



UNIVERSIDAD
Finis Terrae

Predicción de índices accionarios con Google Trends

Belén Castillo Barros
Diego Monasterio Pino

Escuela de Economía y Negocios, Universidad Finis Terrae

Profesor Guía
Nicolás Hardy Hernández

Escuela de Economía y Negocios, Universidad Finis Terrae

Tesis presentada a Facultad de Economía y Negocios de la Universidad Finis Terrae, para optar al título de Ingeniero Comercial.

Santiago, Chile
2021



UNIVERSIDAD
Finis Terrae

UNIVERSIDAD FINIS TERRAE
FACULTAD DE ECONOMÍA Y NEGOCIOS
ESCUELA DE INGENIERÍA COMERCIAL

Predicción de índices accionarios con Google Trends

Belén Castillo Barros
Diego Monasterio Pino

Proyecto de Tesis presentado a la Facultad de Economía y Negocios de la Universidad Finis Terrae, para optar al título de Ingeniería Comercial mención Finanzas.

Profesor(es) Guía:

Carolina Martínez (evaluador): nota 4,5
María José Quinteros (evaluador): nota 5,2
Nicolás Hardy (Guía): nota 5,2

Santiago, Chile
2021

1. Resumen Ejecutivo

El propósito de este trabajo es evaluar la capacidad predictiva de la herramienta Google Trends sobre los índices accionarios de las economías que componen la OCDE. Por tanto, se analiza cómo las frecuencias de búsqueda de algunas palabras con connotación positiva o negativa pueden predecir retornos accionarios. Para evaluar la capacidad predictiva se utiliza un análisis tanto dentro de muestra como fuera de muestra. La base de datos va desde el año 2010 al 2020. Esta investigación es novedosa pues es el primer intento en utilizar la herramienta de Google Trends para la predicción de índices accionarios. La intuición económica es que las frecuencias de búsqueda de Google Trends pueden funcionar como un proxy del sentimiento de mercado y de las expectativas que tienen los agentes sobre el futuro de la economía y por tanto, podría tener una capacidad predictiva sobre los retornos accionarios. En general, los resultados indican que hay escasa capacidad predictiva, ya que entre los resultados más importantes, de un total de 139 palabras para 9 países de las economías OCDE, el 20.5% son significativas para el estudio dentro de muestra, y 5.5% para fuera de muestra, donde se realizó un total de 46.287 regresiones. Dentro de las implicancias prácticas, puede servir a administradores de riesgos, portfolio de managers, etc. ya que este trabajo aporta a entender los drivers en la capacidad predictiva de retornos.

Agradecimientos

Gracias al profesor Nicolás Hardy por el tremendo apoyo que siempre brindó y a las familias por el apoyo que brindaron, en especial a Francisca Monasterio, Ximena Pino, Julio Monasterio, Daniela Arancibia, Manuel Altamirano, Carmen Barros, Andrés Castillo.

2. Introducción

El propósito de esta investigación es evaluar la capacidad predictiva que tiene la herramienta Google Trends sobre los índices accionarios de las economías que componen la OCDE con primera lengua inglés y español. La motivación económica es que las frecuencias de búsqueda en Google Trends pueden funcionar como un proxy del sentimiento de mercado y de las expectativas que tienen los agentes sobre el futuro de la economía y por tanto, podría tener una capacidad predictiva sobre los retornos accionarios.

El hecho estilizado en la literatura de pronósticos de activos financieros es que es sumamente difícil superar a simples modelos como el paseo aleatorio (RW). Un paper seminal en esta línea es Goyal y Welch (2008) quienes muestran que ninguno de los modelos de asset pricing elaborados por la literatura logra superar de manera significativa a un paseo aleatorio en términos de capacidad predictiva. En la misma línea, Timmermann (2008) encuentra que hay escasa capacidad predictiva en los retornos accionarios, esto se encuentra en breves periodos de predictibilidad y usualmente se ve confinada en periodos de crisis. Timmermann (2008), pp 1. Rossi y Sekhposyan (2011) identificaron tres posibles causas del beneficio de predicción de modelo: contenido predictivo, sobreajuste y capacidad de pronóstico variable en el tiempo. En su estudio, respecto a la primera causa, el autor indica si el ajuste dentro de la muestra predice el rendimiento del pronóstico fuera de la muestra. Por consiguiente, el sobreajuste es la etapa en la que un modelo incluye regresores irrelevantes, que mejoran el ajuste en la muestra del modelo y que, sin embargo, penaliza el modelo en un ejercicio de pronósticos fuera de muestra. Por último, la predicción variable en el tiempo es causada por cambios en los parámetros de los modelos. Estos resultados de baja predictibilidad no solamente se suscriben a retornos accionarios, sino también otro tipo de activos, por ejemplo, Meese,

Para el estudio se utilizaron los países de las economías OCDE con primera lengua inglés y español, estos países son: Australia, Canadá, Chile, Colombia, España, Estados Unidos, México, Nueva Zelanda y Reino Unido. Para descargar las series y ver las bases de datos ir a <http://www.wjh.harvard.edu/~inquirer/homecat.htm>

R.A y Rogoff (1983) encuentran que los modelos estructurales no mejoran en el modelo de caminata aleatoria a pesar de que los pronósticos se basan en valores reales realizados de futuras variables explicativas. Asimismo, señala que un prototipo aleatorio ejecuta perfectamente como cualquier modelo estimado en horizontes de uno a doce meses para los tipos de cambio dólar/libra, dólar/marca, dólar/yen y dólar ponderado por el comercio.

Sin embargo, estudios recientes de autores tales como Masuda y Takeda (2017, 2019), Choi y Varian (2012), Bulut (2018), Matsuki y Chang (2016), Adachi y Col. (2017), Onorante y Koop (2016), Bank et al. (2011), Ginsberg y Col (2009), Dimpfl y Jank (2016), Chojnowski y Dybka (2017) han mostrado el beneficio de utilizar nuevas herramientas tecnológicas como Google Trends que es un instrumento de acceso libre, que permite medir la frecuencia de búsqueda de varias palabras a través del tiempo. Las frecuencias de búsqueda son llevadas a una escala de 0 a 100, donde 100 es la fecha con mayor frecuencia de búsqueda de algún término o palabra clave. De esta manera, es posible ver en qué periodo de tiempo esta palabra ha sido buscada con mayor frecuencia según zona demográfica. Por ende, esta tesis intenta explotar el interés de los agentes como proxy de las expectativas que tienen sobre el comportamiento de las economías en el futuro.

