TRIBUNA DE ECONOMÍA

Todos los artículos publicados en esta sección son sometidos
a un proceso de evaluación externa anónima

Juan de Lucio Fernández*

ANÁLISIS DE LOS ARTÍCULOS PUBLICADOS EN ICE: TEMÁTICA Y FACILIDAD DE LECTURA

Las técnicas de procesamiento del lenguaje natural permiten aproximaciones rigurosas y replicables que facilitan el análisis del contenido publicado en las revistas académicas. En este documento se emplean para analizar la temática tratada y la legibilidad de los documentos publicados en Información Comercial Española (Boletín, Cuadernos y Revista) durante los últimos 20 años. Los resultados confirman objetivamente, mediante técnicas cuantitativas: la proximidad temática de las publicaciones a los intereses de la Secretaría de Estado de Comercio que las patrocina, sugiere un descenso a lo largo del tiempo de la facilidad de lectura de los artículos publicados, confirma la cercanía de contenidos entre Revista de Economía y Boletín Económico y muestra el sesgo académico y la mayor variedad temática de Cuadernos Económicos.

An analysis on the topics and the legibility of published articles in ICE

Natural language processing techniques allow to make rigorous and replicable analysis, easing the study of the content of the articles published in academic journals. In this article, we use such techniques to analyze the topics and the legibility of the articles published in Información Comercial Española journals (Boletín, Cuadernos and Revista) during the last twenty years. The results objectively confirm, throught the use of quantitative techniques, that the topics covered in the articles are closely linked to the topics of interest of the Secretariat of State for Trade, the institution that funds ICE journals. The analysis also suggests that the topics covered in Revista de Economía and Boletín Económico are very similar between them and shows the academic approach and the greater variety of topics found in Cuadernos Económicos.

Palabras clave: *Información Comercial Española, temática, legibilidad, procesamiento del lenguaje natural.*
Keywords: *Información Comercial Española, topics, legibility, natural language processing.*
JEL: *A1, F00, Z13.*

* Universidad de Alcalá.
El artículo se ha beneficiado de los comentarios de Ramón Ferrer-i-Cancho. Asimismo, el autor agradece la financiación recibida por la Comunidad de Madrid y la UAH (ref: EPU-INV/2020/006).
Contacto: Juan.deLucio@uah.es
Versión de marzo de 2021.
https://doi.org/10.32796/ice.2021.919.7177

1. Introducción

La Secretaría de Estado de Comercio del Ministerio de Industria, Comercio y Turismo publica tres revistas periódicas que configuran el grupo de revistas de *Información Comercial Española (ICE)*. Los dos objetivos principales de estas publicaciones son: contribuir a la difusión del conocimiento y participar en el debate en materia económica. El grupo de revistas intenta prestar especial atención a los temas de mayor relevancia y actualidad para la economía española. El análisis del contenido de las revistas a lo largo de los años debería permitir identificar aquellos temas que han alcanzado mayor relevancia y actualidad en la economía española. Este trabajo analiza el contenido de los trabajos publicados durante los últimos 20 años en función de los dos objetivos señalados, es decir, analiza los aspectos relacionados con la difusión (en concreto la facilidad de lectura de los documentos) y la temática abordada en las publicaciones (contenido del debate en materia económica).

El trabajo tiene un doble propósito: en primer lugar, se analiza el contenido temático, y en segundo lugar, se estudia el carácter divulgativo de los artículos publicados. Para ello se utilizan técnicas de procesamiento del lenguaje natural (PLN).

Las tres revistas que promueve ICE[1] son:

1) El **Boletín Económico** de Información Comercial Española (en este trabajo, Boletín) se edita desde 1947, actualmente con carácter mensual. Según se indica en su página web tiene un carácter divulgativo de la acción de «la Administración en el ámbito económico y comercial».

2) **Cuadernos Económicos** de ICE (publicación que denominaremos Cuadernos en este documento) se edita desde 1977 y se publica actualmente con periodicidad semestral. La revista está orientada a la investigación y elabora números monográficos. Es la revista más científica entre la publicadas por ICE, aunque presta especial interés a temas de actualidad o innovadores.

3) Información Comercial Española, **Revista de Economía** (Revista, a partir de aquí), es una publicación de referencia para el seguimiento de la economía española con especial interés en su dimensión exterior. Es una publicación monográfica, de periodicidad bimestral y presencia editorial desde 1898. La revista, tal y como se indica en su página web, pretende contribuir al debate económico acercando el rigor académico a las políticas públicas.

Los contenidos del Boletín y la Revista están disponibles en la web (http://www.revistasice.com/) desde mediados de 1999. Para Cuadernos, la web publica contenidos anteriores. Por homogeneidad utilizaremos los últimos 20 años de información. La página web de ICE facilita el acceso a los artículos de cada uno de los números en formato PDF. Este acceso ha permitido descargar los artículos en este formato para su análisis sistemático con técnicas de PLN.

El análisis de texto permite la identificación de temáticas para cada uno de los artículos, también se realiza una identificación de palabras clave y conceptos que pueden ser seguidos a lo largo del tiempo. Adicionalmente, las técnicas de análisis de texto permiten analizar la facilidad de lectura de los textos que se publican. En conjunto, con todo ello se obtiene una visión general, a través del análisis sistemático y cuantitativo, de una colección de publicaciones de carácter económico y comercial de referencia en España. El trabajo proporciona una primera aproximación a los temas que han ocupado el tiempo de los especialistas que finalmente publican su trabajo en ICE. Igualmente, el análisis simultáneo de publicaciones hermanas permite una comparación entre las mismas.

En total se han descargado de la página, utilizando técnicas de «raspado web» (*web scraping*), 3.866 archivos que reflejan el contenido de 602 números. Aproximadamente 53.000 páginas y 12 millones de palabras publicadas en los últimos 20 años.

[1] ICE publica también un número especial del Boletín Económico dedicado al sector exterior de cada año, que no se analiza en este documento.

