicios que apuntan a que la información se difunde gradualmente a través de la población de inversionistas y que esta difusión gradual afecta los precios de las acciones. Estudios recientes demuestran que existe alguna dependencia entre los precios del mercado y las historias publicadas por los medios.

Entre los primeros trabajos que reportan alguna relación, aunque débil, entre las noticias sobre empresas y el movimiento del precio, o la volatilidad, de las acciones de estas empresas tenemos los de Berry y Howe (1994) y de Mitchell y Mulherin (1994). Ambos trabajos, publicados en 1994, no tratan el sentimiento en la noticia, sino la relación del volumen de estas con la volatilidad o el volumen de negociación. Chan (2003) sí que analiza el tono de los titulares de noticias sobre compañías de diferentes capitalizaciones y observa un incremento en la variabilidad del precio de las acciones

de compañías de más baja capitalización después de la publicación de noticias negativas. El trabajo de Antweiler y Frank (2004) es pionero en el desarrollo de medidas de sentimiento de noticias mediante métodos de aprendizaje automático, para explicar el comportamiento del precio de las acciones. Estos investigadores emplean un algoritmo Naive Bayes para clasificar el sentimiento de los documentos a partir de un conjunto de entrenamiento construido manualmente, y obtienen de esta clasificación unas series de señales de compra y venta para el mercado. Concluyen que si bien estas señales pueden predecir la volatilidad del mercado, su capacidad predictiva de la variación del precio de las acciones es mínima.

En 2007, Tetlock publica un serio análisis de la técnica basada en léxicos, o “bolsa de palabras” (*bag-of-words*), para clasificar los sentimientos implícitos en las noticias del ámbito financiero (Tetlock, 2007). Esta técnica consiste en construir listas (bolsas) de palabras, donde cada una de estas listas se asocia con una categoría (p.ej. positivo o negativo). Mediante una clasificación de palabras basada en categorías extraídas del *Diccionario Psicosocial de Harvard*, Tetlock cuantifica el optimismo y el pesimismo contenido en la columna “Abreast of the Market” del *Wall Street Journal* y observa que los niveles altos de pesimismo reflejado por el conjunto de las noticias predicen caídas de los precios del mercado, a los que les sigue a continuación un movimiento inverso. Empleando técnicas similares basadas en diccionarios, Tetlock, Saar-Tsechansky y Macskassy (2008) descubren que la proporción de palabras de carga negativa en las noticias del *Dow Jones News* y *Wall Street Journal* predicen futuras variaciones en los beneficios de las compañías, lo cual explican como consecuencia del hecho de que el contenido lingüístico en las noticias financieras captura los aspectos difíciles de cuantificar de los fundamentales de una compañía que son rápidamente incorporados en el precio de las acciones.

Una mejora a los métodos de análisis de sentimiento basados en diccionarios fue aportada posteriormente por Loughran y McDonald (2011), quienes demostraron que las listas de palabras desarrolladas para disciplinas sociales o humanísticas en general (como los diccionarios Harvard) clasifican erróneamente las palabras comunes en textos financieros. Como alternativa Loughran y McDonald desarrollan listas de palabras extraídas de los formularios anuales 10-K, con una carga sentimental positiva, negativa y otras cuatro categorías que reflejan más fielmente el tono en los textos financieros, y muestran que la proporción de palabras negativas en estos formularios 10-K afecta a la baja los rendimientos de las acciones.

Si bien no existe un consenso general sobre la posibilidad de predecir los movimientos en los precios de acciones en los mercados financieros mediante el análisis del sentimiento de las noticias (véase una extensa discusión al respecto en (Schoen *et al.*, 2013)), el hecho relevante es que esta área de investigación es actualmente una de las más populares en el desarrollo de modelos de pronóstico que hacen uso de grandes masas de datos, tanto a nivel académico como industrial.

## 2. LA MECÁNICA DEL ANÁLISIS DE SENTIMIENTO DE TEXTOS

Nuestra exposición se centra en el análisis de sentimientos de textos con referencia a un sujeto. Esto significa que nuestro objetivo es determinar si un documento, o una frase dentro de un documento, expresa emociones positivas, negativas o de otro tipo *hacia un sujeto específico*. Existen otras categorías de análisis de sentimiento textual, las cuales se describen en el libro de texto de Liu (2015). En aplicaciones financieras, nuestros sujetos de interés son empresas, mercados financieros, materias primas o cualquier otra entidad con valor económico. Esta información sentimental se utiliza para alimentar modelos de pronóstico de variables estocásticas que cuantifican el comportamiento del sujeto de interés; p.ej. precios, volatilidad, u otros indicadores estadísticos financieros.

Se propone a continuación un flujo de trabajo para construir modelos de pronóstico financiero basados en sentimientos extraídos de datos textuales. Este comprende las siguientes etapas:

- Creación y procesamiento de *corpus* textuales.
- Cálculo del sentimiento.
- Agregación de valores de sentimiento y construcción de indicadores.
- Modelización.

### 2.1. Gestión del *corpus* textual

La primera tarea fundamental es la recopilación de textos adecuados, y la aplicación de técnicas de minería de texto para limpiar y categorizar los términos dentro de cada documento. Los textos han de estar en formato electrónico y cada documento ha de tener un identificador único (por ejemplo, un nombre de archivo singular) y una marca de tiempo (e.g. fecha y/o hora de su publicación). Además, utilizando algún esquema de reconocimiento de entidades, se ha de identificar dentro de cada documento los sujetos de interés (e.g. el nombre o *ticker* de compañías cotizadas). De esta manera podremos agrupar los documentos por sujeto común, siendo posible que un mismo documento aparezca en dos grupos diferentes pertenecientes a dos sujetos diferentes.

*Ejemplo 2.1.* Los sujetos (por ejemplo, el nombre de una empresa o un indicador de su cotización-ticker) se pueden identificar mediante la concordancia de palabras clave o técnicas de reconocimiento de nombres de entidades<sup>1</sup>. Alternativamente, algunos proveedores de noticias como Dow Jones Newswires incluyen etiquetas en sus archivos xml que indican la empresa de la que trata la noticia.

<sup>1</sup> Consulte el software Stanford NER <https://nlp.stanford.edu/software/CRF-NER.shtml>

## 2.2. Cálculo del sentimiento

El análisis de sentimientos es básicamente un problema de clasificación de textos. Por lo tanto, se puede abordar este problema algorítmico de las dos formas posibles de clasificación mecánica: (1) aplicando un algoritmo de aprendizaje automático supervisado que se entrena con textos ya etiquetados como positivo o negativo (o cualquier otra emoción); o (2) utilizar un método de clasificación sin supervisión basado en el reconocimiento de algunos patrones sintácticos fijos, o palabras claves, que expresan un sentimiento específico (un diccionario o léxico de sentimiento). Esta última solución es utilizada con mayor frecuencia por investigadores y profesionales de las finanzas. Por lo tanto, en esta exposición priorizamos la descripción del método no supervisado basado en léxicos de sentimiento y únicamente daremos algunas indicaciones a la literatura sobre el enfoque de aprendizaje automático para la clasificación de sentimientos.

### 2.2.1. Método no supervisado basado en léxicos para la clasificación de sentimiento

El componente clave de este método de clasificación de texto es un conjunto de palabras, o patrones sintácticos, que denotan un sentimiento específico. Por ejemplo, la positividad se expresa con palabras como *bueno*, *admirable*, *mejor*, ... y emoticonos como : -) o ; -], y otros símbolos similares que se utilizan a menudo en mensajes cortos como los de Twitter (Bifet y Frank, 2010; Go, Bhayani, y Huang, 2009). Estas *bolsas de palabras* (como comúnmente se les refiere en la jerga computista) conforman un *lexicón de sentimiento o diccionario*. Un ejemplo de diccionario de términos que caracterizan los sentimientos positivo y negativo, específico al ámbito financiero, es el recopilado por Loughran y McDonald (2011), una muestra del cual se presenta en la Tabla 1.