En la misma línea, Google Trends recientemente ha captado la atención de académicos como Choi y Varian (2012) que descubren evidencia en la aplicación, siendo utilizada para anticipar y disminuir errores en las predicciones de la tasa de cambio, esto implica, una recopilación extensa de datos, creando un diccionario con aquellos términos relacionados al impacto en la tasa de cambio, para así, crear regresiones que arrojen resultados de la intensidad de estas palabras, es decir, búsquedas en Internet que expliquen el comportamiento de las fluctuaciones determinadas en un lapsus de tiempo. Matsuki y Chang (2016) analizan si existe capacidad predictiva sobre los índices accionarios de múltiples países usando Google Trends por sobre los benchmark clásicos que son RW, DRW y AR (1). Masuda y Takeda (2017) indican que Google Trends puede ser utilizado para explicar y predecir fenómenos de la vida real. Bulut (2018) utiliza un listado de términos de Google Trends que corresponden a los Índices de Sentimiento (IVS) con palabras que tengan relación con términos económicos, para lograr capturar de forma directa la percepción del mercado, aplicado a tres modelos que lograron vencer a un Random Walk. Adachi y Col. (2017) han encontrado hallazgos en el campo de las finanzas, ya que esta herramienta se ha utilizado para explicar los movimientos de precios de las acciones utilizando las frecuencias de búsqueda de nombres de compañías como proxy de la atención de los inversores. Onorante y Koop (2016) investigaron si los datos de los buscadores de Internet como Google pueden ayudar a mejorar los pronósticos actuales o las predicciones a corto plazo de las variables macroeconómicas. Bank et al. (2011) encuentran hallazgos en que las búsquedas en Google no sólo sirven como un proxy intuitivo para el reconocimiento general de la empresa, sino que también capta la atención de los inversores del mercado de valores ya que un aumento en las consultas de búsqueda está asociado con un aumento en la actividad comercial y liquidez de las acciones. Ginsberg y Col (2009) han demostrado que los datos de Google Trends pueden usarse para explicar y predecir ciertos fenómenos de la vida real, en este caso lo utilizan para detectar posibles epidemias de influenza monitoreando el comportamiento de búsqueda de salud, buscando frecuencias de consultas en línea como la aplicación Google. Dimpfl y Jank (2016) concluyen que las consultas de búsqueda están fuertemente correlacionadas contemporáneamente con la volatilidad del mercado y que las consultas de búsqueda de hoy (positivamente) predicen la volatilidad de mañana y estos hallazgos son consistentes con la hipótesis del operador de ruido. Chojnowski y Dybka (2017) amplían el modelo del valor actual, basado en los resultados observados incluyendo tres principios no estudiados: sentimientos del mercado de crédito, del mercado financiero y el mercado de precios, esto lo hicieron desarrollando un método de extracción de sentimientos de los datos de Google Trends sobre búsquedas en diferentes mercados, este método de extracción de sentimientos consiste en medir los fundamentos no observables (sentimientos del mercado) sobre la base de las series de tiempo de Google Trends en consultas específicas buscadas en Google, observando el optimismo y el pesimismo de los consumidores reflejados en búsquedas de Google.

En este trabajo, para evaluar la capacidad predictiva se utiliza un análisis tanto dentro como fuera de muestra. Para el análisis fuera de muestra se utiliza la prueba ENC-NEW de Clark y McCracken (2001) y la prueba ENC-T de Clark y West (2006, 2007).

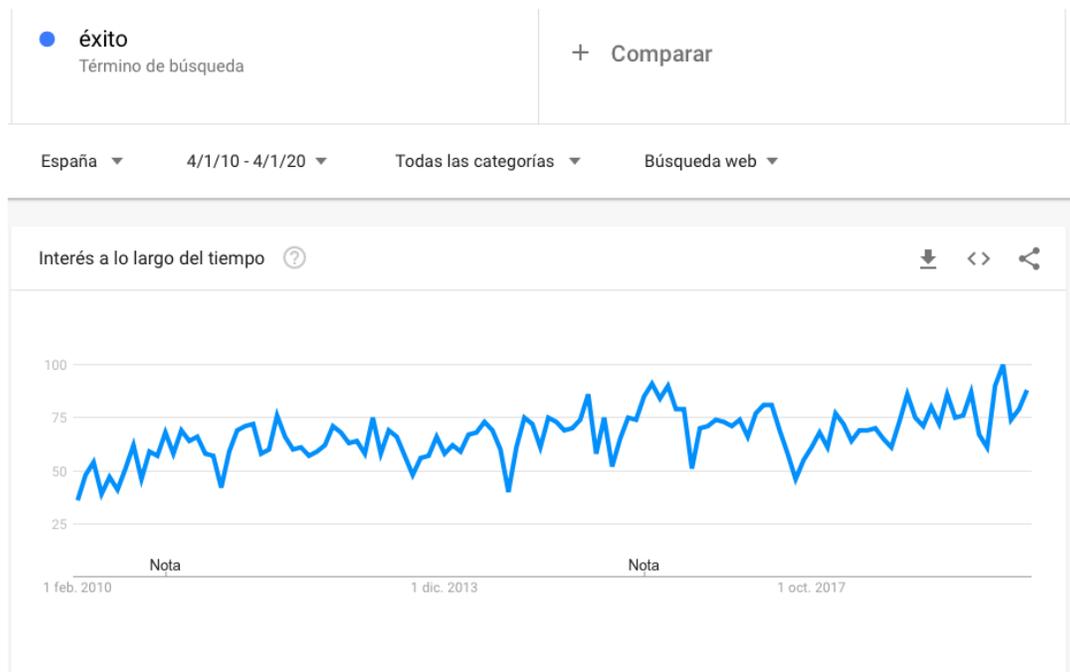
Sin embargo, también existe escasa evidencia de capacidad predictiva con autores como Mark (1995), Engel et al. (2007) que muestran que algunos modelos estructurales tienen capacidad predictiva sobre el paseo aleatorio. Además, Mark (1995) descubre hallazgos en que los pronósticos de análisis fuera de muestra de los tipos de cambio superan el DRW.

La estructura del artículo está dividido en cuatro partes: Resumen ejecutivo, introducción, datos y metodología y finalmente, resultados y discusión.

4. Datos y metodología

La base de datos va desde el año 2010 hasta el 2020 en frecuencias mensuales. Siguiendo a Da, Engelberg y Gao (2015) y Masuda y Takeda (2019) se utilizó el Diccionario de Harvard IV-4 para obtener las palabras con connotación positiva y negativa. Cabe destacar que las palabras están en inglés y se realizó la traducción literal para los países en español. Para los índices bursátiles se descargó data en la misma frecuencia desde Thomson Reuters Datastream.

Figura 1: Gráfico de búsqueda de palabras con la herramienta Google Trends.



Nota: Esta es una serie de la palabra éxito para el país de España con un periodo de búsqueda de 10 años.

Fuente: Google Trends.

En este gráfico se observa cómo varían las frecuencias de búsqueda de palabras de la herramienta Google Trends.

Para los análisis dentro y fuera de muestra se utilizaron las siguientes especificaciones:

$$\Delta \ln(Y_{t+1}) = c + \beta \Delta \ln(X_t) + \rho \Delta \ln(Y_t) + \varepsilon_{1,t+1} \quad (1)$$

$$\Delta \ln(Y_{t+1}) = c + \beta \Delta \ln(X_t) + \varepsilon_{2,t+1} \quad (2)$$

$$\Delta \ln (Y_{t+1}) = \beta \Delta \ln (X_t) + \varepsilon_{3,t+1} \quad (3)$$

Donde:

$$\begin{aligned} \Delta \ln (Y_{t+1}) &= \ln (Y_{t+1}) - \ln (Y_t) \\ \Delta \ln (X_{t+1}) &= \ln (X_{t+1}) - \ln (X_t) \end{aligned}$$

Donde $\Delta (Y_t)$ representa la diferencia logarítmica del retorno del índice accionario, $\Delta \ln (X_t)$ corresponde a la diferencia logarítmica de la frecuencia de búsqueda de Google Trends, c representa a la constante, $\varepsilon_{i,t}$, $i = 1, 2, 3$ son los términos de error, y por último, ρ y β son los coeficientes de regresión.

En la especificación (1) el pronóstico de los retornos de los índices accionarios de un periodo se establece como una constante (c) más un rezago de la variable Google Trends que se expresa en diferencia logarítmica y más una componente autorregresiva para el retorno de los índices accionarios de cada país en un tiempo t . En la especificación (2) es similar a la especificación (1) pero con la diferencia que no se incluye la componente autorregresiva, es decir, el rezago del retorno de Y . Finalmente, en la especificación (3), se elimina la constante c y la componente autorregresiva de orden 1.