El resto del trabajo se articula de la siguiente manera. El segundo apartado presenta en mayor detalle la base de datos. En el apartado tres se presenta la metodología y los resultados que se derivan de la misma. Esta sección contiene los resultados principales del trabajo en dos partes; la primera sobre temática y conceptos clave, y la segunda sobre facilidad de lectura. El documento se cierra con una sección de conclusiones y un Anexo con información complementaria a la presentada en el texto principal.

2. La base de datos

Para construir la base de datos ha sido necesario recurrir a técnicas de «raspado web» de manera que para cada una de las revistas y cada uno de los números se identifican los enlaces a los artículos publicados para posteriormente descargar los documentos. Las técnicas de extracción de información de la web han sido utilizadas anteriormente para analizar el contenido de las mejores revistas de economía (Card & DellaVigna, 2013 o Anauati *et al.*, 2016).

El Boletín incorpora varias secciones, algunas de las mismas tienen un carácter más institucional y en algunos casos se dedican a contenidos propios elaborados por el Ministerio. El trabajo pretende capturar aquellos contenidos complementarios a las motivaciones institucionales. Buscando homogeneidad con el contenido del resto de publicaciones analizadas, en el Boletín nos centramos en el apartado de colaboraciones, habitualmente externas al Ministerio. Quedan fuera del trabajo los asuntos recurrentes y aquellos elaborados por la entidad que edita las publicaciones. Por este motivo, este análisis se centra en la sección de «colaboraciones» cuyo proceso de elaboración de contenidos es más homogéneo con el que cabría esperar de las publicaciones de la Revista o de Cuadernos; generalmente contribuciones de especialistas externos al Ministerio con estructura de artículo científico preocupado por la difusión.

Cada uno de los artículos se ha tratado de la siguiente manera. En primer lugar, el fichero en formato PDF de cada documento se ha pasado a texto, desechando las referencias bibliográficas y los anexos[2]. En segundo lugar, si el artículo original estaba en inglés se ha traducido al español para hacer todo el *corpus* homogéneo en un único idioma[3]. En tercer lugar, se han armonizado los textos, en concreto, por ejemplo, se han unido aquellas palabras separadas por guion de cambio de línea, se han transformado todas las letras a minúsculas y se han realizado otras modificaciones para homogeneizar el formato de los textos. En cuarto lugar, el texto se divide en párrafos, frases, palabras, sílabas y letras. En quinto lugar, se tratan las palabras de la siguiente manera, se eliminan las que tienen menos de tres caracteres, los números y las «palabras vacías»[4]. Los textos en tercera persona se cambian a primera persona y los verbos en tiempo pasado y futuro se cambian a presente (p. ej., «él comió» se sustituye por «yo comer»). Finalmente, los términos se reducen a su forma raíz. Estos pasos nos proporcionan unos textos homogéneos para los distintos artículos sobre los que aplicaremos una serie de técnicas de tratamiento del lenguaje natural.

En términos generales este proceso permite crear una base de datos en la que cada uno de los artículos tiene un registro con sus características más relevantes para el análisis. Un resumen de la base de datos se presenta en la Tabla 1. En total se han analizado 3.866 artículos, el 56 % de los documentos han sido publicados en la Revista, el 35 % de los artículos en el Boletín y solo un 9 % en Cuadernos. Dado que para Cuadernos se dispone de un número sustancialmente menor de documentos, en algunas secciones del análisis se ha optado por una presentación independiente de las otras dos publicaciones[5]. En algunos casos se

[2] Se desechan también encabezados y pies de página, así como imágenes, esquemas, cuadros y similares.

[3] Se ha utilizado Googletrans 2.4.0.

[4] Las palabras vacías, por ejemplo «y» o «la» carecen de información específica sobre el contenido del artículo. Las palabras vacías se eliminan con la librería NLTK.

[5] Para algunos ejercicios de Cuadernos se han eliminado los números 683-685 por problemas de formato en PDF.

TABLA 1

CARACTERÍSTICAS BÁSICAS DE LOS TEXTOS ANALIZADOS DE LAS REVISTAS DE INFORMACIÓN COMERCIAL ESPAÑOLA

	Revista	Boletín	Cuadernos	**Total**
Artículos	2.161	1.357	348	**3.866**
Números	135	434	33	**602**
Páginas	28.783	16.966	7.695	**53.444**
Palabras	6.531.977	3.881.838	1.525.377	**11.939.192**
Frases	1.381.609	768.269	290.953	**2.440.831**
Párrafos	163.757	138.838	32.020	**334.615**

FUENTE: Elaboración propia.

desplazan los resultados del análisis de Cuadernos, referidos a un menor número de documentos, al Anexo.

Los artículos tienen 12 millones de palabras lo que supone algo más de 3.000 palabras por artículo sin tener en cuenta las palabras vacías. Los artículos de la Revista son un poco más largos que las colaboraciones de Boletín, aunque los publicados en Cuadernos son los más extensos de todos. Durante el periodo estudiado, aproximadamente dos décadas, se han publicado más de 53.000 páginas (54 % en la Revista, 32 % en el Boletín y el 14 % en Cuadernos).

A partir de los artículos descargados de internet se lleva a cabo el análisis de los textos en dos dimensiones: contenido temático y capacidad divulgativa (facilidad de lectura). Este análisis se realiza de manera independiente para cada uno de los artículos. Los resultados obtenidos para los artículos de manera aislada se agregan para cada una de las publicaciones y pueden explotarse, igualmente, en función del año de publicación. Estos procesos de análisis son independientes entre ellos y se presentan, junto con sus resultados, en la siguiente sección. Con ello podremos ver los temas de interés, la coincidencia entre publicaciones y la evolución de la facilidad de lectura de los textos publicados.

3. Resultados

En este apartado se detallan las metodologías y los resultados obtenidos para dos aproximaciones: contenidos (temática) y legibilidad (divulgación).