Dado un sentimiento  $S$  (por ejemplo, positivo, negativo, ...), determinado por algún léxico  $L(S)$ , un algoritmo básico para asignar un valor numérico del sentimiento  $S$  a un documento, es contar el número de apariciones de términos de  $L(S)$  en el documento. Este número da una medida de la fuerza del sentimiento  $S$  en el documento. Para comparar las fortalezas de dos sentimientos diferentes en un documento, es aconsejable relativizar estos números con el número total de términos en el documento. Hay muchas mejoras de esta función básica de valoración sentimental. Por ejemplo, si se considera una escala de valores predeterminados para los términos en el léxico (en lugar de que cada uno tenga el mismo valor de 1), o si se permiten valores negativos, además de positivos, para capturar la dirección del sentimiento, o se hacen otras consideraciones sobre el contexto donde, dependiendo de las palabras vecinas, los términos del léxico pueden cambiar sus valores o incluso pasar de un sentimiento a otro. Por ejemplo, *bueno* es positivo, pero *no es bueno* es negativo. Repasaremos algunas de estas variantes, pero para una exposición más detallada, consulte el libro de texto de Liu (2015) y sus referencias.

TABLA 1.

UNA MUESTRA DEL DICCIONARIO RECOPIADO POR LOUGHRAN Y MCDONALD (2011) A PARTIR DE REPORTES 10-K. CONSTA DE 2355 TÉRMINOS NEGATIVOS Y 354 POSITIVOS

Negativo	Positivo
Forbid	Collaborates
Adversity	Pleased
Intimidation	Rewarding
Barred	Stabilize
Distracted	Gaining
Diminishes	Profitably
Accusations	Achieve
Overstating	Effective
Declines	Inventive
Aftermaths	Revolutionize

Fuente: University of Notre Dame, Software Repository for Accounting and Finance (<https://sraf.nd.edu/textual-analysis/resources/>)

Formalicemos ahora un esquema general para definir una serie temporal de valoraciones de sentimiento para documentos con respecto a un objetivo específico (por ejemplo, una empresa o un instrumento financiero), basados en algún diccionario sentimental. Tenemos a mano  $\lambda = 1, \dots, \Lambda$  léxicos  $L_{\lambda}$ , donde cada uno define un sentimiento. Tenemos  $K$  posibles sujetos y recopilamos un flujo de documentos en diferentes instantes de tiempo  $t = 1, \dots, T$ . Sea  $N_t$  el número total de documentos en el instante de tiempo  $t$ . Sea  $D_{n,t,k}$  el  $n$ -ésimo documento, en el tiempo  $t$  y que menciona al  $k$ -ésimo sujeto, para  $n = 1, \dots, N_t$ ,  $t = 1, \dots, T$  y  $k = 1, \dots, K$ .

Fijemos un lexicón  $L_{\lambda}$  y un sujeto  $G_k$ . Se define un valor numérico del sentimiento basado en el lexicón  $L_{\lambda}$  para el documento  $D_{n,t,k}$ , en tiempo  $t$  y referente al sujeto  $G_k$  mediante la siguiente ecuación:

$$S_{n,t}(\lambda, k) = \sum_{i=1}^{I_d} w_i S_{i,n,t}(\lambda, k) \quad [1]$$

donde  $S_{i,n,t}(\lambda, k)$  es el valor de sentimiento asignado al  $i$ -ésimo término singular que aparece en el documento y de acuerdo con el lexicón  $L_{\lambda}$ , siendo este valor cero si el término no está en el lexicón.  $I_d$  es el número total de términos en el documento  $D_{n,t,k}$  y  $w_i$  es un peso, para cada término, que determina la forma en que se agregan las valoraciones de sentimiento en el documento.

*Ejemplo 2.2.* Si  $S_{i,n,t} = 1$  (o 0 si es el caso que el  $i$ -ésimo término no está en el lexicón), para todo  $i$ , y  $w_i = 1/I_d$  tenemos entonces la muy básica estimación de la densidad del

sentimiento en el texto utilizado, por ejemplo, en (Tetlock, Saar-Tsechansky y Macskassy, 2008; Loughran y McDonald, 2011) y muchos otros trabajos sobre análisis de sentimiento de textos, el cual da igual importancia a todos los términos en el lexícón. Un esquema de ponderación más refinado, que refleja diferentes niveles de relevancia de cada término con respecto al sujeto-objetivo, es considerar  $w_i = \text{dist}(i, k)^{-1}$ , donde  $\text{dist}(i, k)$  es una métrica de distancia entre el  $i$ -ésimo término y el  $k$ -ésimo sujeto, como en (Ding, Liu y Yu, 2008).

El valor de sentimiento  $S_{i,n,t}$  es un número real y puede descomponerse en factores  $v_i \cdot s_i$ , donde  $v_i$  es un número que representa un cambio de sentimiento (un *cambiador de valencia*: una palabra que cambia los sentimientos en la dirección opuesta) y  $s_i$  el valor de sentimiento *per se*.

### Sobre los cambiadores de valencia

Definidos originalmente y analizado su efecto contrario sobre el sentimiento calculado en documentos de habla inglesa en (Polanyi y Zaenen, 2006), estas son palabras que pueden alterar el significado de una palabra polarizada, y que pertenecen a una de las cuatro categorías básicas siguientes: *negadores*, *amplificadores*, *de-amplificadores* y *conjunciones adversativas*. Un negador invierte el signo de una palabra polarizada, como en la frase: “esa empresa *no* es una *buena* inversión”. Un amplificador intensifica la polaridad de una oración, como por ejemplo el adverbio *definitivamente* amplifica la negatividad en el ejemplo anterior: “esa empresa *no* es *definitivamente* una *buena* inversión”. Los de-amplificadores (también conocidos como reductores), por otro lado, disminuyen la intensidad de una palabra polarizada (p. ej., “la empresa es *apenas* buena como inversión”). Una conjunción adversativa anula la polaridad de sentimiento de la cláusula precedente, p. ej., “Me gusta la empresa *pero* no vale la pena”.

¿Debemos preocuparnos por los cambiadores de valencia? Si los cambiadores de valencia ocurren con frecuencia en nuestros conjuntos de textos, entonces no considerarlos en el cálculo de las valoraciones de sentimiento en la ecuación [1] generará una valoración inexacta del sentimiento del texto. Más aún en el caso de negadores y conjunciones adversativas que invierten o anulan la polaridad del sentimiento de la oración.

Para conjuntos de textos extraídos de redes sociales como Twitter o Facebook, se ha observado que la ocurrencia de cambios de valencia, en particular de negadores, es considerablemente alta (aproximadamente un 20% para varios trending topics<sup>2</sup>), por lo que en este contexto es importante tener en consideración este fenómeno gramatical.

En el ámbito financiero, hemos calculado la ocurrencia de los cambiadores de valencia (en inglés) en una muestra de 1,5 millones de documentos del conjunto de Dow Jones

<sup>2</sup> <https://cran.r-project.org/web/packages/sentimentr/readme/README.html>

Newswires. Los resultados de estos cálculos, que se pueden ver en la Tabla 2, muestran una baja incidencia de de-amplificadores y conjunciones adversativas (alrededor del 3%), pero los negadores aparecen en un número que puede merecer cierta atención.

## Creación de léxicos de sentimiento

TABLA 2.

**PORCENTAJE DE OCURRENCIA DE CAMBIADORES DE VALENCIA EN 1,5MM DOCS. DJN**

Texto	Negadores	Amplificadores	De-amplificadores	Adversativas
DJN noticias	7,00	14,13	3,02	3,02

Un punto de partida para compilar un conjunto de palabras con una carga sentimental específica es utilizar un diccionario estructurado (preferiblemente en línea como WordNet<sup>3</sup>) que enumere sinónimos y antónimos para cada palabra. Luego comience con algunas palabras seleccionadas (palabras claves) que conlleven un sentimiento específico y continúe agregando algunos de los sinónimos al conjunto, y al conjunto que caracteriza al sentimiento contrario agregue los antónimos. Hay muchas formas inteligentes de hacer esta expansión desde palabras claves (o semillas) de sentimiento utilizando algún algoritmo de clasificación supervisado para encontrar más palabras que conlleven una emoción similar. Un ejemplo es el trabajo de Tsai y Wang (2014) donde tratan la expansión de un conjunto de palabras claves financieras utilizando el modelo continuo de bolsa de palabras aplicado a los informes financieros anuales obligatorios, conocidos como 10-K en EE. UU. Otro ingenioso esquema supervisado para construir léxicos basado en la teoría de redes se expone en (Rao y Ravichandran, 2009). Para una descripción más extensa sobre la creación de léxicos de sentimientos veáse (Liu, 2015, Cap. 7) y las referencias que allí se listan.