La intuición de estos resultados es compararse contra los benchmarks clásicos de la literatura, que es un AR(1), paseo aleatorio con drift, paseo aleatorio sin drift y se utilizan estos benchmarks por la vasta literatura que ha mostrado que los modelos autoregresivos simples y los paseos aleatorios son benchmarks sumamente difíciles de vencer. Goyal y Welch (2008), Meese y Rogoff (1983).

Al final de la investigación hay un apéndice suplementario de las palabras que fueron eliminadas de la investigación ya que no tenían relación con índices de sentimiento de mercado económico.

4.1 Análisis dentro de muestra (IS)

La siguiente hipótesis establece bajo la nula que las palabras descargadas por Google Trends no tienen capacidad predictiva sobre el índice accionario en el futuro, por tanto un rechazo en la hipótesis nula implica la existencia de capacidad predictiva. Los t-stadistic para esta prueba de hipótesis se calcularon utilizando estimadores de la varianza a largo plazo de HAC de Newey y West (1987, 1994).

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

Por ende, se busca encontrar hallazgos y ver si Google Trends tiene capacidad predictiva sobre los benchmark clásicos Random Walk (RW), paseo aleatorio sin drift (DRW) y el prototipo autorregresivo simple AR (1).

4.2 Análisis fuera de muestra (OOS)

Para realizar las evaluaciones fuera de muestra se aplicaron las pruebas ENC-NEW y ENC-T de Clark y McCracken (2001) y Clark y West (2006, 2007). Siguiendo la línea de hallazgos de estos autores, se dividió la base de datos en dos, una ventana de estimación de tamaño R y una ventana de evaluación de tamaño P , tal que $P + R = T$, donde T es el número total de observaciones. Con la ventana inicial R se realizaron las estimaciones en nuestros modelos y se evaluaron los pronósticos en la ventana de evaluación P . Para evaluar estos pronósticos, se utilizaron las pruebas ENC-NEW y ENC-T. Ambas pruebas poseen distribuciones asintóticas no estándar, pero los valores críticos están disponibles en el paper de Clark y McCracken

(2001). En general, la distribución asintótica de estos test depende del método de estimación de los modelos, ya sea, ventana rodantes, recursivas o fijas. Ver West (2006) para más detalles. El parámetro π que se define como el límite de P/R y del número de excesos de parámetro del modelo a evaluar que en nuestro caso es 1. Para la evaluación de pronósticos se empleó una vasta literatura que prueba que derrotar modelos autorregresivos y paseos aleatorios es un trabajo sumamente complejo. Ver Meese, R.A y Rogoff (1983, 1987), Goyal y Welch (2008), Timmermann (2008). Con este propósito la figura 2 muestra las especificaciones econométricas que se proponen utilizando la serie Google Trends y se comparó contra los benchmark clásicos, que son AR(1), RW y DRW en cada caso.

Figura 2: Especificaciones econométricas:

Especificación econométrica	Benchmark
$Y_{t+1} = c + \rho Y_t + \beta x_t + \varepsilon_{t+1}$	$Y_{t+1} = c + \rho Y_t + \varepsilon_{t+1}$
$Y_{t+1} = c + \beta x_t + \varepsilon_{t+1}$	$Y_{t+1} = c + \varepsilon_{t+1}$
$Y_{t+1} = \beta x_t + \varepsilon_{t+1}$	$Y_{t+1} = \varepsilon_{t+1}$

Nota: En la tabla se presentan las especificaciones econométricas usadas para compararnos contra los benchmark.

Fuente: Elaboración propia.

Así para evaluar la capacidad predictiva en cada comparación, el modelo va a derrotar al benchmark en la medida en que se rechaza la hipótesis nula $H_0: \beta = 0$ y la hipótesis alternativa $H_1: \beta \neq 0$. Como sugiere Rossi (2005) mostrar resultados fuera de muestra no elimina completamente las preocupaciones de overfitting y siguiendo sus sugerencia se evaluaron nuestros pronósticos fuera de muestra con más de una posible edición de ventana, así se realizaron tres evaluaciones distintas, la primera utilizando 50% para estimar y 50% para evaluar (P/R=1), luego 70% para estimar y 30% para evaluar (P/R≈2) y finalmente, 30% para estimar y 70% para evaluar (P/R≈0.4). Se Dispone de los niveles de confianza de los autores Clark y McCracken (2001) y Clark y West (2007). En la especificación econométrica se asumió el ruido blanco como 0. Además, Y_{t+1} es variable dependiente y las otras variables son independientes.

5. Resultados y discusión

5.1 Resultados dentro de muestra (IS)

Considerando que son 9 países y 139 palabras, en este capítulo se muestra una sección de los resultados más interesantes por palabra y país y los apéndices complementarios a esta tesis muestran la totalidad de los resultados.

Todas las regresiones son siguiendo la especificación 1 evaluando la hipótesis nula $H_0: \beta = 0$ y la hipótesis alternativa $H_1: \beta \neq 0$ utilizando el estadístico t, con estimadores de la varianza a largo plazo siguiendo a HAC Newey & West (1987, 1984).

Las tablas 1-3 presentan los resultados para los 9 países bajo análisis y cada tabla reporta los resultados para 2 palabras que vale la pena destacar.

La tabla 1 reporta resultados para dos palabras que tiene connotación positiva en el Diccionario de Harvard IV-4, mientras que la tabla 2 reporta resultados para dos palabras que tiene connotación negativa en el Diccionario de Harvard IV-4, finalmente, la tabla 3 reporta resultados para dos palabras sin connotación en el Diccionario de Harvard IV-4.

En general, la componente autorregresiva da significativa pocas veces, eso es consistente con la vasta evidencia de que los retornos accionarios son poco persistentes. Por otro lado, los R^2 en la tabla van entre 0.001 y 0.145 por lo que en general

son pequeños y es consistente porque es difícil explicar buena parte de la varianza de los retornos futuros. Respecto a las palabras de Google Trends, en cuanto al signo, es poco intuitivo en las regresiones ya que en las palabras positivas se espera que el signo será positivo, sin embargo, no ocurre siempre.

Tabla 1: Regresiones palabras positivas

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Australia</i>	<i>Canadá</i>	<i>Chile</i>	<i>Colombia</i>	<i>España</i>	<i>Estados Unidos</i>	<i>México</i>	<i>Nueva Zelanda</i>	<i>Reino Unido</i>
<i>Éxito</i>	-0,003 (0,002)	0 (0,002)	0,01*** (0,003)	-0,001 (0,004)	-0,011 (0,011)	0,006*** (0,002)	0,007 (0,007)	-0,001 (0,006)	0,013* (0,007)
<i>AR(1)</i>	-0,145** (0,055)	-0,284*** (0,099)	0,016 (0,081)	0,107 (0,087)	-0,08 (0,074)	-0,252** (0,096)	-0,134 (0,105)	-0,225 (0,152)	-0,18** (0,075)
<i>Constante</i>	0,001 (0,004)	0,002 (0,004)	0,004 (0,003)	0 (0,004)	-0,004 (0,006)	0,009*** (0,003)	0,006 (0,007)	0,005 (0,006)	0,001 (0,004)
<i>Observaciones</i>	119	119	119	119	119	119	119	119	119
<i>Invertir</i>	-0,002 (0,005)	-0,002 (0,003)	-0,004 (0,003)	0,007*** (0,002)	0,002 (0,005)	0,002 (0,005)	0,003 (0,007)	0,006* (0,003)	0,007*** (0,001)
<i>AR(1)</i>	-0,165*** (0,053)	-0,285*** (0,089)	0,02 (0,086)	0,113 (0,078)	-0,089 (0,069)	-0,26*** (0,095)	-0,14 (0,118)	-0,244 (0,152)	-0,234*** (0,066)
<i>Constante</i>	0,001 (0,004)	0,002 (0,005)	0,004 (0,003)	0 (0,004)	-0,004 (0,006)	0,009** (0,004)	0,007 (0,006)	0,006 (0,006)	0,001 (0,004)
<i>Observaciones</i>	119	119	119	119	119	119	119	119	119

Nota: Esta tabla presenta las regresiones de 2 palabras y 9 países de las economías OCDE, las palabras están basadas en el Diccionario de Harvard IV-4. Se usó 10%, 5% y 1% de niveles de significancia. Los paréntesis son las desviaciones estándar de las regresiones. El término AR(1) es el coeficiente autorregresivo de la regresión. Siendo $p^* < 10\%$, $p^{**} < 5\%$, $p^{***} < 1\%$.