Contenidos

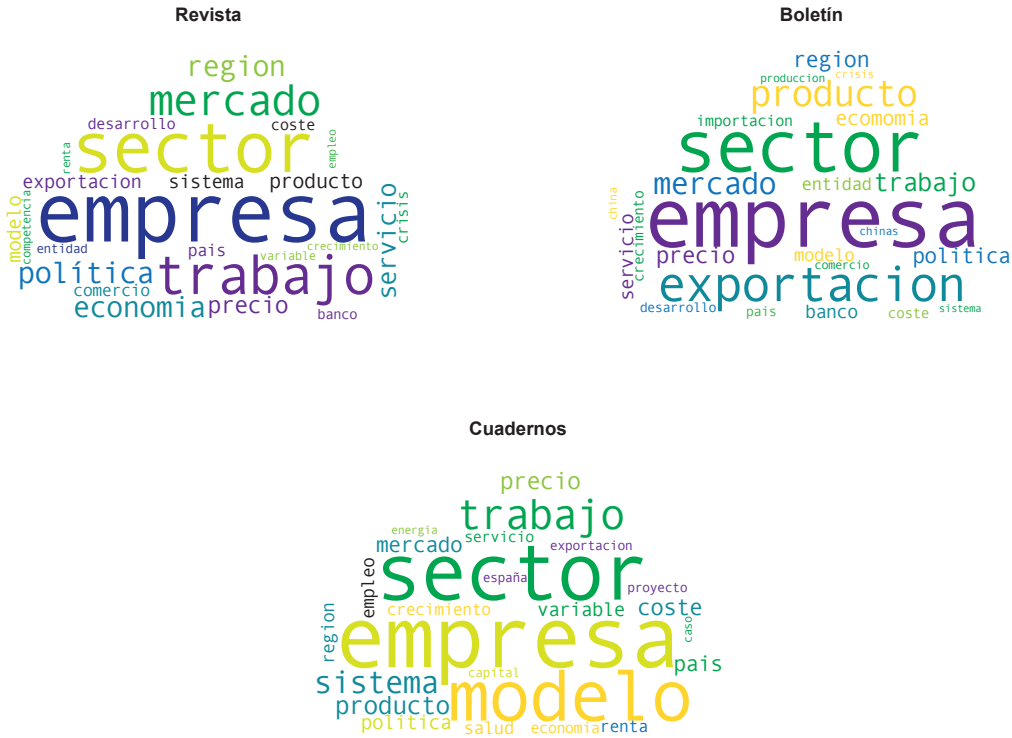
Para analizar los contenidos se proponen tres aproximaciones: una primera identifica conceptos y palabras clave; la segunda determina los temas abordados; finalmente se visualiza la evolución de la presencia en los textos de algunos conceptos. En todos los casos se realizan los ejercicios utilizando técnicas de PLN.

Palabras clave

La primera propuesta de análisis de los textos es extraer las palabras clave de cada uno de los documentos. Las palabras clave no corresponden con las extraídas por los autores, que están sujetas a criterios subjetivos de cada uno de los autores, sino que son identificadas mediante la utilización de un algoritmo matemático elaborado por Mihalcea y Tarau (2004) y mejorado por Barrios *et al.* (2016). El algoritmo utiliza

FIGURA 1

NUBE DE PALABRAS CLAVE EN LAS PUBLICACIONES DE ICE



FUENTE: Elaboración propia.

un mecanismo de clasificación basado en grafos en el que las palabras son los vértices y las relaciones entre ellos ilustran las relaciones entre los distintos conceptos. Esta metodología proporciona un sistema automático homogéneo para todos los textos[6] y devuelve la importancia relativa de cada concepto en el artículo.

Los conceptos clave extraídos mediante este procedimiento para cada artículo se agrupan por publicación y se representan en una nube de palabras en las que el tamaño de cada palabra se corresponde con la frecuencia que aparece entre los conceptos clave de la publicación. En la Figura 1 se pueden observar estas nubes de palabras. En la parte superior tenemos las correspondientes a la Revista y al Boletín, como veremos bastantes similares entre sí: este resultado es especialmente significativo si consideramos los procesos paralelos que se han llevado para el análisis.

En las tres revistas observamos que su foco se encuentra en temas relacionados principalmente con las siguientes palabras: «empresa», «mercado», «sector», «producto» y «exportación». Aunque también son frecuentes las palabras derivadas de «país» y «región». Estos términos se aproximan considerablemente a los intereses de la Secretaría de Estado de Comercio (SEC) que promueve las publicaciones. A modo ilustrativo,

[6] Se ha utilizado Gensim (Řehůřek & Sojka, 2010). Esta rutina funciona adecuadamente en castellano.

podemos referirnos a lo que establece el Real Decreto 998/2018 en su artículo 2 sobre las competencias de la SEC «relativas a la definición, desarrollo y ejecución de la política comercial del Estado, en lo que se refiere al comercio exterior e interior, incluido el intracomunitario, así como a la estrategia competitiva de la política de internacionalización, las inversiones exteriores y las transacciones exteriores, y a las actividades de promoción e internacionalización de las empresas españolas».

Esta línea editorial tiene matices entre las distintas publicaciones que en algunos casos tienen que ver con la definición de las actividades de la SEC que realiza el Real Decreto que se acaba de mencionar. Por ejemplo, el Boletín contiene entre las palabras clave más frecuentes «comercio» e «importación» (esta última palabra aparece de manera exclusiva en esta publicación). Además, el Boletín es la publicación que pone más foco en el sector financiero; en su nube de puntos aparecen «entidades» y «banco». Por su parte, en la Revista son más característicos las temáticas relacionadas con: «política», «competencia», «servicios» y «desarrollo», alguno de estos términos también presente en el Real Decreto. Igualmente, en la Revista han tenido mayor presencia temas asociados a los conceptos de «crisis» y «trabajo».