### 2.2.2. Método de aprendizaje automático supervisado para la clasificación de sentimientos

Otra forma de clasificar textos es mediante el uso de algoritmos de aprendizaje automático, que se fundamentan en un modelo previamente entrenado para generar predicciones. A diferencia del método basado en diccionarios, estos algoritmos no están programados para responder de cierta manera según las entradas recibidas, sino para extraer patrones de comportamiento de conjuntos de datos de entrenamiento pre Etiquetados. Los algoritmos internos que dan forma a la base de este proceso de aprendizaje tienen algunas componentes sólidamente fundamentadas en la Matemática y la Estadística, lejos de ser heurísticas arbitrarias. Algunos de los algoritmos de aprendizaje

<sup>3</sup> <https://wordnet.princeton.edu/>

automático más populares y robustos son Naïve Bayes, *Support Vector Machines* y *Deep Learning*. Las etapas para la clasificación de sentimientos en textos, utilizando modelos de aprendizaje automático son las siguientes:

*Desarrollo y preprocesamiento de un corpus de textos.* El proceso de aprendizaje parte de un corpus de textos clasificados manualmente, que después de la extracción de características, será utilizado por el algoritmo de aprendizaje automático para encontrar los parámetros que mejor se ajusten al modelo y evaluar la precisión en una etapa de prueba. Es por esto que la parte más importante de este proceso es el desarrollo de un buen conjunto de textos de entrenamiento. Este debe ser lo más grande posible y representativo del conjunto de datos que se analizarán. Una vez obtenido el corpus, se deben aplicar técnicas para reducir el ruido que generan las palabras sentimentales sin sentido, así como para aumentar la frecuencia de cada término mediante la derivación y lematización. Estas técnicas dependen del contexto al que se aplique. Esto significa que un modelo entrenado para clasificar textos de un campo determinado no podría aplicarse directamente a otro. Por tanto, es de suma importancia tener un corpus clasificado manualmente lo mejor posible.

*Extracción de características.* El enfoque general para extraer características consiste en transformar el texto preprocesado en una expresión matemática basada en la detección de la coocurrencia de palabras o frases. El texto transformado se divide en una serie de características, cada una de las cuales corresponde a un elemento del texto original de entrada.

*Clasificación.* Durante esta etapa, el modelo entrenado recibe un conjunto de características aún no vistas para estimar una nueva clase.

Para más detalles sobre todo este procedimiento de aprendizaje consúltese (Sebastiani, 2002; Liu, 2015).

Un ejemplo de algoritmo de aprendizaje automático para el análisis de sentimiento en textos es *Deep-MLSA* (Deriu *et al.*, 2016, 2017). Este modelo consta de un clasificador de red neuronal convolucional multicapa con tres estados que corresponden a sentimientos negativos, neutrales y positivos. *Deep-MLSA* se adapta muy bien a la corta longitud y el carácter informal de los tweets de las redes sociales, y ha resultado ser el mejor algoritmo de clasificación de polaridad de mensajes en la sección de "Sentiment Analysis in Twitter" de la competición *SemEval* (Nakov *et al.*, 2016).

### 2.3. Métodos para agregar valores de sentimiento y construir indicadores

Sean  $L_\lambda$  un lexicón que caracteriza algún sentimiento y  $G_k$  un sujeto-objetivo. Una vez calculados los valores del  $L_\lambda$ -sentimiento para cada documento que menciona el sujeto

$G_k$  y siguiendo la rutina expuesta en la ecuación [1], se procede a agregar estos valores para cada instante de tiempo  $t$  disponible, y así obtener el  $L_\lambda$ -sentimiento sobre  $G_k$  y en el instante de tiempo  $t$ , que denotaremos por  $S_t(\lambda, k)$ , y formalmente describimos mediante la siguiente ecuación:

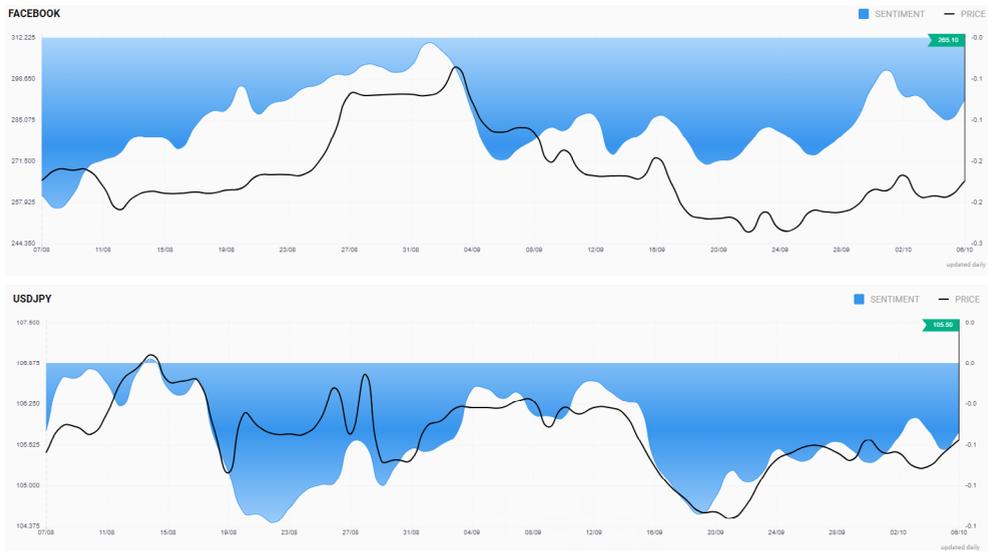
$$S_t(\lambda, k) = \sum_{n=1}^{N_t} \beta_n S_{n,t}(\lambda, k) \quad [2]$$

Al igual que en la ecuación [1], los pesos  $\beta_n$  determinan la manera en que son agregados los valores de sentimiento para cada documento. Por ejemplo, si se considera  $\beta_n = 1/\text{card}(D_{n,t,k})$ , esta ponderación da más relevancia a los documentos de menor longitud. ( $\text{card}(D_{n,t,k})$  es la cardinalidad de  $D_{n,t,k}$ , esto es el número de términos que lo componen.)

Se obtiene de esta manera una serie temporal de valores de sentimiento, o un *indicador de sentimiento*,  $\{S_t : t = 1, \dots, T\}$ , basado en el léxico  $L_\lambda$  que define un sentimiento específico sobre el sujeto  $G_k$ . Variaciones de este indicador de sentimiento (basado en léxico  $L_\lambda$ ) para  $G_k$  se pueden obtener aplicando cualquier filtro  $F$  a la serie  $S_t$ , y así obtener una nueva serie indicatriz  $\{F(S_t) : t = 1, \dots, T\}$ . Por ejemplo, si se aplica una media móvil se obtendrá una versión suavizada de la serie de sentimientos original. En la figura 1 se muestran ejemplos de activos financieros (*stock* de la compañía Facebook y la tasa de cambio USD/JPY) con indicador de sentimiento (en azul) sobrepuesto al

FIGURA 1

PRECIO DE FACEBOOK Y USD/JPY (EN NEGRO) SOBRE UN INDICADOR DE SENTIMIENTO COMPUESTO (EN AZUL)



Fuente: Cortesía de Acuity Trading Inc.

precio (negro). El indicador de sentimiento es una combinación de positivo y negativo, elaborado por la empresa *Acuity Trading Inc.*

## 2.4. Modelización

Existen básicamente dos enfoques para la modelización: utilizar los indicadores de sentimiento como información exógena que alimentamos a nuestros modelos de pronóstico, y se comprueba su relevancia para pronosticar movimientos de precios, retornos de precios u otras estadísticas del precio; o utilizarlos como criterios para seleccionar los sujetos (objetivos) de las noticias, que en nuestro caso de interés serán compañías cotizadas, y formar una cartera. Algunos ejemplos escogidos de la gran cantidad de investigaciones publicadas sobre el tema de la predicción y la gestión de carteras con análisis de sentimiento de textos son (Arias, Arratia y Xuriguera, 2013; Baker y Wurgler, 2007; Beckers, Kholodilin y Ulbricht, 2017; Heston y Sinha, 2017; Loughran y McDonald, 2011; Tetlock, 2007; Tetlock, Saar-Tsechansky y Macskassy, 2008; Uhl, Pedersen y Malitius, 2015). En la sección tercera damos un ejemplo (académico) de cómo utilizar indicadores de sentimiento para generar señales de inversión.