Fuente: Elaboración propia.

Tabla 2: Regresiones palabras negativas

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Australia</i>	<i>Canadá</i>	<i>Chile</i>	<i>Colombia</i>	<i>España</i>	<i>Estados Unidos</i>	<i>México</i>	<i>Nueva Zelanda</i>	<i>Reino Unido</i>
<i>Cobrar</i>	-0,001 (0,005)	0,008** (0,003)	0,003 (0,002)	-0,008*** (0,003)	-0,012 (0,012)	0,011** (0,005)	0,002 (0,006)	-0,005 (0,005)	-0,004 (0,002)
<i>AR(1)</i>	-0,152** (0,058)	-0,272*** (0,096)	0,018 (0,088)	0,134* (0,074)	-0,101 (0,083)	-0,234*** (0,077)	-0,15 (0,109)	-0,213 (0,152)	-0,206*** (0,063)
<i>Constante</i>	0,001 (0,004)	0,002 (0,005)	0,004 (0,003)	0,000 (0,004)	-0,004 (0,006)	0,009** (0,004)	0,007 (0,004)	0,005 (0,006)	0,001 (0,005)
<i>R²</i>	0,024	0,010	0,005	0,051	0,032	0,145	0,020	0,057	0,054
<i>Observaciones</i>	119	119	119	119	119	119	119	119	119
<i>Jefe</i>	0,000 (0,002)	0,006*** (0,002)	-0,001 (0,002)	-0,012*** (0,003)	0,007** (0,004)	0,000 (0,006)	-0,008* (0,004)	0,008 (0,008)	0,005** (0,002)
<i>AR(1)</i>	-0,154*** (0,057)	-0,273*** (0,101)	0,014 (0,087)	0,131* (0,067)	-0,056 (0,067)	-0,263** (0,106)	-0,147 (0,111)	-0,201 (0,129)	-0,211*** (0,067)
<i>Constante</i>	0,001 (0,004)	0,002 (0,004)	0,004 (0,003)	0,000 (0,004)	-0,004 (0,006)	0,009** (0,004)	0,007 (0,004)	0,005 (0,006)	0,001 (0,004)
<i>R²</i>	0,024	0,092	0,001	0,058	0,026	0,063	0,036	0,067	0,055
<i>Observaciones</i>	119	119	119	119	119	119	119	119	119

Nota: Esta tabla presenta las regresiones de 2 palabras y 9 países de las economías OCDE, las palabras están basadas en el Diccionario de Harvard IV-4. Se usó 10%, 5% y 1% de niveles de significancia. Los paréntesis son las desviaciones estándar de las regresiones. El término AR(1) es el coeficiente autorregresivo de la regresión. Siendo $p^* < 10\%$, $**p < 5\%$, $***p < 1\%$.

Fuente: Elaboración propia.

Tabla 3: Regresiones palabras sin connotación

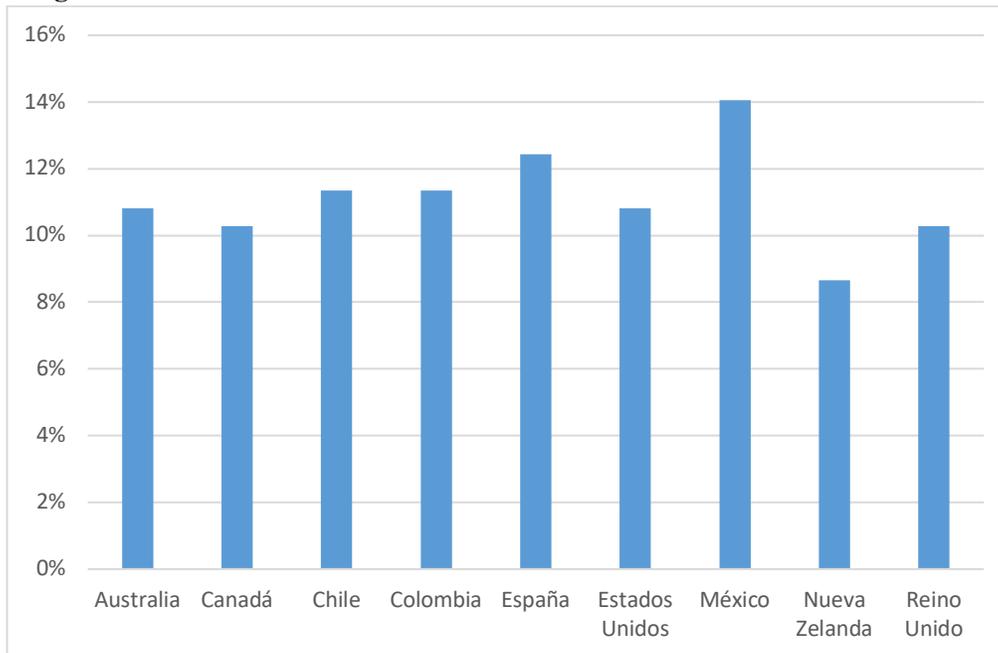
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Australia</i>	<i>Canadá</i>	<i>Chile</i>	<i>Colombia</i>	<i>España</i>	<i>Estados Unidos</i>	<i>México</i>	<i>Nueva Zelanda</i>	<i>Reino Unido</i>
<i>Gastar</i>	-0,004*** (0,001)	0,014 (0,014)	-0,005*** (0,001)	0,007 (0,013)	-0,033** (0,015)	0,007 (0,008)	0,074* (0,042)	0,000 (0,001)	-0,016 (0,011)
<i>AR(1)</i>	-0,113* (0,057)	-0,279*** (0,079)	0,049 (0,083)	0,118* (0,069)	-0,012 (0,011)	-0,206*** (0,064)	-0,135 (0,115)	-0,174 (0,115)	-0,216*** (0,064)
<i>Constante</i>	0,002 (0,003)	0,002 (0,004)	0,004 (0,003)	0,000 (0,004)	-0,003 (0,006)	0,010*** (0,003)	0,006 (0,004)	0,006 (0,006)	0,001 (0,004)
<i>R²</i>	0,002	0,089	0,023	0,016	0,010	0,043	0,045	0,032	0,063
<i>Observaciones</i>	119	119	119	119	119	119	119	119	119
<i>Importar</i>	0,127** (0,050)	0,156** (0,061)	-0,039 (0,031)	0,011 (0,020)	0,129* (0,067)	0,322*** (0,093)	0,087 (0,082)	-0,024 (0,063)	0,099 (0,066)
<i>AR(1)</i>	-0,130*** (0,047)	-0,283*** (0,089)	0,039 (0,091)	0,108 (0,068)	-0,001 (0,011)	-0,180*** (0,034)	-0,157 (0,106)	-0,173 (0,118)	-0,205*** (0,064)
<i>Constante</i>	0,002 (0,003)	0,002 (0,004)	0,004 (0,003)	0,000 (0,004)	-0,003 (0,006)	0,011*** (0,002)	0,007* (0,004)	0,006 (0,006)	0,001 (0,004)
<i>R²</i>	0,041	0,103	0,014	0,015	0,023	0,117	0,028	0,032	0,058
<i>Observaciones</i>	119	119	119	119	119	119	119	119	119

Nota: Esta tabla presenta las regresiones de 2 palabras y 9 países de las economías OCDE, las palabras están basadas en el Diccionario de Harvard IV-4. Se usó 10%, 5% y 1% de niveles de significancia. Los paréntesis son las desviaciones estándar de las regresiones. El término AR(1) es el coeficiente autorregresivo de la regresión. Siendo $p^* < 10\%$, $**p < 5\%$, $***p < 1\%$.