El carácter más académico de Cuadernos se refleja en una presencia comparativamente más elevada de palabras como «modelo», «variable» y «sistema» poniendo de manifiesto su aproximación más teórica. Cuadernos también presenta una temática más amplia que el resto de las publicaciones como pone de manifiesto que entre los términos clave aparezcan palabras como «salud», «energía», «crecimiento», «coste» y varios términos relacionados con el mercado de trabajo. Cuadernos es, por lo tanto, la publicación con mayor variedad temática de las tres analizadas.

En conjunto, los resultados de la Figura 1 ponen de manifiesto una línea editorial conjunta de las publicaciones ICE pero con diferencias entre ellas, entre las que podemos destacar:

- Hay una mayor concentración de los asuntos comerciales en el Boletín.
- Cuadernos seguido de Revista tienen un carácter más técnico.
- Cuadernos tiene una temática más variada, el Boletín tiene una mayor especialización en términos financieros y la Revista pone un interés algo diferencial en asuntos de política económica.

Para contrastar estas impresiones preliminares, en la siguiente sección se realiza un análisis adicional que pone su foco en la identificación de las temáticas de las tres publicaciones.

Temática

El segundo proceso de análisis del contenido se centra en extraer el tema de cada uno de los artículos. Utilizando estos términos se aplica la metodología denominada Asignación Latente de Dirichlet (LDA, por las siglas en inglés de Latent Dirichlet Allocation) propuesto por Blei *et al.* (2003).

En grandes líneas el procedimiento de LDA es el siguiente: a cada palabra del documento se le asigna una probabilidad en función de las veces que aparece. De acuerdo con estas probabilidades podemos considerar que las palabras que aparecen en unos documentos específicos, pero no son muy frecuentes en el *corpus* de documentos, nos estarían indicando temas en común en el subconjunto de documentos que contienen estos términos[7]. Con la distribución de palabras por tema podemos extraer en qué medida cada uno de los documentos aborda cada tema. El método, en definitiva, descubre los patrones estructurales de manera endógena sin que sea necesario la predefinición de palabras para cada tema. En este proceso el investigador debe determinar el número de temas que considera adecuado para un conjunto de documentos. Como primera aproximación y por simplicidad para

[7] Para llevar a cabo esta tarea se ha utilizado spaCy (*software* para el procesamiento avanzado de lenguaje natural en Python y Cython).

TABLA 2

TEMAS PRINCIPALES DE LA REVISTA Y EL BOLETÍN

Temática	Revista	Boletín
Comercio	País, exportación, producto, sector, millón, importación, comerciar, España, euro, europeo.	Empresa, país, mercado, sector, servicio, resultar, inversión, producto, comerciar, España.
Financiera	País, mercado, financiero, económico, inversión, banco, economía, crecimiento, político, sector.	Mercado, país, financiero, político, sistema, económico, tipo, desarrollar, banco, europeo.
Sociopolítica y empresarial	Empresa, desarrollar, país, mercado, económico, servicio, información, producto, actividad, social.	Económico, político, país, desarrollar, social, europeo, economía, público, España, año.

FUENTE: Elaboración propia.

poder presentar los resultados en un espacio reducido y demostrar la similitud temática entre publicaciones se ha considerado que tres temas podrían ser suficientes para estos objetivos.

Esta metodología es la utilizada por Azqueta-Gavaldon *et al.* (2020) para la construcción de indicadores de incertidumbre por temáticas (fiscal, monetaria, geopolítica…). También utilizan LDA los autores Tobback *et al.* (2017) en su trabajo para analizar la percepción de los medios de comunicación sobre el tono de los comunicados del BCE con objeto de detectar los temas dominantes en los artículos de noticias. Hansen, McMahon y Prat (2017) utilizan LDA para identificar temas en los comunicados del Comité Federal de Mercado Abierto de la Reserva Federal.

Las palabras que devuelve la técnica LDA cuando se aplica de manera independiente sobre el conjunto de artículos de cada una de las tres publicaciones son las que figuran en la Tabla 2 y en la Tabla 3. Cabe destacar que, pese a que los análisis se realizan de manera independiente para el Boletín y la Revista, los temas que se extraen mediante LDA son, como anticipaba el análisis de palabras clave, muy similares (Tabla 2). Estas dos publicaciones tienen términos e interpretaciones compartidas, lo que pone de manifiesto una proximidad temática entre ambas publicaciones. Así, por ejemplo, los términos «banco» y «financiero» solo aparecen en el grupo de temas financieros, común a ambas publicaciones y no aparece en los términos que caracterizan las temáticas de Cuadernos. La metodología identifica temáticas diferentes para Cuadernos (Tabla 3).

Las combinaciones de palabras «comerciar», «producto» y «sector», aparecen de manera común en el Boletín y la Revista, siendo identificativos del tema que se ha dado en llamar «comercio». Finalmente, hay una tercera temática de temas sociopolíticos y empresariales.

La temática abordada por Cuadernos es bastante diferente a la de las dos publicaciones anteriores, tal y como se puede observar en la Tabla 3. En primer lugar, observamos una serie de artículos cuyo interés en «modelo», «mostrar», «dato», «variable» y «resultado» es diferencial; este conjunto de documentos tiene una aproximación más orientada a la obtención de resultados probablemente mediante la utilización de modelos y variables. En segundo lugar, hay una temática que podríamos caracterizar como empresarial que viene determinada por temas relacionados con la «empresa», «sector», «país», «mercado», «crecimiento» y «desarrollo»; este grupo de artículos estaría próximo a la temática que denominamos comercio y estaría más en línea con los artículos clasificados en la misma temática en las otras dos publicaciones de

TABLA 3

TEMAS PRINCIPALES DE CUADERNOS

Temática	Cuadernos
Académica	Variable, resultado, efecto, modelar, casar, individuo, mostrar, dato, utilizar, país.
Empresarial	Empresa, país, mercado, resultado, crecimiento, sector, coste, efecto, económico, desarrollar.
Política económica española	Político, competencia, mercado, servicio, país, salud, precio, público, resultado, España.