Para un tratamiento más extenso de los bloques de construcción para producir modelos de pronóstico basados en datos textuales descritos en esta sección consulte (Algaba et al., 2020), y el tutorial para el paquete de R *sentometric* en (Ardia et al., 2019).

## 3. NEGOCIACIONES BURSÁTILES CON SENTIMIENTO

Comencemos por recordar una regla de inversión muy popular entre los analistas técnicos (aquellos que se guían por patrones en las gráficas de precios para realizar inversiones [Achelis, 2001]). Esta es la estrategia de cruces de medias móviles (en notación  $MM(p)$ , donde  $p$  es el periodo de días sobre el que se calcula la media), que consiste en trazar dos medias móviles sobre el precio del activo, una de corto plazo  $MM(c)$  y otra de mediano a largo plazo  $MM(m)$ , y la regla consiste en tomar posiciones largas (comprar) cuando la media móvil corta esté por encima de la larga ( $MM(c) > MM(m)$ ), o tomar posiciones cortas (vender) en caso contrario.

Nuestra idea es aplicar esta regla de inversión, no al precio, sino a un indicador de sentimiento "bullish" (BULL) sobre el banco J.P. Morgan (NYSE:JPM), permitiendo posiciones en corto, y durante el periodo 1 de febrero 2018 a 15 de mayo 2020. Este indicador BULL se ha construido a partir de un diccionario de términos financieros con carga sentimental positiva. La media móvil corta que utilizamos es un  $MM(10)$  y la larga un  $MM(25)$ , y contrastamos nuestra estrategia con la estrategia básica de no hacer nada, es decir "comprar-y-mantener" ("buy-and-hold"). La tabla 3 presenta los resultados de nuestra estrategia (MAX) y de "comprar-y-mantener" (BH) en términos de: retorno

acumulado (un 88% para nuestra estrategia contra un -14% para BH, debido en buena parte por tomar posiciones cortas en el momento de la crisis bursátil ocasionada por la COVID-19); retorno anualizado; cociente de Sharpe (anualizado); volatilidad (anualizada); caída porcentual máxima y de longitud máxima.

TABLA 1.

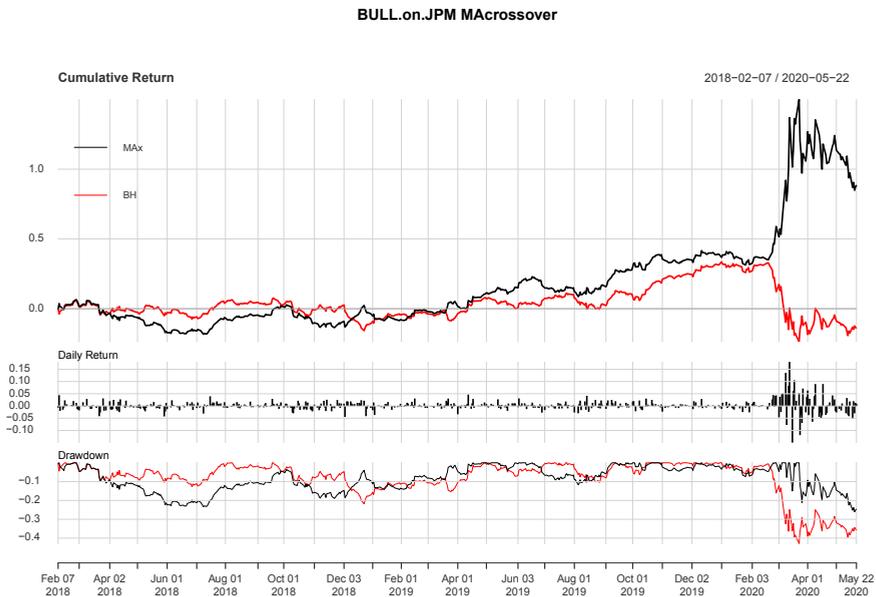
**RESULTADOS DE TRADING CON SENTIMIENTO BULL SOBRE JPM CONTRA "COMPRAR-Y-MANTENER" (BH)**

	MAx	BH
Retorno Acum.	0.8866467	-0.14067188
Retorno Anual	0.3246418	-0.06493845
Cociente de Sharpe (anual.)	0.9167684	-0.18304064
% veces ganancia	0.5106383	0.50531915
Volatilidad (anual.)	0.3541154	0.35477613
Caída % max.	-0.2603982	-0.42870098
Longitud caída max.	284	149

Fuente: Elaboración propia con el software R.

FIGURA 2

**RETORNO ACUMULADO Y CAÍDA MÁXIMA DE LAS ESTRATEGIAS DE INVERSIÓN MAX (NEGRO) Y BH (ROJO)**

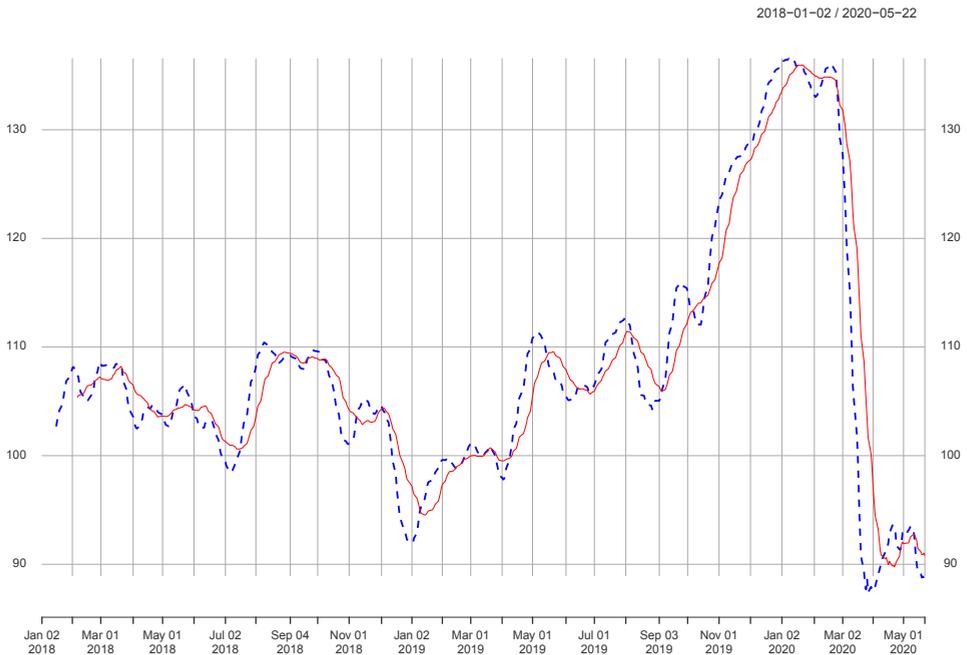


Fuente: Elaboración propia con el software R.

Es evidente que nuestra estrategia de inversión sentimental en JPM gana por goleada a una actitud pasiva. En la figura 2 se muestran la evolución del retorno acumulado y caídas de las estrategias M<sub>Ax</sub> y BH. En la figura 3 se pueden observar los puntos de cruces de las medias móviles M M (10) y M M (25) para JPM en el periodo considerado de inversión.

FIGURA 3

**CRUCES DE M M (10) (AZUL SEGMENTADA) Y M M (25) (ROJO CONTINUA) PARA JPM**



Fuente: Elaboración propia con el software R.