Fuente: Elaboración propia.

En la tabla 1 se presentan dos palabras con connotación positiva, donde sólo en cinco países las dos palabras dan significativas y sólo para uno de ellos la constante se comporta igual mostrando significancia. Por otra parte, la tabla 2 representa dos palabras con connotación negativa, de las cuales en seis países las palabras fueron significativas, y, sólo en un país la constante fue significativa para ambas palabras. Finalmente, en la tabla 3 se presentan dos palabras sin connotación, donde ocho veces se obtiene significancia de las palabras en seis países de nueve en total, y, en algunos casos la constante es significativa para las dos palabras en un mismo país. En conclusión, del total sólo el 20.5% de palabras son significativas en cada país y aportan a la predicción del índice accionario correspondiente.

Figura 3: Palabras significativas dentro de muestra



Fuente: Elaboración propia.

Nota: Este gráfico representa las palabras significativas en el modelo dentro de muestra por país.

En la gráfica se percibe el total de palabras significativas que logran predecir el índice accionario del país correspondiente del modelo dentro de muestra y finalmente, un 20.5% en promedio de palabras por país entrega resultados alentadores. Igualmente, se puede ver que los nueve países OCDE tienen resultados similares entre ellos.

5.2 Resultados fuera de muestra (OOS)

En esta sección se muestra una colección de los resultados más llamativos del estudio, el resultado para el resto de las palabras está disponible en un apéndice suplementario a la tesis.

Las tablas 4 a 12 muestran una selección de 10 palabras con resultados interesantes por país. En particular, se realizaron las pruebas ENC-NEW y ENC-T donde se utilizaron métodos de estimación rodante y recursivo. Por brevedad en esta tesis solamente se presentan resultados para el ENC-T con ventanas recursivas, mientras que los resultados para el ENC-NEW con ventana rodante y recursiva están disponibles en un apéndice suplementario.

Cada tabla muestra 3 paneles, un panel donde la división es 50% para estimar y 50% para evaluar (P/R=1), luego 70% para estimar y 30% para evaluar (P/R≈2) y finalmente, 30% para estimar y 70% para evaluar (P/R≈0.4).

Tabla 4: Tablas fuera de muestra para los mejores resultados de Australia

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Contaminación</i>	<i>Corredor de bolsa</i>	<i>Crédito</i>	<i>Desempleado</i>	<i>Emprendedor</i>	<i>Gran éxito</i>	<i>Hombre de negocios</i>	<i>Intervención</i>	<i>Negociar</i>	<i>Transparente</i>
	50/50									
DRW	1.056*	1.388**	1.538**	1.069*	3.435***	1.110*	1.447**	1.318*	2.042**	1.479**
RW	1.062*	1.412**	1.719**	1.086*	4.786***	1.112*	1.445**	1.466**	2.096***	1.491**
AR(1)	1.109*	1.320*	1.406**	1.058*	2.737***	1.102*	1.502**	1.283*	1.664**	1.577**
	70/30									
DRW	1.212*	1.218*	1.731**	1.088*	3.098***	1.120*	1.387**	1.462**	1.511**	1.706**
RW	1.194*	1.145*	1.834**	1.085*	8.435***	1.116*	1.363**	1.282*	1.298*	1.624**
AR(1)	1.256*	1.155*	1.484**	1.073*	2.892***	1.115*	1.448**	1.460**	1.434**	1.705**
	30/70									
DRW	1.238*	1.193*	1.300*	1.028*	2.389***	1.142*	1.377**	0.991*	1.346*	1.108*
RW	1.239*	1.165*	1.415**	1.027*	2.945***	1.143*	1.360**	1.010*	1.383**	1.115*
AR(1)	1.306*	1.215*	1.277*	1.005*	2.342***	1.151*	1.406**	1.035*	1.162*	1.269*

Nota: Esta tabla presenta las regresiones fuera de muestra para Australia, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 5: Tablas fuera de muestra para los mejores resultados de Canadá

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Asociado</i>	<i>Escasez</i>	<i>Gerente</i>	<i>Jefe</i>	<i>Nobleza</i>	<i>Pago</i>	<i>Poblacional</i>	<i>Preocupación</i>	<i>Recompensa</i>	<i>Subasta</i>
	50/50									
DRW	1.742**	1.124*	1.849**	1.413**	1.152*	0.994*	1.546**	1.250*	1.526**	1.376**
RW	1.697**	1.116*	1.928**	1.446**	1.104*	0.998*	1.583**	1.245*	1.598**	1.375**
AR(1)	1.239*	1.092*	2.294***	1.035*	1.418**	0.960*	1.233*	1.205*	1.618**	1.316*
	70/30									
DRW	2.292***	1.063*	1.456**	1.484**	1.052*	1.050*	1.254*	1.278*	1.498**	1.366**
RW	1.946**	1.050*	1.501**	1.523**	1.035*	1.048*	1.270*	1.253*	1.541**	1.311*
AR(1)	1.569**	1.052*	1.995**	1.077*	0.855	1.042*	0.866	1.618**	1.557**	1.323*
	30/70									
DRW	1.98**	1.116*	1.280*	1.310*	1.901**	1.066*	1.372**	1.365**	1.794**	1.644**
RW	1.874**	1.065*	1.261*	1.323*	1.807**	1.063*	1.360**	1.302*	1.821**	1.520**
AR(1)	2.105***	1.090*	2.257***	0.974*	2.168***	1.061*	1.171*	1.366**	1.874**	1.444**

Nota: Esta tabla presenta las regresiones fuera de muestra para Canadá, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 6: Tablas fuera de muestra para los mejores resultados de Chile

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Ambición</i>	<i>Capitalismo</i>	<i>Cooperativa</i>	<i>Corredor de bolsa</i>	<i>Cuenta Bancaria</i>	<i>Extravagante</i>	<i>Gerente</i>	<i>Gran éxito</i>	<i>Profesión</i>	<i>Transparente</i>
	50/50									
DRW	1.460**	2.054***	1.214*	2.894***	1.484**	2.101***	1.174*	1.972**	2.034**	2.045**
RW	1.518**	1.993**	1.213*	2.773***	1.479**	2.190***	0.963*	2.020**	2.335***	0.485
AR(1)	1.557**	2.249***	1.219*	2.985***	1.449**	1.660**	1.186*	2.095***	1.999**	1.929**
	70/30									
DRW	1.142*	1.274*	0.983*	4.336***	1.160*	2.173***	1.090*	1.615**	1.754**	1.345*
RW	1.169*	1.004*	0.876	3.168***	1.092*	2.138***	0.656	1.503**	1.755**	-0.748
AR(1)	1.202*	1.545**	0.949*	11.92***	1.361**	2.066***	1.170*	1.613**	1.827**	1.727**
	30/70									
DRW	2.374***	2.005**	1.288*	2.631***	0.981*	1.214*	0.800	1.363**	1.251*	1.650**
RW	2.510***	2.163***	1.483**	2.748***	1.117*	1.498**	1.054*	1.445**	1.658**	1.156*
AR(1)	2.388***	2.229***	1.313*	2.776***	1.124*	1.157*	1.000*	1.267*	1.280*	0.909