FUENTE: Elaboración propia.

ICE. Finalmente, un grupo de artículos aborda temas de carácter más «político», «competencia», «servicio», o «salud»; en este grupo caben artículos más variados con una sensibilidad por la política económica española.

De acuerdo con esta distribución temática se han analizado los 3.866 artículos para asignarlos a una de ellas. Este procedimiento nos permite aproximarnos al interés que por los distintos temas se ha manifestado implícitamente en las publicaciones ICE. La Figura 2 representa el porcentaje de artículos que, en cada uno de los años, se han publicado sobre cada uno de los tres temas principales, tanto en la Revista como en el Boletín: estas dos son las publicaciones que mantienen temáticas similares y un número elevado de artículos en cada uno de los años. En el Anexo, Figura 5, se presentan los resultados para Cuadernos, algo menos estables por el menor número de artículos.

En primer lugar, debemos hacer notar que tanto el año 1999 como el año 2020 son años incompletos. Estos dos años muestran una distribución más desigual de los temas, aunque no modifican el mensaje general. En segundo lugar, según ha pasado el tiempo hay una menor presencia de los temas relativos a comercio exterior tanto en el Boletín como en la Revista, la correlación Kendall del porcentaje de artículos clasificados en comercio con la variable tiempo es en ambos casos negativa y significativa (-72 % y -50 % en ambos casos con *p-valor* = 0,000)[8]. En la Revista, los temas empresariales son los que han mantenido una tendencia creciente durante el periodo analizado (correlación Kendall del 53 %, *p-valor* = 0,000).

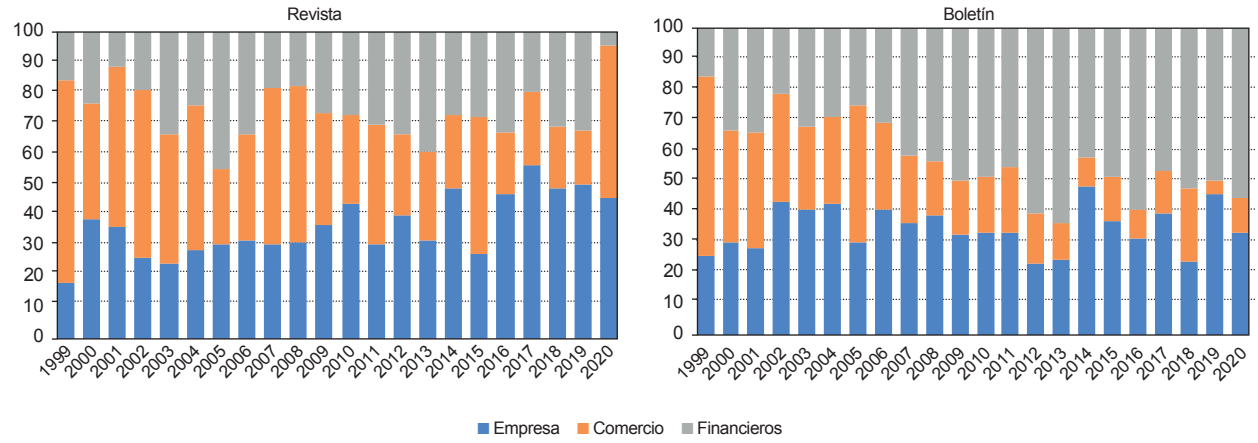
Aunque el número de observaciones es reducido para llevar a cabo test estadísticos entre periodos, en términos comparativos, durante las dos décadas analizadas, en la Revista los temas comerciales han sido la materia principal en el 38 % de los artículos, mientras que en el Boletín los artículos con temática comercial son el 27 %. En el Boletín parecen haber ganado presencia relativa asuntos financieros, el 52 % de los artículos se clasifican en esta temática después de la crisis financiera de 2008 frente al 31 % del periodo anterior. Por su parte, el gran colapso del comercio mundial no parece haber dado lugar a una mayor presencia relativa de temas comerciales entre los artículos publicados en el periodo posterior a la crisis.

La Figura 2 pone de manifiesto que, pese a que los temas de comercio tienen una presencia destacada en las publicaciones analizadas, no concentran la mayor parte de los artículos; las publicaciones proporcionan espacio a una amplia variedad de temáticas principalmente: financiera, empresarial o sociopolítica.

[8] Solo se proporcionan correlaciones y *p-valores* a título ilustrativo sin pretender exhaustividad en el análisis estadístico que es, en cualquier caso, exploratorio.

FIGURA 2

DISTRIBUCIÓN PORCENTUAL POR TEMÁTICAS DE LOS ARTÍCULOS PUBLICADOS EN EL BOLETÍN Y EN LA REVISTA DURANTE EL PERIODO 1999-2020



FUENTE: Elaboración propia.

Finalmente, a modo ilustrativo, se han considerado la aparición de una serie de términos específicos en las publicaciones ICE, Figura 3. El primero de los gráficos, esquina superior izquierda, se corresponde con palabras relacionadas con empresa (empresario, empresarial, etc.); estos términos tienen una de las mayores frecuencias tanto en la Revista como en el Boletín, en esta última publicación habría disminuido en la última década (pasa de una frecuencia cercana a 10, a una frecuencia próxima a 6 cada 1000 palabras). «País» es otra palabra que aparece con mucha frecuencia y ha permanecido estable o con una ligera tendencia descendente en el periodo analizado en el Boletín. Descienden la presencia (correlación Kendall negativa del -50 % con *p-valor* = 0,000 en el Boletín, y del -26 % con *p-valor* = 0,102 en la Revista) de los términos que comienzan por «export» (exportador, exportación, exportar, etc.) y «sector» (correlación Kendall negativa del -26 % con *p-valor* = 0,100 en la Revista, y del -40 % con *p-valor* = 0,010 en el Boletín). Aumenta (correlación positiva del 47 %, *p-valor* = 0,002) en el Boletín aquellos que empiezan por «internacion» (internacionalización, internacional, etc.) y se mantiene (sin correlación) en la Revista.