#### 4. PROPIEDADES EMPÍRICAS DE LOS INDICADORES DE SENTIMIENTO

En esta sección se presentan algunas de las propiedades empíricas más notables observadas sobre los indicadores de sentimiento público construidos en distintos contextos. Sirvan estas observaciones de orientación para una mejor comprensión de los datos, su potencial predictivo y utilidad en la construcción de modelos de pronóstico o sistemas de inversión.

Antes debemos advertir que, según un estudio realizado por Kumar y Lee (2006), estas propiedades empíricas de los indicadores de sentimiento parecen ser causados por y afectar principalmente a los inversores minoristas. Porque se acepta que los inversores institucionales son más racionales en la toma de decisiones de inversión basadas en

información pública (no por poseer una psique más estable, menos emocional, sino más bien y en gran parte debido a una mayor automatización de sus procesos de negociación y toma de decisiones, lo cual minimiza la intervención del factor humano en estos eventos). En consecuencia, es el inversor minorista quien se ve más afectado por el tono del sentimiento en las noticias financieras y más propenso a actuar en consecuencia, lo que hace que los precios de las acciones se desvíen de sus valores fundamentales. Por lo tanto, el análisis del sentimiento de los textos financieros y sus aplicaciones tiene más sentido en los mercados con una alta participación de inversores minoristas (principalmente de economías desarrolladas, como EE. UU. y Europa), en contraposición a los mercados emergentes. En estos mercados desarrollados, los inversores institucionales aún podrían explotar las desviaciones de los precios de las acciones de los valores fundamentales debido al comportamiento de los inversores minoristas impulsado por las noticias.

Listamos y comentamos a continuación aquellos hechos observables más comunes para un amplio espectro de indicadores de sentimiento en el ámbito financiero.

#### **4.1. Correlación entre el volumen de noticias y volatilidad del precio**

Cuanto más noticias hay sobre una empresa cotizada, mayor será la volatilidad del precio de su acción en el mercado.

Esta dependencia entre el volumen de noticias y volatilidad de precios se ha observado para varias empresas cotizadas, en distintos mercados y utilizando diferentes fuentes de texto. Por ejemplo, esta relación se ha observado con mensajes extraídos de Twitter sobre empresas que aparecen en el S&P 500 por Arias, Arratia y Xuriguera (2013). Adicionalmente, Aouadi, Arouri y Teulon (2013), Dimpfl y Jank (2015), y Hamid y Heiden (2015) demuestran que para diferentes mercados de valores (Francia, S&P 500, Dow Jones) el volumen de búsquedas en Google (o "Google Trends"), tomado como indicador de atención sobre una compañía o tema económico, tiene poder predictivo de la volatilidad del mercado (en el sentido de una relación causa-efecto). Más recientemente Arratia y López (2020) demostraron esta relación causal del volumen de búsquedas en Google sobre bitcoins y el precio de las cripto-monedas de mayor liquidez.

#### **4.2. Mayor volumen de noticias alrededor de las fechas de publicación de beneficios**

El volumen de noticias sobre una empresa tiende a aumentar significativamente en los días cercanos a la publicación de beneficios obtenidos por la empresa (informes periódicos, usualmente trimestrales, de obligatoriedad por ley). Este hecho ha sido reportado y analizado por Tetlock, Saar-Tsechansky y Macskassy (2008) para las noticias que aparecen en el *Wall Street Journal* y el *Dow Jones Newswires* entre 1980 y 2004, para empresas listadas en el índice S&P 500. En su estudio los autores elaboraron un histograma

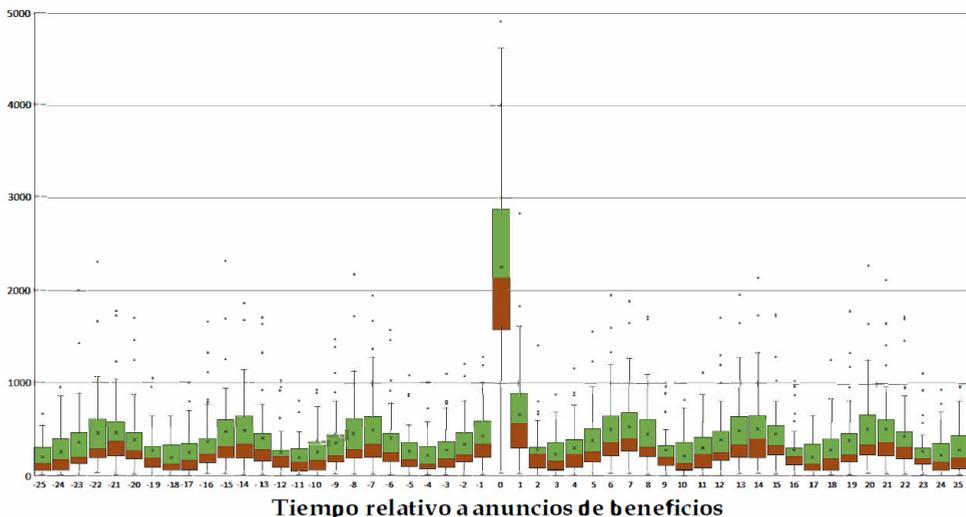
que describe la relación entre el número de noticias específicas sobre la empresa y el número de días desde (respectivamente, hasta) el último (respectivamente, próximo) anuncio de beneficios de la empresa, observando una mayor concentración en los registros cercanos a las fechas de beneficios, que se representan por el 0 del histograma. Este hecho lleva a sugerir a los autores una posible relación de dependencia estadística entre las noticias específicas sobre una empresa y los fundamentales de la empresa.

Hemos reproducido este análisis con datos actuales, para corroborar (y actualizar) las observaciones de Tetlock, Saar-Tsechansky y Macskassy (2008) (figura 4). En nuestra versión consideramos todas las noticias de empresas cotizadas en el S&P 500 que se publican en el *Dow Jones Newswires* desde 2012 a 2019. Sin embargo, consideramos que el efecto de las noticias en el comportamiento de los mercados financieros es de corto alcance y, por tanto reducimos el intervalo de análisis a 25 días antes y 25 días después de la publicación de beneficios.

El día de publicación de beneficios se representa en el 0. Para cada noticia correspondiente a una empresa, se calcula el número de días *hasta* el próximo reporte de beneficios, o el número de días *después* del último reporte de beneficios por esa empresa. Cada noticia contribuye solamente una vez al volumen de noticias en día  $t \in [-25, 25]$ , antes o después del reporte de beneficios (y no dos veces como en (Tetlock, Saar-Tsechansky y Macskassy, 2008)), y se realiza un *boxplot* de este volumen para cada período  $t$  de días.

FIGURA 4

**VOLUMEN DE NOTICIAS PUBLICADAS EN EL DOW JONES NEWSWIRES, DESDE 2012 A 2019, ALREDEDOR DE DÍAS DE ANUNCIO DE BENEFICIOS PARA EMPRESAS COTIZADAS EN EL S&P 500**



Fuente: Elaboración propia.

En todos los períodos, la distribución del número de noticias es muy asimétrica (todas las medias son más grandes que las medianas) y sus colas son pesadas por la derecha, excepto en el día que se hacen públicos los beneficios, donde parece más simétrica. En esta nueva gráfica podemos ver que el aumento más destacado en el volumen de noticias es el día exacto de publicación de beneficios, y el día inmediatamente posterior también tiene un aumento anormal con respecto al resto de la serie de volúmenes, lo que indica un aumento de noticias después de los hechos. Por lo demás se observa una estacionalidad semanal, seguramente debido a que el número de noticias depende del día de la semana, con menos volumen los fines de semana. El número de observaciones extremas cada día es pequeño: como máximo cinco empresas superan el límite estándar (1,5 veces el rango intercuartílico) para declarar el valor como un "valor atípico". Entonces, no podemos concluir de nuestra representación de la cobertura de los medios sobre los anuncios de beneficios que los sentimientos en las noticias pueden pronosticar indicadores fundamentales de una empresa (por ejemplo, precio-beneficios, precio-valor-libro, etc.) como se concluye en (Tetlock, Saar-Tsechansky y Macskassy, 2008), excepto quizás por las pocas empresas más populares en las noticias, los valores atípicos en nuestra gráfica. Sin embargo, especulamos que el sentimiento en las noticias después de las publicaciones de beneficios es el tipo de información útil para operar en corto, tal como ha sido considerado por Engelberg, Reed y Ringgenberg (2012).