Nota: Esta tabla presenta las regresiones fuera de muestra para Chile, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 7: Tablas fuera de muestra para los mejores resultados de Colombia

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Comprobar</i>	<i>Consejo</i>	<i>Corredor de bolsa</i>	<i>Desempleado</i>	<i>Estafa</i>	<i>Gerente</i>	<i>Jefe</i>	<i>Negociar</i>	<i>Oro</i>	<i>Vendedor</i>
DRW	50/50									
	1.214*	1.065*	1.683**	1.325*	2.198***	0.609	1.543**	1.765**	1.857**	2.001**
RW	0.652	1.029*	1.62**	0.882	1.809**	0.504	1.462**	1.364**	1.751**	1.848**
AR(1)	1.018*	1.018*	1.016*	1.043*	1.021*	1.016*	1.013*	1.014*	1.015*	1.016*
DRW	70/30									
	1.818**	1.478**	0.955*	1.159*	0.778	1.316*	0.815	1.744**	1.914**	2.372***
RW	1.265*	1.426**	0.909	0.803	0.573	1.200*	0.633	1.731**	1.746**	2.309***
AR(1)	1.030*	1.030*	1.030*	1.058*	1.022*	1.029*	1.025*	1.028*	1.028*	1.029*
DRW	30/70									
	0.023	1.117*	0.115	1.386**	2.346***	1.736**	1.809**	1.406**	2.340***	1.092*
RW	0.174	1.132*	0.088	1.401**	2.302***	1.653**	1.907**	1.442**	2.204***	1.118*
AR(1)	1.012*	1.013*	1.011*	1.037*	1.026*	1.011*	1.008*	1.009*	1.010*	1.011*

Nota: Esta tabla presenta las regresiones fuera de muestra para Colombia, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 8: Tablas fuera de muestra para los mejores resultados de España

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Adquisición</i>	<i>Bienes</i>	<i>Empresarial</i>	<i>Empresario</i>	<i>Escasez</i>	<i>Invertir</i>	<i>Patrocinador</i>	<i>Pertenecer</i>	<i>Pobre</i>	<i>Privilegio</i>
	50/50									
DRW	0.941	1.119*	2.167***	0.648	1.364**	0.965*	1.330*	1.439**	1.356**	1.072*
RW	0.645	1.094*	1.995**	0.677	1.369**	0.982*	1.353**	1.490**	1.333**	1.056*
AR(1)	1.580**	1.073*	1.975**	0.722	1.313*	1.026*	1.061*	1.467**	1.328*	1.138*
	70/30									
DRW	1.181*	1.305*	2.358***	1.774**	1.385**	1.11*	1.049*	1.162*	1.243*	0.989*
RW	1.158*	1.287*	2.332***	1.837**	1.359*	1.106*	1.040*	1.142*	1.227*	0.972*
AR(1)	1.654**	0.082	2.378***	1.686**	0.100	1.115*	0.521	1.142*	1.307*	1.049*
	30/70									
DRW	1.110*	1.143*	1.646**	0.996*	1.993**	0.911	1.992**	1.372**	-1.193	1.159*
RW	0.960	1.101*	1.816**	1.060*	2.011**	0.874	1.864**	1.258*	-1.233	1.072*
AR(1)	1.574**	1.413**	1.755**	1.609**	2.181***	0.972*	1.632**	0.254	-1.364	1.237*

Nota: Esta tabla presenta las regresiones fuera de muestra para España, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 9: Tablas fuera de muestra para los mejores resultados de Estados Unidos

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Benefactor</i>	<i>Burocracia</i>	<i>Colonia</i>	<i>Emprendedor</i>	<i>Fortuna</i>	<i>Inflación</i>	<i>Jefe</i>	<i>Jugar</i>	<i>Legar</i>	<i>Patrocinador</i>
	50/50									
DRW	2.229***	1.060*	1.338**	1.678**	1.910**	0.886	1.490**	1.672**	1.126*	1.182*
RW	2.545***	1.078*	1.350**	1.740**	1.779**	2.178***	1.492**	1.894**	1.128*	1.200*
AR(1)	2.462***	1.101*	1.208	1.412**	1.324*	1.482**	1.577**	1.356**	1.155*	1.231*
	70/30									
DRW	1.414**	1.045*	1.232*	1.433**	1.163*	1.342*	1.457**	1.356*	1.054*	1.155*
RW	1.577**	1.065*	1.208*	1.400**	1.155*	1.428**	1.461**	1.486**	1.066*	1.164*
AR(1)	1.417**	1.035*	1.172*	1.259*	1.123*	1.189*	1.468**	0.990*	1.082*	1.177*
	30/70									
DRW	2.323***	1.087*	1.308*	1.163*	1.515**	1.159*	1.432**	1.571**	1.106*	1.159*
RW	2.551***	1.154*	1.330*	1.298*	1.665**	2.607***	1.441**	1.871**	1.122*	1.172*
AR(1)	2.332***	1.149*	1.245*	1.300*	1.413**	1.580**	1.492**	1.301*	1.158*	1.191*

Nota: Esta tabla presenta las regresiones fuera de muestra para Estados Unidos, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 10: Tablas fuera de muestra para los mejores resultados de México

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Auditor</i>	<i>Caridad</i>	<i>Corporación</i>	<i>Desempleado</i>	<i>Económico</i>	<i>Empresario</i>	<i>Heredar</i>	<i>Tarifa</i>	<i>Trampa</i>	<i>Ventaja</i>
	50/50									
DRW	1.877**	1.271*	1.663**	1.179*	2.028**	1.295*	1.047*	2.131***	3.150***	1.269*
RW	1.853**	1.266*	1.707**	1.174*	1.579**	1.317*	1.052*	2.112***	3.215***	1.275*
AR(1)	1.030*	1.179*	1.734**	1.187*	1.255*	1.412**	1.055*	2.070***	2.683***	1.146*
	70/30									
DRW	1.066*	1.356*	2.069***	1.171*	0.631	1.066*	1.230*	1.924**	1.654**	1.436**
RW	0.990*	1.353*	2.119***	1.162*	0.519	1.078*	1.225*	1.910**	1.724**	1.448**
AR(1)	-0.664	1.315*	2.117***	1.169*	0.993*	1.158*	1.246*	1.890**	1.886**	1.499**
	30/70									
DRW	1.739**	1.278*	1.106*	1.154*	2.180***	1.111*	1.075*	2.248***	3.795***	1.491**
RW	1.766**	1.261*	0.995*	1.154*	1.997**	1.103*	1.075*	2.243***	4.080***	1.541**
AR(1)	0.818	1.225*	0.967	1.159*	1.343*	1.153*	1.018*	2.133***	3.552***	1.119*

Nota: Esta tabla presenta las regresiones fuera de muestra para México, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 11: Tablas fuera de muestra para los mejores resultados de Nueva Zelanda

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Banco</i>	<i>Benefactor</i>	<i>Capital</i>	<i>Comprobar</i>	<i>Consumir</i>	<i>Cooperativa</i>	<i>Gran éxito</i>	<i>Pago</i>	<i>Presupuestario</i>	<i>Ruina</i>
DRW	50/50									
	1.019*	1.736**	1.205*	1.074*	1.293*	0.873	1.026*	1.063*	1.403**	0.773
	RW	0.992*	1.295*	1.028*	1.015*	1.205*	0.849	1.015*	1.052*	1.333**
AR(1)	1.057*	1.381**	1.415**	1.008*	1.372**	0.849	1.026*	1.073*	1.495**	0.709
DRW	70/30									
	1.000*	2.85***	1.279*	1.211*	1.162*	1.148*	1.028*	1.077*	1.339*	2.039***
	RW	1.021*	3.01***	1.323*	1.235*	1.187*	1.151*	1.036*	1.084*	1.392**
AR(1)	1.019*	1.876**	1.594**	1.233*	0.953*	1.157*	1.037*	1.086*	1.508**	2.48***
DRW	30/70									
	0.812	1.716**	0.680	1.059*	0.984*	1.110*	1.008*	1.026*	1.631**	1.946**
	RW	0.661	0.222	-0.148	0.883	0.452	1.033*	0.956	0.969*	1.407**
AR(1)	0.778	1.296*	1.214*	0.985*	1.002*	1.106*	0.997*	1.020*	1.860**	1.770**