La última línea de gráficos en la Figura 3 presenta palabras con menor presencia en publicaciones ICE. El minigráfico correspondiente a «crisis» pone de manifiesto un cierto retraso en la publicación de artículos relacionados con la crisis en relación con el inicio de la misma; las publicaciones analizadas tardan un cierto tiempo en considerar temas de actualidad como lo fueron en su momento la crisis financiera y el colapso comercial mundial. El Boletín parece reaccionar con mayor agilidad a los temas de actualidad que la Revista. En la Revista se abordan con una cierta mayor intensidad temas relativos al mercado de trabajo (asunto de carácter estructural de la economía española). Finalmente, el grafico inferior derecho estaría relacionado con las palabras que empiezan por «politic» (política, político, etc.), cuya presencia ha sido más intensa (correlación Kendall del 24 % con *p-valor* = 0,129 en el Boletín y del 66 % con *p-valor* = 0,000 en la Revista) especialmente en el periodo más reciente.

FIGURA 3

PRESENCIA DE PALABRAS SELECCIONADAS A LO LARGO DEL TIEMPO. FRECUENCIA CADA MIL PALABRAS (LÍNEA DISCONTINUA) Y CURVA SUAVIZADA SOBRE EL DATO ORIGINAL MEDIANTE MEDIA MÓVIL (MM) DE CUATRO AÑOS (LÍNEA CONTINUA)

FUENTE: Elaboración propia.

Legibilidad

Los indicadores de legibilidad analizan la estructura léxica y lingüística del texto (párrafos, oraciones y palabras). Los artículos pueden ser más (menos) inteligibles en función de si sus frases y palabras son más cortas (largas) o más (menos) frecuentes. En este apartado utilizamos tres indicadores, dos habituales de la literatura para el análisis de la lengua española y se propone otro nuevo para analizar la perspicuidad.

En concreto, este trabajo se propone un nuevo índice de similitud entre el vocabulario utilizado en los

artículos analizados y el que se utiliza con más frecuencia en el idioma español según la información de la Real Academia de la Lengua Española. La institución española elabora el Corpus de Referencia del Español Actual, CREA[9], en el que se proporciona la distribución de probabilidades de las distintas palabras (tipos). El indicador propuesto calcula el porcentaje de apariciones de palabras (ocurrencias) que corresponden a palabras que están entre las 10.000 palabras más utilizadas en español (tipos), de acuerdo con la Ecuación [1]. El numerador de la ecuación suma la frecuencia absoluta, f_w, para el tipo, w, para aquellas palabras distintas, W_a, de en un artículo, a, que figuran entre las 10.000 palabras más frecuentes en español, $RAE_{10.000}$[10]. El denominador de la Ecuación [1] es el número total de palabras en el artículo. Cuanto mayor sea la proporción de apariciones de palabras dentro de la lista de las 10.000 más frecuentes de la RAE, mayor será la legibilidad. Por lo tanto, este indicador varía entre cero y uno y es mayor cuanto mayor es la simplicidad del vocabulario utilizado[11].

$$Frequentes_a = 100 \bullet \frac{\Sigma_w \varepsilon W \cap RAE_{10.000} f_w}{\Sigma_w \varepsilon W_a f_w} \qquad [1]$$

El indicador de palabras «frecuentes» en un año determinado será la media de $Frequentes_a$ para todos los artículos de ese año.

Adicionalmente, utilizamos dos indicadores conocidos en la literatura. En primer lugar, el indicador μ propuesto por Muñoz y Muñoz (2006) se calcula siguiendo la Ecuación [2]:

$$\mu_a = 100 \bullet \left(\frac{\Sigma_w \varepsilon W_a f_w}{\Sigma_w \varepsilon W_a f_w - 1} \right) \left(\frac{\overline{Le}_a}{\sigma^2_{Le_a}} \right) \qquad [2]$$

donde a, hace referencia al artículo concreto; $\sum w \, \varepsilon W_a f_w$, corresponde con el número de palabras del artículo; $\overline{Le}$, es el número medio de letras por palabra, σ^2_{Le}, hace referencia a la varianza del número de letras por palabra[12]. El indicador μ varía entre 0 y 100 para cada artículo, aunque en circunstancias excepcionales puede ser superior a 100. Niveles más elevados del indicador μ están asociados con niveles mayores de legibilidad. Niveles inferiores del indicador se corresponden con artículos más difíciles de leer.

Finalmente, se calcula el indicador de entropía propuesto por Katz y Bommarito (2014) y Shannon (1951). La entropía nos indica la energía dentro de un sistema, mayor entropía mayores niveles de incertidumbre y menor legibilidad. Debemos hacer notar que existe un límite inferior a partir del cual se reduce el potencial comunicativo de un texto (Bentz *et al.*, 2017; Ferrer-i-Cancho, 2018), en este sentido es una medida parcial del coste de la comunicación relacionada con la diversidad del vocabulario. El indicador para un artículo, Ecuación [3], se define como:

$$Entropía_a = -\Sigma_{w \varepsilon W} p_w \, log_2 \, p_w \qquad [3]$$

donde p_w, es la probabilidad de aparición de cada palabra (tipo). El valor final es la suma, para todas las palabras del texto, de la probabilidad de aparición de cada palabra multiplicada por su logaritmo en base 2.