#### **4.3. Sentimientos de orientación negativa están más vinculados a movimientos en los mercados financieros que los sentimientos de orientación positiva**

Este hecho se ha observado para varios mercados y empresas cotizadas en distintos sectores industriales. Algunas fuentes donde se han reportado ejemplos de este fenómeno son (Arias, Arratia y Xuriguera, 2013; Loughran y McDonald, 2011; Tetlock, Reed y Ringgenberg, 2008) y (Chan, 2003), aunque este último usa datos físicos de la prehistoria de internet. Allí se observa que esta mayor frecuencia en la dependencia entre negatividad y mercados es más evidente cuando se trata de empresas de media y pequeña capitalización.

#### **4.4. Mayor efecto de los sentimientos sobre las compañías de capitalización media y pequeña**

Este hecho se sugiere y es analizado por Chan (2003). Está relacionado con el hecho de que son los inversores minoristas quienes principalmente negocian en función del sentimiento de las noticias, y este tipo de inversores no mueven acciones de empresas de gran capitalización, un hecho que se argumenta extensamente en (Kumar y Lee, 2006).

## 5. HERRAMIENTAS PARA LA INFERENCIA ESTADÍSTICA

Para inferir algunas propiedades de los datos, útiles en la modelización de predictores, yendo más allá de la pura estadística descriptiva, es deseable tener a mano y poner en práctica un conjunto de contrastes estadísticos que desglosen las posibles relaciones entre los indicadores de sentimiento y las variables a predecir. En las próximas secciones pasamos revista a esos contrastes más utilizados en la inferencia de propiedades de las series temporales de nuestro interés.

Ejemplos de aplicación de estos contrastes estadísticos para determinar las posibles relaciones entre los indicadores de sentimiento de noticias y variables financieras se encuentran en los trabajos (Mendoza, 2018) y (Arratia y López, 2020).

### 5.1. Estacionaridad

En el contexto de variables económicas o sociales, se deberían intercambiar? normalmente observamos únicamente una realización del proceso estocástico subyacente que define estas variables. No es posible obtener muestras sucesivas o realizaciones independientes de dicho proceso. Para poder estimar las características “transversales” del proceso, como su media y su varianza, debemos asumir a partir de su evolución “longitudinal” que las propiedades transversales (distribución de las variables en cada instante en el tiempo) son estables en el tiempo. Esto conduce al concepto de estacionaridad.

Un proceso estocástico (serie temporal) es estacionario (o estrictamente estacionario) si las distribuciones marginales de todas las variables son idénticas y las distribuciones finito-dimensional de cualquier conjunto de variables dependen solo de la longitud de los retrasos que las separan. En particular, si la varianza es finita, la media y la varianza de todas las variables son las mismas. Más aún, la distribución conjunta de cualquier conjunto de variables es invariante por traslaciones (en el tiempo).

En general, estacionaridad estricta es una condición muy fuerte y difícil de verificar en la práctica, ya que se han de tener las distribuciones conjuntas para cualquier conjunto de variables dadas por el proceso. Nos conformamos entonces con una versión débil de estacionaridad, a saber, que sean invariantes por traslaciones temporales el primer y segundo momento de la distribución. Una serie temporal es débilmente estacionaria si  $EX_t$  es constante y  $EX_{t+h}X_t$  solamente depende de  $h$ .

La estacionaridad de una serie temporal puede ocasionalmente ser evaluada mediante un contraste de Dickey y Fuller (1979). Este no es exactamente un contraste de la hipótesis nula de estacionaridad, sino más bien un contraste de la existencia de una raíz unitaria en un proceso autoregresivo. La hipótesis alternativa es que el proceso, o bien es estacionario, o que es estacionario en tendencia (*i.e.*, estacionario después de remo-

verle la tendencia). Es importante tener en cuenta que la aplicación del contraste de Dickey-Fuller asume implícitamente que el proceso observado se rige por un modelo autoregresivo. Así que se debe al menos realizar previamente un contraste de autoregresión tipo ACF o PACF a los datos (cf. (Arratia, 2014, Ch. 2)).

## 5.2. Independencia

Antes de usar cualquier indicador como predictor, es importante determinar si existe alguna dependencia, en un sentido estadístico, entre el objetivo  $Y$  y el predictor  $X$ . De todos los contrastes estadísticos de dependencia existentes nuestro predilecto es el (relativamente nuevo) contraste basado en *correlaciones de distancias*, concebido por Székely *et al.* (2007).

Dadas dos variables aleatorias  $X$  e  $Y$  (posiblemente multivariadas), la correlación de distancias entre  $X$  e  $Y$  se calcula a partir de una muestra  $(X_1, Y_1), \dots, (X_n, Y_n)$ , mediante el siguiente esquema:

- Calcular la distancia euclídea entre todos los pares de observaciones de cada vector de muestras,  $X_i - X_j$  y  $Y_i - Y_j$ ,  $1 \leq i, j \leq n$ , para así obtener 2 matrices  $n \times n$  de distancias, una por cada vector.
- Centrar doblemente cada elemento de la matriz de distancia: a cada elemento restar la media de su fila y la media de su columna, y sumar la media de la matriz.
- Finalmente, calcular las covarianzas de las  $n^2$  distancias centradas.

La correlación de distancias se obtiene al normalizar las covarianzas anteriores de tal forma que cuando  $X = Y$  el resultado es 1. Se puede demostrar que cuando  $n \rightarrow \infty$ , las covarianzas de las distancias convergen a 0 si y solamente si los vectores  $X$  e  $Y$  son independientes.

A partir de esta somera descripción que hemos dado, deben ser obvios los puntos fuertes de la correlación de distancias, a saber: (1) caracteriza independencia; (2) se puede calcular para vectores numéricos, no únicamente para escalares; (3) al basarse en distancias,  $X$  e  $Y$  pueden tener dimensiones diferentes; (4) es invariante por rotaciones.

El contraste de independencia consiste en probar la hipótesis nula de cero correlación de distancias. Los  $p$  valores se obtienen mediante técnicas de *bootstrap*. El paquete *energy* de **R** (Rizzo y Székely, 2018) incluye las funciones `dcor` y `dcor.test` para calcular la correlación de distancias y el correspondiente contraste de independencia.

### 5.3. Causalidad de Granger

Es también importante evaluar la posibilidad de causalidad (y no solo dependencia) de un proceso aleatorio  $X_t$  hacia otro proceso aleatorio  $Y_t$ . En nuestro caso  $X_t$  es un indicador temporal de sentimiento e  $Y_t$  la serie de retornos del precio de un activo, o cualquier otro estadístico del precio que deseamos predecir. La idea básica de causalidad se debe a Granger (1969), quien la formuló de la siguiente manera:  $X_t$  causa  $Y_t$ , si  $Y_t$  se puede predecir mejor utilizando la información del pasado de  $X_t$  junto con el pasado de  $Y_t$ , que con solamente el pasado de  $Y_t$ . Formalmente se ha de considerar un modelo bivariado de autoregresión sobre  $X_t$  y  $Y_t$ , con  $Y_t$  dependiente de los pasados de  $X_t$  y  $Y_t$ , junto con un modelo lineal autoregresivo solo para  $Y_t$ , y probar la hipótesis nula de " $X_t$  no causa  $Y_t$ ", lo cual significa probar que todos los coeficientes que acompañan las observaciones pasadas de  $X_t$  en el modelo bivariado autoregresivo son cero. Bajo la presunción de que los datos tienen una distribución normal, se puede evaluar esta hipótesis nula mediante un F-test. Este modelo vectorizado de autoregresión para probar la causalidad de Granger se debe a Toda y Yamamoto (1995), y tiene la ventaja de funcionar bien aún en presencia de series no estacionarias. Para mayor detalle y ejemplos de aplicaciones en series temporales financieras ver (Arratia, 2014, Ch. 3).