Nota: Esta tabla presenta las regresiones fuera de muestra para Nueva Zelanda, las palabras están basadas en el Diccionario de Harvard IV-4. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Tabla 12: Tablas fuera de muestra para los mejores resultados de Reino Unido

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	<i>Ambición</i>	<i>Asociación</i>	<i>Caritativo</i>	<i>Caro</i>	<i>Contrato</i>	<i>Cooperativa</i>	<i>Corporación</i>	<i>Gran éxito</i>	<i>Ingreso</i>	<i>Pago</i>
DRW	50/50									
	1.048*	1.895**	1.256*	1.663**	2.043**	1.861**	0.988*	0.995*	2.251***	1.408**
RW	1.056*	1.968**	1.261*	1.791**	2.059***	1.838**	0.998*	0.997*	2.323***	1.515**
AR(1)	1.018*	1.854**	1.211*	1.331*	1.563**	5.032***	1.413**	1.011*	3.054***	1.001*
DRW	70/30									
	1.007*	1.767**	1.073*	1.406**	1.642**	1.646**	2.147***	1.063*	2.06***	2.071***
RW	1.012*	1.854**	1.079*	1.408**	1.621**	1.652**	2.101***	1.068*	2.066***	2.19***
AR(1)	1.042*	3.150***	1.055*	1.002*	1.217*	4.088***	2.109***	1.087*	2.409***	1.741**
DRW	30/70									
	1.055*	2.302***	1.217*	1.707**	2.376***	2.27***	1.518**	1.011*	2.15***	1.973**
RW	1.065*	2.386***	1.246*	1.892**	2.409***	2.278***	1.539**	1.019*	2.302***	2.063***
AR(1)	0.988*	2.181***	1.261*	1.096*	1.979**	3.820***	1.905**	0.981*	3.008***	1.54**

Nota: Esta tabla presenta las regresiones fuera de muestra para Reino Unido, las palabras están basadas en el Diccionario de Harvard IV-4.. Significancia: ENC-T recursiva P/R≈1 0.955< * 1.331< ** 2.052< *** P/R≈2 0.939< * 1.360< ** 2.033< *** P/R≈0.4 0.968< * 1.350< ** 2.054< ***.

Fuente: Elaboración propia.

Las tablas 1-3 compuestas por las economías OCDE de Australia, Canadá y Chile, muestran las 10 mejores palabras que logran vencer a los benchmark, en el caso de los dos primeros países los resultados son excepcionales, las palabras logran ser significativas en las 3 divisiones de ventanas. Sin embargo, en Chile, se presentan sólo siete palabras con significancia en las 3 divisiones de ventanas. No obstante, las tres palabras restantes derrotan al menos un benchmark clásico dentro de alguna división de ventana.

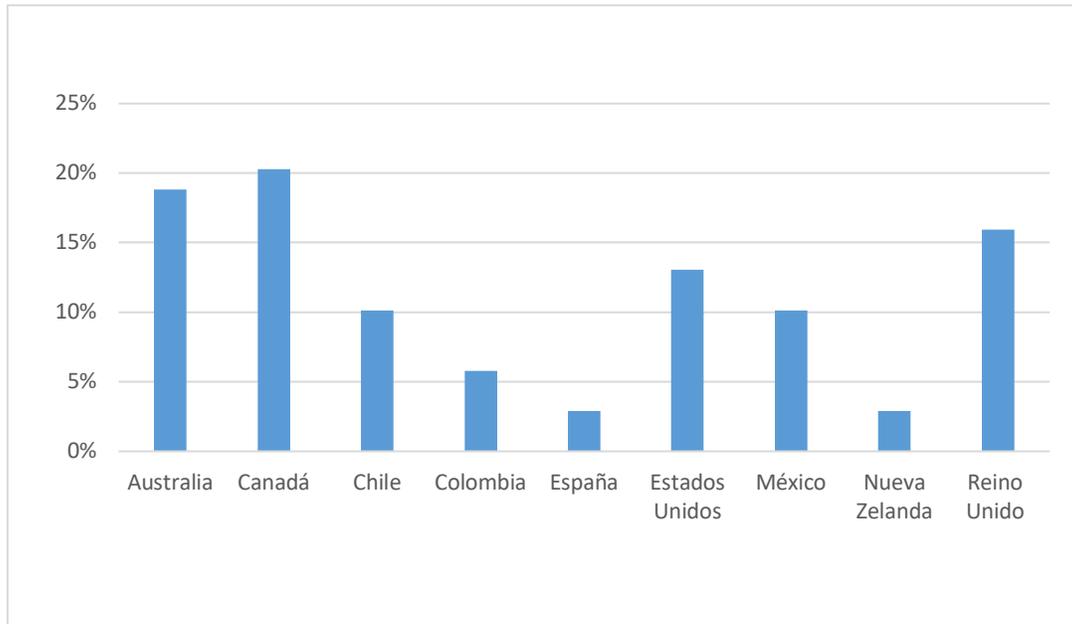
Las tablas 3-6 formadas por las economías OCDE de Colombia, España y Estados Unidos presentan una selección de las diez mejores palabras que se obtuvieron. Para Colombia y España los resultados fueron un total de seis palabras significativas entre los dos países para las 3 divisiones de ventanas. Por otra parte, Estados Unidos obtiene 9 de 10 palabras significativas. Cabe destacar que la selección de estas palabras es bastante estricta, ya que, filtra por tener significancia contra todos los benchmark y para cada ventana de división.

Las tablas 6-9 compuestas por las economías OCDE de México, Nueva Zelanda y Reino Unido presentan mejores resultados que las tablas mencionadas anteriormente ya que entre los 3 países logran un 66% de palabras significativas que vencen a los benchmark. Asimismo, hay un bajo reporte por parte de Nueva Zelanda, con tan sólo dos palabras que cumplen con ser

significativas en las 3 divisiones de ventanas para los 3 benchmark, no obstante, al igual que países anteriores, obtiene significancia frente algunos benchmark en distintas divisiones de ventanas.

En las tablas 4-12, se muestra un fuerte rechazo a la hipótesis nula para ambos paseos aleatorios (RW y DRW) y el AR (1), donde el 77% de las palabras seleccionadas derrotaron a los benchmark en las 3 divisiones de ventanas, y el 33% restante vence por lo menos a un benchmark en alguna de las 3 distintas divisiones de ventanas.