El valor final correspondiente para cada año es la media correspondiente a los artículos publicados ese año. La Figura 4 presenta la evolución de la facilidad de lectura de los textos publicados. Se presenta el dato de cada año (línea discontinua) y la media móvil de cuatro años (línea continua). En relación con el Boletín se observa una menor legibilidad en la actualidad que en el año 2000. En relación con la Revista,

[9] https://www.rae.es/recursos/banco-de-datos/crea

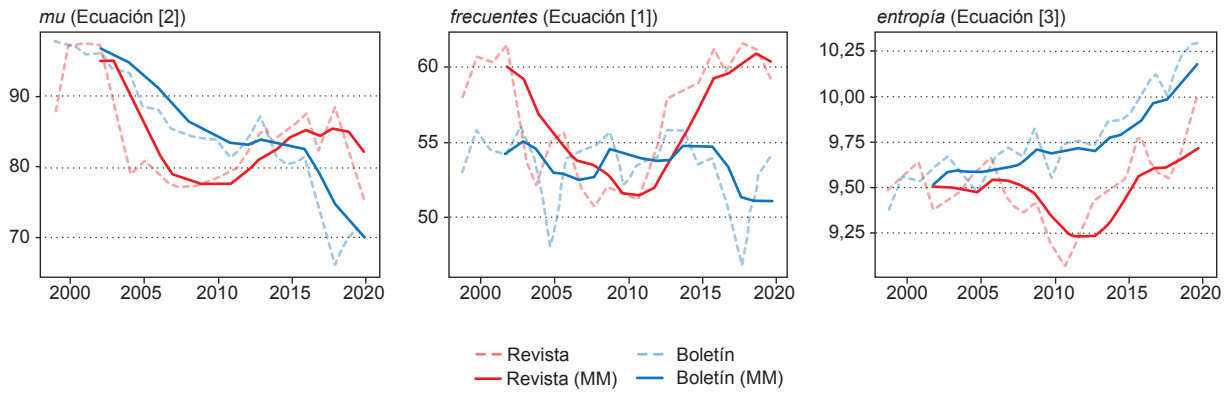
[10] Pese a que la selección del número de palabras frecuentes es arbitraria, existe evidencia que indica que el conjunto de palabras nucleares en un idioma se comportaría de manera diferente al resto de las palabras y la transición entre ambos conjuntos, de palabras frecuentes y poco frecuentes, se produciría antes de la palabra 10.000 (ver Ferrer-i-Cancho & Solé, 2001).

[11] El mismo indicador puede ser calculado utilizando las palabras únicas, es decir, el porcentaje de términos diferentes sobre el número total de términos, sin que los resultados varíen.

[12] Para textos con una cantidad de palabras como las consideradas en este documento, el primer paréntesis tiende a uno y, por lo tanto, la fórmula es la inversa del coeficiente de variación de Pearson, que indica la relación entre la desviación estándar de una muestra y su media, dividido todo ello entre la variación estándar.

FIGURA 4

EVOLUCIÓN DE LOS INDICADORES DE LEGIBILIDAD. DATO ANUAL (LÍNEA DISCONTINUA) Y CURVA SUAVIZADA SOBRE EL DATO ORIGINAL MEDIANTE MEDIA MÓVIL (MM) DE CUATRO AÑOS (LÍNEA CONTINUA)



FUENTE: Elaboración propia.

en las líneas rojas, se observan dos periodos, durante la primera década hay un descenso de la facilidad de lectura que parece corregirse en la segunda década. Los resultados para Cuadernos, con un periodo y un número de artículos más reducido, se encuentran en la Figura 6 en el Anexo.

La evidencia facilitada es compatible con lo indicado por Plavén-Sigray *et al.* (2017) para un periodo más largo demuestran que la legibilidad de la ciencia ha disminuido durante el siglo pasado y en lo que llevamos de este. En cualquier caso, este ejercicio debe considerarse como primera evidencia pues el periodo analizado es solo de 20 años y la base solo recoge publicaciones ICE por lo que no puede extenderse a otras publicaciones de carácter académico.

4. Conclusiones

Este artículo utiliza técnicas cuantitativas de análisis de texto para examinar los últimos 20 años de publicaciones de ICE. En concreto, se analiza el Boletín, Cuadernos y la Revista, estas publicaciones son decanas y referencias en España en los ámbitos empresariales, políticos y académicos.

Las técnicas de PLN abren un campo de creciente interés en la academia y la empresa. El análisis de texto puede confirmar o desmentir de manera cuantitativa apreciaciones subjetivas. Las técnicas de PLN aplicadas a las publicaciones ICE ponen de manifiesto que la línea editorial se aproxima a los intereses de la Secretaría de Estado de Comercio que edita las publicaciones. En concreto, el Boletín y la Revista ponen especial énfasis en temas relacionados con el comercio, aunque también prestan espacio a asuntos sociopolíticos, empresariales y financieros. Por su parte, la temática de Cuadernos es más amplia y algo más próxima a la academia. En general, se observa una cierta dificultad para contribuir en tiempo real a los debates más pujantes de la economía española; la capacidad de las publicaciones ICE para captar temas de actualidad en tiempo real es aparentemente baja, aunque quizás algo superior en el Boletín. Las técnicas utilizadas permiten analizar la legibilidad de los textos publicados para contrastar, en línea con artículos previos, un paulatino descenso en la facilidad

de comprensión de los textos académicos. Los resultados preliminares apuntan que podría haberse producido un cierto descenso de la comprensibilidad en el Boletín (principalmente utilizando los indicadores; *μ* y *entropía*). Los datos para la Revista son menos concluyentes con dos periodos diferenciados, el primer y el segundo decenio, con deterioro y recuperación de la perspicuidad, respectivamente. El análisis puede ser ampliado utilizando otros indicadores y periodos.

Este análisis es preliminar y debe ser recibido con cautela. La creación de la base de datos mediante técnicas de rastreado web, la traducción de textos y la traslación de ficheros con formato PDF a textos, son procesos que incorporan un cierto grado de imprecisión. Sin embargo, la robustez de los instrumentos utilizados y la convergencia de los resultados desarrollados de manera individual, para cada una de las publicaciones, dan un grado de confiabilidad suficiente como para dar validez a las consideraciones más generales.