Existen otras propuestas más recientes para probar causalidad entre series temporales basados en métodos no paramétricos, métodos de kernel y teoría de la información, entre otros, que hacen frente a la no linealidad y no estacionariedad, pero sin tener en cuenta la presencia de información secundaria (causalidad condicional), ver p. ej. (Diks y Wolski, 2015; Marinazzo, Pellicoro, y Stramaglia, 2008; Wibral et al., 2013). Para un contraste no paramétrico de causalidad condicional, ver (Arratia, Cabaña y Serès, 2016).

### 5.4. Selección de variables

El análisis de causalidad revela cualquier relación causa-efecto entre los indicadores de sentimiento y cualquiera de las funciones del precio de los activos financieros que tengamos como objetivo a predecir. El siguiente paso es analizar estos indicadores de sentimiento, individualmente o en conjunto, como variables independientes en un modelo de regresión para cualquiera de las variables financieras. Una razón fundamental para poner las variables juntas podría ser, al menos, lo que podrían tener en común semánticamente. Por ejemplo, juntar en un modelo de regresión todas las variables que expresan un sentimiento pesimista hacia el mercado (*bearish*), o todas aquellas que expresan un sentimiento optimista (*bullish*).

No obstante, en un período de tiempo determinado, no todas las variables de uno de estos grupos pueden causar tanto a la variable objetivo como a algunas del grupo, y su adición en el modelo podría agregar ruido en lugar de información de valor. Por tanto, conviene tratar con un modelo de regresión que *discrimine la importancia de las variables*.

Aquí es donde entonces proponemos implementar una regresión LASSO con todas las variables bajo consideración que potencialmente expliquen (causen) la variable objetivo. El método LASSO, debido a Tibshirani (Tibshirani, 1996), optimiza el error cuadrático medio entre la variable objetivo y la combinación lineal de los regresores, sujeto a una penalización de tipo  $L_1$  sobre los coeficientes de los regresores, lo que equivale a eliminar aquellos que son significativamente pequeños, eliminando así aquellas variables que aportan poco al modelo. LASSO no tiene en cuenta las posibles dependencias lineales entre los predictores, que pueden dar lugar a inestabilidades numéricas, por lo que recomendamos la verificación previa de las posibles correlaciones entre las variables a incluir en la regresión, descartando una de cada par de variables que estén altamente correlacionadas. Alternativamente, se puede intentar agregar una penalización de tipo  $L_2$  en los coeficientes de los regresores, lo que conduce a una red elástica (*elastic net*).

## 6. GUÍA DE SOFTWARE

Se dan aquí algunas sugerencias de software desarrollado en los lenguajes *R* y Python para el análisis de sentimiento de textos.

### *R*

En la actualidad existen varias funciones integradas en diversos paquetes de *R* para analizar el sentimiento en un documento y construir indicadores de sentimiento. A continuación, hacemos una breve revisión de las herramientas a disposición en *R*, diseñadas exclusivamente para el análisis de sentimientos en textos. Esta lista no es de ninguna manera exhaustiva, ya que continuamente se publican nuevas actualizaciones debido al creciente interés en el campo. Además otras herramientas de análisis de sentimientos ya están incluidas implícitamente en paquetes de minería de textos más generales como *tm* (Meyer Hornik y Feinerer, 2008), *openNLP* (Hornik, 2019) y *qdap* (Rinker, 2020). De hecho, muchos de los paquetes actuales específicos para el análisis de sentimiento tienen una fuerte dependencia en los paquetes de minería de textos antes mencionados, como también en otros paquetes pertenecientes a la *CRAN Task View on Natural Language Processing*<sup>4</sup>.

*SentimentAnalysis (2019-03)*: diseñado para el análisis de sentimiento de textos basado en diccionarios. El paquete contiene varios diccionarios generales (p. ej. Harvard IV), o específicos del lenguaje financiero (como el diccionario recopilado por Lougran-McDonald), y permite la creación de diccionarios propios. Para esto último incorpora una rutina de regularización LASSO, como herramienta estadística para seleccionar términos relevantes basado en variables características de contexto (Feuerriegel y Proellocks, 2019).

<sup>4</sup> <https://cran.r-project.org/web/views/NaturalLanguageProcessing.html>

*RSentiment (2018-07)*: sirve para analizar el sentimiento de una oración en inglés y asignar una valoración correspondiente al nivel del sentimiento. Puede clasificar oraciones en las siguientes categorías de sentimientos: Positivo, Negativo, Muy Positivo, Muy Negativo, Neutral. Para un vector de oraciones, cuenta el número de oraciones en cada categoría de sentimiento. Al calcular la valoración, se tienen en cuenta los cambiadores de valencia (Bose, 2018).

*sentimentr (2019-03)*: diseñado exclusivamente para calcular la polaridad en el sentimiento en un texto (Rinker, 2019).

*sentometrics (2019-11)*: un sistema integrado para la construcción y agregación de series temporales de sentimiento textual. Contiene todas las funciones necesarias para implementar cada una de las etapas del proceso descrito en la sección 2 para crear modelos de predicción basados en el sentimiento extraído de las noticias (Ardia *et al.*, 2019).

*quanteda (2019-11)*: diseñado para realizar análisis cuantitativo de datos textuales (Benoit y *et al.*, 2019).

*syuzhet (2017)*: sirve para extraer sentimiento y graficar el árbol de derivación de sentimiento de textos (Jockers, 2017).

## Python

Para los programadores de Python, también hay una gran cantidad de opciones para el análisis de sentimientos. De hecho, una búsqueda rápida de “Sentiment Analysis” en el índice de paquetes de Python (PyPI)<sup>5</sup> devuelve alrededor de 6000 registros. Aquí incluimos una lista reducida de los módulos más relevantes.

*Vader: Valence Aware Dictionary for sEntiment Reasoning* es un modelo basado en reglas (Hutto y Gilbert, 2014), principalmente entrenado en el análisis de textos sociales (por ejemplo, textos de medios sociales, reseñas de películas, etc.). Vader clasifica las oraciones en tres categorías: positivas, negativas y neutrales, con valores que representan las proporciones de las partes del texto que caen en cada categoría (la suma es 1 o cercana). También proporciona una valoración *compuesta* que se calcula sumando los valores normalizados de cada palabra en el léxico<sup>6</sup>. Una implementación de Vader se encuentra en la librería de uso general para el Procesamiento Natural del Lenguaje *nltk*.

*TextBlob*: para cualquier texto (en inglés) la librería TextBlob<sup>7</sup> provee de funciones para calcular el sentimiento en términos de polaridad, con valores continuos en  $[-1, 1]$ ,

<sup>5</sup> <https://pypi.org/>

<sup>6</sup> <https://github.com/cjhutto/vaderSentiment/#about-the-scoring>

<sup>7</sup> <https://textblob.readthedocs.io/en/dev/quickstart.html#sentiment-analysis>

y subjetividad, con valores en  $[0, 1]$ . Para los valores de subjetividad, 0 significa muy objetivo y 1 es muy subjetivo.

*Pattern*: es un paquete multipropósitos para minería de la web, tareas de procesamiento natural de lenguaje (NLP, en inglés), aprendizaje automático y análisis de redes. El sentimiento se estima como combinación de polaridad y subjetividad, y estos se pueden obtener a nivel de documento o de palabras (Smedt y Daelemans, 2012).

*pycorenlp*: provee de una interface para la librería en Java de la Standford CoreNLP de la cual hereda varias funcionalidades<sup>8</sup>. La lista completa de las funciones en el coreNLP pueden verse en el sitio web<sup>9</sup>.

## Referencias

ACHELIS, S. B. (2001). *Technical Analysis from A to Z*. New York: McGraw Hill.

ALGABA, A., ARDIA, D., BLUTEAU, K., BORMS, S., y BOUDT, K. (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys*, 34(3), pp. 512–547.

ANTWEILER, W. y FRANK, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of Finance*, 59(3), pp. 1259–1294.

AOUADI, A., AROURI, M., y TEULON, F. (2013). Investor attention and stock market activity: Evidence from france. *Economic Modelling*, 35, pp. 674–681.

ARDIA, D., BLUTEAU, K., BORMS, S., y Boudt, K. (2019). *The R Package sentometrics to Compute, Aggregate and Predict with Textual Sentiment*.