Figura 4: Palabras significativas para análisis fuera de muestra

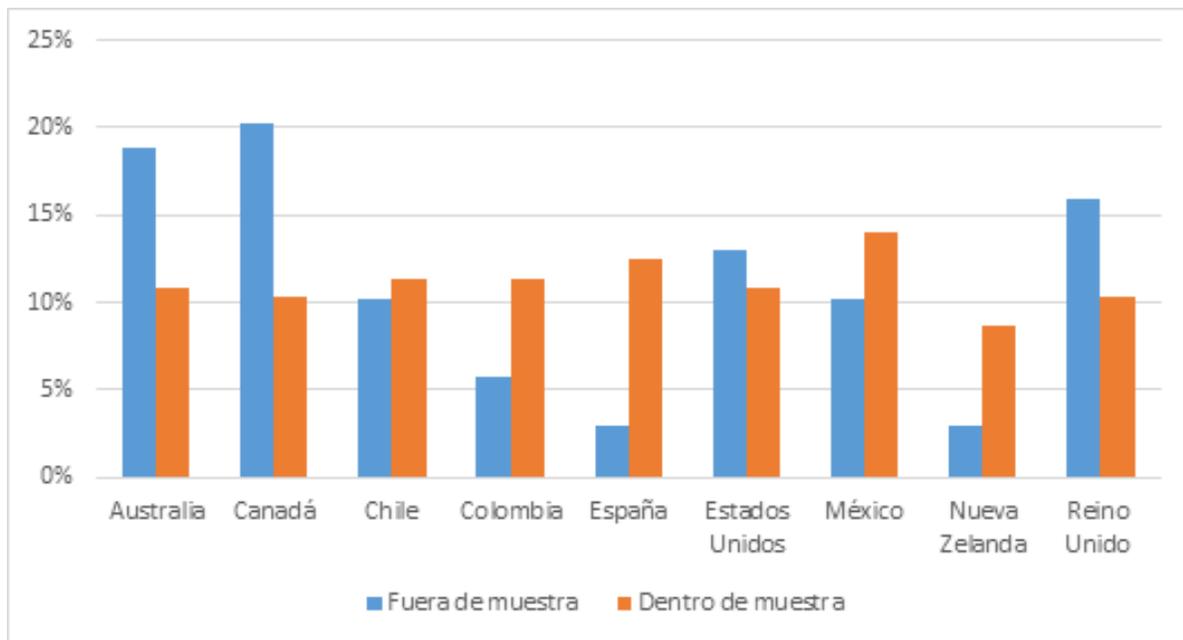


Nota: Este gráfico representa las palabras significativas en las 3 divisiones de datos de cada país para la prueba ENC-T en la ventana recursiva.

Fuente: Elaboración propia.

Sin embargo, esta es una representación de los resultados más interesantes, pero considerando todas las palabras analizadas, cabe destacar que los resultados no son tan potentes ya que sólo un 5.5% da significativas.

Figura 5: Resultados para análisis dentro y fuera de muestra



Nota: Este gráfico representa la comparación de resultados dentro de muestra y fuera de muestra de las palabras significativas de los 9 países de las economías OCDE.

Fuente: Elaboración propia.

En vías generales los resultados están en línea con lo que ha sido probado en la literatura y por lo general, los resultados dentro de muestra tienden a tener más poder con un 20.5%, pero a la vez, rechazan falsamente la hipótesis nula que las evaluaciones fuera de muestra con un resultado de 5.5%.

6. Conclusión

El propósito de esta investigación es evaluar la capacidad predictiva que tiene la herramienta Google Trends sobre los índices accionarios de las economías que componen la OCDE con primera lengua inglés y español. La motivación económica es que las frecuencias de búsqueda en Google Trends pueden funcionar como un proxy del sentimiento de mercado y de las expectativas que tienen los agentes sobre el futuro de la economía y por tanto, podría tener una capacidad predictiva sobre los índices accionarios. No obstante, la cantidad de palabras que logran vencer a los benchmark es muy baja, respecto al total de 139 palabras de cada país, estos resultados confirman lo complejo que es predecir índices accionarios, tipos de cambio y otros índices financieros, demostrando similitudes con la extensa literatura de forecasting.

Se descargó el Diccionario de Harvard IV-4, obteniendo un total de 150 palabras en inglés de tipo económico, de las cuales se escogieron 139 que fueron traducidas.

Se realizó análisis tanto dentro de muestra como fuera de muestra, donde se utilizaron los t-stadistic para evaluar el poder de predicción de las frecuencias de búsqueda. En general, hay palabras que tienen fuerte capacidad predictiva, sin embargo, esas palabras son un porcentaje muy menor del diccionario total disponible de Harvard.

Asimismo, una posible limitación de esta investigación es que el Diccionario de Harvard IV-4 con connotaciones positivas y negativas es un diccionario hecho en inglés y se utilizaron economías de la OCDE que tienen como primera lengua español e inglés y por tanto, se realizó la traducción de estos términos y en muchos casos, la traducción puede perder algo en la connotación, por lo que los resultados en español deben ser vistos con precaución.

Una vía interesante para la investigación futura podría ser extender el análisis para considerar otros países de la OCDE, o también sería interesante explorar si esta relación predictiva arroja mejores resultados en horizontes más largos, o también podría ser realizar un diccionario propio.

Bibliografía

1. Bulut, L. (2018). Google Trends and the forecasting performance of exchange rate models. *Journal of Forecasting*, 37(3), 303-315.
2. Dimpfl, T., & Jank, S. (2016). Can internet search queries help to predict stock market volatility?. *European Financial Management*, 22(2), 171-192.
3. Rossi, B., & Sekhposyan, T. (2011). Understanding models' forecasting performance. *Journal of Econometrics*, 164(1), 158-172.
4. Matsuki, T., & Chang, M. J. (2016). Out-of-Sample Exchange Rate Forecasting and Macroeconomic Fundamentals: The Case of Japan. *Australian Economic Papers*, 55(4), 409-433.
5. Meese, R.A., & Rogoff, K. (1983). Empirical exchange rate models of the seventies. *Journal of International Economics*, 14(1-2), 3-24.
6. Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic record*, 88, 2-9.
7. Adachi, Y., Masuda, M., & Takeda, F. (2017). Google search intensity and its relationship to the returns and liquidity of Japanese startup stocks. *Pacific-Basin Finance Journal*, 46, 243-257.
8. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012-1014.
9. Mark, N. C. (1995). Exchange rates and fundamentals: Evidence on long-horizon predictability. *The American Economic Review*, 201-218.
10. Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. *Financial markets and portfolio management*, 25(3), 239.
11. Chojnowski, M., & Dybka, P. (2017). Is exchange rate moody? forecasting exchange rate with google trends data. *Econometric Research in Finance*, 2(1), 1-21.
12. Masuda, M., & Takeda, F. (2019). Application of Google Trends Data in Exchange Rate Prediction.
13. Onorante, L., & Koop, G. (2016, June). Macroeconomic nowcasting using Google probabilities. In CARMA 2016-1st International Conference on Advanced Research Methods and Analytics.
14. Rossi, B. (2005). Testing long-horizon predictive ability with high persistence, and the Meese-Rogoff puzzle. *International Economic Review*, 46(1), 61-92.
15. Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction.
16. Clark, T. E., & West, K. D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of econometrics*, 135(1-2), 155-186.
17. West, K. D. (2006). Forecast evaluation. *Handbook of economic forecasting*, 1, 99-134.
18. Meese, R., & Rogoff, K. (1983). The out-of-sample failure of empirical exchange rate models: sampling error or misspecification?. In *Exchange rates and international macroeconomics* (pp. 67-112). University of Chicago Press.
19. Timmermann, A. (2008). Elusive return predictability. *International Journal of Forecasting*, 24(1), 1-18.
20. Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1), 291-311.

21. West, R. L. (1996). An application of prefrontal cortex function theory to cognitive aging. *Psychological bulletin*, 120(2), 272.
22. Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of econometrics*, 105(1), 85-110.

Apéndice suplementario

Tabla 13: Tabla de palabras eliminadas del estudio

Alto precio	Costo	Efectivo
Bondad	Defecto	Emprendedor
Café	Deudor	Vínculo
Cooperativa	Económico	

Nota: Esta tabla contiene las palabras eliminadas del Diccionario de Harvard IV-4 ya que no aportan al estudio.

Fuente: Elaboración propia.