Las posibilidades que se abren para mejorar la investigación mediante la utilización de PLN son innumerables, proporcionando un campo fascinante para el desarrollo de propuestas e iniciativas hasta hace poco impensables. El análisis de texto es posible utilizarlo en múltiples dimensiones, por ejemplo, para identificar países, sectores o mercados en expansión, para reconocer temas emergentes en tratados internacionales, para analizar la gestión y percepción de los usuarios de políticas comerciales, entre otras posibilidades.

Referencias bibliográficas

Anauati, V., Galiani, S. & Gálvez, R. H. (2016). Quantifying the life cycle of scholarly articles across fields of economic research. *Economic Inquiry*, *54*(2), 1339-1355.

Azqueta-Gavaldon, A., Hirschbühl, D., Onorante, L. & Saiz, L. (2020). Economic policy uncertainty in the Euro area: An unsupervised machine learning approach. *European Central Bank*, Working Paper No. 2359.

Barrios, F., López, F., Argerich, L. & Wachenchauzer, R. (2016). Variations of the similarity function of textrank for automated summarization. *arXiv* preprint *arXiv*: 1602.03606.

Bentz, C., Alikaniotis, D., Cysouw, M. & Ferrer-i-Cancho, R. (2017). The entropy of words—Learnability and expressivity across more than 1000 languages. *Entropy*, *19*(6), 275.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(4-5), 993-1022.

Card, D. & DellaVigna, S. (2013). Nine facts about top journals in economics. *Journal of Economic Literature*, *51*(1), 144-161.

Ferrer-i-Cancho, R. (2018). Optimization Models of Natural Communication. *Journal of Quantitative Linguistics*, *25*(3), 207-237.

Ferrer-i-Cancho, R. & Solé, R. V. (2001). Two regimes in the frequency of words and the origins of complex Lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, *8*(3), 165-173.

Hansen, S., McMahon, M. & Prat, A. (2017). Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801-870.

Katz, D. M. & Bommarito, M. J. (2014). Measuring the complexity of the law: The United States code. *Artificial intelligence and law*, *22*(4), 337-374.

Mihalcea, R. & Tarau, P. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411).

Muñoz, M. & Muñoz, J. (2006). *Legibilidad Mμ*. Viña del Mar, Chile.

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C. & Thompson, W. H. (2017). The readability of scientific texts is decreasing over time. *Elife*, *6*, e27725.

Řehůřek, R. & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell system technical journal*, *30*(1), 50-64.

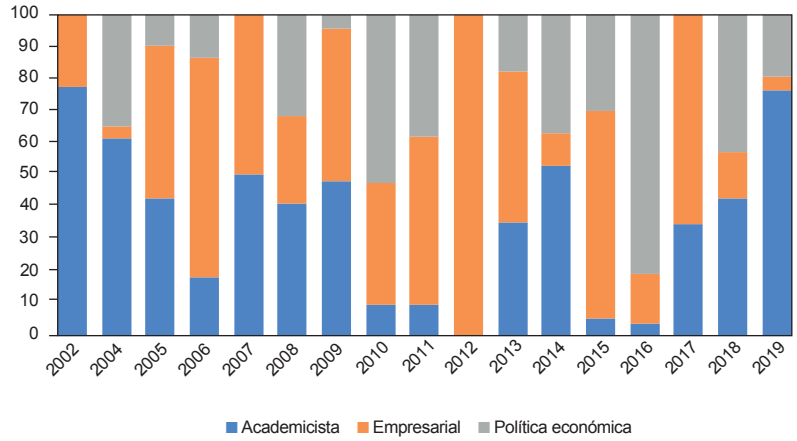
Tobback, E., Nardelli, S. & Martens, D. (2017). Between hawks and doves: measuring central bank communication. *European Central Bank*, Working Paper No. 2085.

ANEXO

Resultados adicionales

FIGURA 5

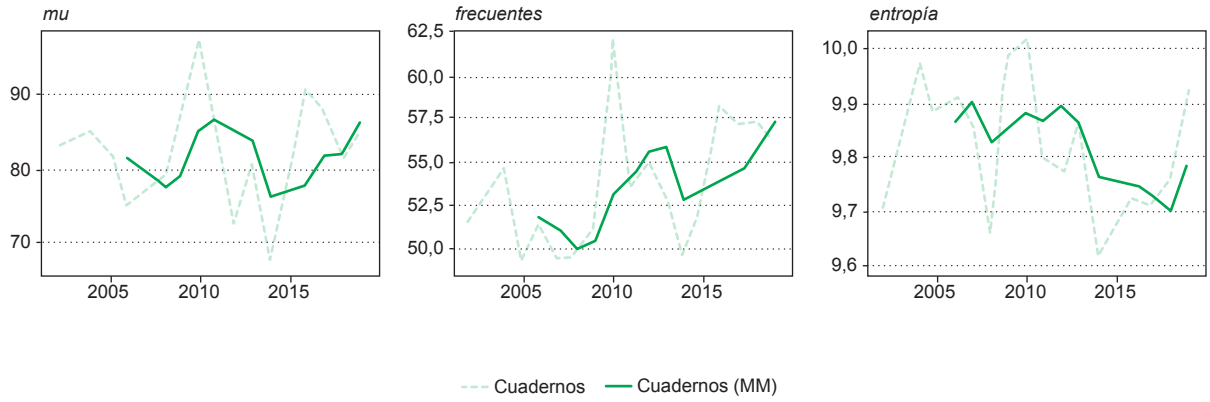
DISTRIBUCIÓN PORCENTUAL POR TEMÁTICAS DE LOS ARTÍCULOS PUBLICADOS EN CUADERNOS



FUENTE: Elaboración propia.

FIGURA 6

EVOLUCIÓN DE LOS INDICADORES DE LEGIBILIDAD EN CUADERNOS. AÑO BASE 2004 = 100. DATO ANUAL Y MEDIA MÓVIL DE CUATRO AÑOS (MM)



FUENTE: Elaboración propia.