ARIAS, M., ARRATIA, A. y XURIGUERA, R. (2013). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1), p. 8.

ARRATIA, A. (2014). *Computational Finance: An Introductory Course with R*. Atlantis Press.

ARRATIA, A., CABAÑA, A. y SERÈS, A. (2016). Towards a sharp estimation of transfer entropy for identifying causality in financial time series. En: ECML-PKDD, *Proc. 1<sup>st</sup> Workshop MIDAS*. CEUR-WS. org.

ARRATIA, A. y LÓPEZ, A. (2020). Do google trends forecast bitcoins? stylized facts and statistical evidence. *Journal of Banking and Financial Technology*.

BAKER, M. y WURGLER, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2), pp. 129–152.

<sup>8</sup> <https://pypi.org/project/pycorenlp/>

<sup>9</sup> <https://stanfordnlp.github.io/CoreNLP/other-languages.html>

- BAUMEISTER, R. F., BRATSLAVSKY, E., FINKENAUER, C., y VOHS, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5(4), pp. 323–370.
- BECKERS, B., KHOLODILIN, K. A., y ULBRICHT, D. (2017). Reading between the lines: Using media to improve german inflation forecasts. *Technical report, DIW Berlin Discussion Paper*.
- BENOIT, K. y ET AL. (2019). *quanteda: Quantitative Analysis of Textual Data*. Version 1.5.2.
- BERRY, T. D. y HOWE, K. M. (1994). Public information arrival. *The Journal of Finance*, 49(4), pp. 1331–1346.
- BIFET, A. y FRANK, E. (2010). Sentiment knowledge discovery in Twitter streaming data. En: *International Conference on Discovery Science*, pages 1–15. Springer.
- BOSE, S. (2018). *Rsentiment: Analysis of Sentiment of English Sentences*. Version 2.2.2.
- BROWN, G. W. y CLIFF, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 11(1), pp. 1–27.
- CHAN, W. S. (2003). Stock price reaction to news and no-news: Drift and reversal after headlines. *Journal of Financial Economics*, 70(2), pp. 223–260.
- DERIU, J., GONZENBACH, M., UZDILLI, F., LUCCHI, A., LUCA, V. D., y JAGGI, M. (2016). Swisscheese at semeval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. En: *Proceedings of the 10th international workshop on semantic evaluation*, pp. 1124–1128.
- DERIU, J., LUCCHI, A., DE LUCA, V., SEVERYN, A., MÜLLER, S., CIELIEBAK, M., HOFMANN, T., y JAGGI, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. En: *Proceedings of the 26<sup>th</sup> international conference on world wide web*, pp. 1045–1052.
- DICKEY, D. A. y FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a), pp. 427–431.
- DIKS, C. y WOLSKI, M. (2015). Nonlinear granger causality: Guidelines for multivariate analysis. *Journal of Applied Econometrics*.
- DIMPFL, T. y JANK, S. (2015). Can internet search queries help to predict stock market volatility? *European Financial Management*, 22(2), pp. 171–192.
- DING, X., LIU, B., y YU, P. S. (2008). A holistic lexicon-based approach to opinion mining. En *Proc. of the 2008 Int. Conf. on Web Search and Data Mining*, pp. 231–240. ACM.
- ENGELBERG, J. E., REED, A. V., y RINGGENBERG, M. C. (2012). How are shorts informed?: Short sellers, news, and information processing. *Journal of Financial Economics*, 105(2), pp. 260–278.
- FEUERRIEGEL, S. y PROELLOCHS, N. (2019). *SentimentAnalysis: Dictionary-Based Sentiment Analysis*. Version 1.3-3.

- GO, A., BHAYANI, R., y HUANG, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12).
- GRANGER, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37, pp. 424–438.
- HAMID, A. y HEIDEN, M. (2015). Forecasting volatility with empirical similarity and google trends. *Journal of Economic Behaviour and Organization*, 117, pp. 62–81.
- HESTON, S. L. y SINHA, N. R. (2017). News vs. sentiment: Predicting stock returns from news stories. *Financial Analysts Journal*, 73(3), pp. 67–83.
- HORNIK, K. (2019). *openNLP: Apache OpenNLP Tools Interface. R Package Version 0.2.7*.
- HUTTO, C. J. y GILBERT, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. En: *8<sup>th</sup> Int. AAAI Conf. on Weblogs and Social Media*.
- JOCKERS, M. L. (2017). *Syuzhet: Extract Sentiment and Plot Arcs from Text. Version 1.0.4*.
- KUMAR, A. y LEE, C. M. (2006). Retail investor sentiment and return comovements. *The Journal of Finance*, 61(5), pp. 2451–2486.
- LIU, B. (2015). *Sentiment Analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.
- LOUGHRAN, T. y McDONALD, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1), pp. 35–65.
- MARINAZZO, D., PELLICORO, M., y STRAMAGLIA, S. (2008). Kernel method for nonlinear granger causality. *Physical Review Letters*, 100(14), pp. 144103.
- MENDOZA, D. (2018). Indices de sentimiento en el ámbito financiero. Master's thesis, Dept. Matemáticas y Estadística, Universitat Autònoma de Barcelona.
- MEYER, D., HORNIK, K., y FEINERER, I. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), pp. 1–54.
- MITCHELL, M. L. y MULHERIN, J. H. (1994). The impact of public information on the stock market. *The Journal of Finance*, 49(3), pp. 923–950.
- NAKOV, P., RITTER, A., ROSENTHAL, S., SEBASTIANI, F., y STOYANOV, V. (2016). Semeval-2016 task 4: Sentiment analysis in twitter. En: *Proceedings of the 10<sup>th</sup> International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1–18.
- POLANYI, L. y ZAENEN, A. (2006). Contextual valence shifters. En: *Computing Attitude and Affect in Text: Theory and Applications*, pp. 1–10. Springer.

- RAO, D. y RAVICHANDRAN, D. (2009). Semi-supervised polarity lexicon induction. En: *Proceedings of the 12<sup>th</sup> Conference of the European Chapter of the Association for Computational Linguistics*, pp. 675–682. Association for Computational Linguistics.
- RINKER, T. W. (2019). *sentimentr: Calculate Text Polarity Sentiment*. Version 2.7.1.
- (2020). *qdap: Quantitative Discourse Analysis*. Version 2.3.6. Buffalo, New York.
- RIZZO, M. L. y SZEKELY, G. J. (2018). *energy: E-Statistics: Multivariate Inference via the Energy of Data*. R package version 1.7-4.
- ROZIN, P. y ROYZMAN, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4), pp. 296–320.
- SCHOEN, H., GAYO-AVELLO, D., METAXAS, P. T., MUSTAFARAJ, E., STROHMAIER, M., y GLOOR, P. (2013). The power of prediction with social media. *Internet Research*.
- SEBASTIANI, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), pp. 1–47.
- SMEDT, T. D. y DAELEMANS, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13(Jun), pp. 2063–2067.
- SZÉKELY, G. J., RIZZO, M. L., BAKIROV, N. K., ET AL. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), pp. 2769–2794.
- TETLOCK, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62, pp. 1139–1168.
- TETLOCK, P. C., SAAR-TSECHANSKY, M., y MACSKASSY, S. (2008). More than words: Quantifying language to measure firm's fundamentals. *The Journal of Finance*, 63(3), pp. 1437–1467.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp. 267–288.
- TODA, H. Y. y YAMAMOTO, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, 66(1-2), pp. 225–250.
- TSAI, M.-F. y WANG, C.-J. (2014). Financial keyword expansion via continuous word vector representations. En: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1453–1458.
- UHL, M. W., PEDERSEN, M., y MALTIUS, O. (2015). What's in the news? using news sentiment momentum for tactical asset allocation. *The Journal of Portfolio Management*, 41(2), pp. 100–112.
- WIBRAL, M., PAMPU, N., PRIESEMAN, V., SIEBENHÜHNER, F., SEIWERT, H., LINDER, M., LIZIER, J., y VICENTE, R. (2013). Measuring information-transfer delays. *PLoS ONE*, 8(2), e55809.