ANÁLISIS DE SENTIMIENTO DEL *INFORME DE ESTABILIDAD FINANCIERA*

2020

BANCO DE ESPAÑA
Eurosistema

Documentos de Trabajo
N.º 2011

Ángel Iván Moreno Bernal y Carlos González Pedraz

ANÁLISIS DE SENTIMIENTO DEL *INFORME DE ESTABILIDAD FINANCIERA* (*)

Ángel Iván Moreno Bernal y Carlos González Pedraz

BANCO DE ESPAÑA

(*) Los autores agradecen a José Manuel Marqués, Ricardo Gimeno, Carlos Conesa, Jesús Ibáñez, Adrian van Rixtel y Carlos Pérez sus comentarios y sugerencias; y a Alicia Aguilar, José Manuel Carbó, María Teresa González, Roberto Pascual, Jara Quintanero y José Luis Romero, su ayuda en la definición y elaboración del diccionario.

Documentos de Trabajo. N.º 2011

2020

El objetivo de la serie de Documentos de Trabajo es la difusión de estudios originales de investigación en economía y finanzas, sujetos a un proceso de evaluación anónima. Con su publicación, el Banco de España pretende contribuir al análisis económico y al conocimiento de la economía española y de su entorno internacional.

Las opiniones y análisis que aparecen en la serie de Documentos de Trabajo son responsabilidad de los autores y, por tanto, no necesariamente coinciden con los del Banco de España o los del Eurosistema.

El Banco de España difunde sus informes más importantes y la mayoría de sus publicaciones a través de la red INTERNET, en la dirección http://www.bde.es.

ISSN: 1579-8666 (edición electrónica)

Resumen

En este artículo se muestra una aplicación de la minería de textos para extraer información de documentos financieros y usar esta información para crear índices de sentimiento. En particular, el análisis se centra en los diferentes números del *Informe de Estabilidad Financiera* (IEF) del Banco de España desde 2002 hasta 2019 en su versión en español, y en la reacción de la prensa a este Informe. Para calcular los índices, se ha creado, hasta donde conocemos, el primer diccionario en español de palabras con connotación positiva, negativa o neutra dentro del contexto de la estabilidad financiera. Se analiza la robustez de los índices aplicándolos a distintas secciones del Informe, y usando diversas variaciones del diccionario y de la definición del índice. Finalmente, se mide también el sentimiento de las noticias de los periódicos los días siguientes a la publicación del Informe. Los resultados muestran que la lista de palabras recogida en el diccionario de referencia constituye una muestra robusta para estimar el sentimiento de estos textos. Esta herramienta constituye un valioso instrumento para analizar la repercusión del IEF, y también para cuantificar de forma objetiva el sentimiento que se está trasladando en él.

Palabras clave: minería de textos, análisis de sentimiento, procesado de lenguaje natural, comunicaciones de bancos centrales, estabilidad financiera.

Códigos JEL: C82, G28.

Abstract

This article shows a text mining application to extract information from financial texts and use this information to create sentiment indices. In particular, the analysis focuses on the Banco de España's financial stability reports from 2002 to 2019 in their Spanish version and on the reaction of the press to these reports. To calculate the indices, the first Spanish dictionary of words with a positive, negative or neutral connotation has been created, as far as we know, within the context of financial stability. The robustness of the indices is analyzed by applying them to different sections of the report, and using different variations of the dictionary and the definition of the index. Finally, sentiment is also measured for newspaper news in the days following the publication of the report. The results show that the list of words collected in the reference dictionary constitutes a robust sample to estimate the sentiment of these texts. This tool constitutes a valuable methodology to analyze the repercussion of financial stability reports, while objectively quantifying the sentiment that is being transferred in them.

Keywords: text mining, sentiment analysis, natural language processing, central bank communications, financial stability.

JEL classification: C82, G28.

1. Introducción

Entre las diferentes publicaciones del Banco de España, el *Informe de Estabilidad Financiera* (IEF) destaca como herramienta de comunicación esencial, no solo en lo relativo a los riesgos del sistema financiero español y a la rentabilidad y solvencia de las entidades de depósito españolas, sino también en relación con la política y las medidas macroprudenciales institucionales. Publicado desde 2002, el Informe constituyó inicialmente el primer capítulo de la *Revista de Estabilidad Financiera,* y pasó a ser una publicación independiente en noviembre de 2004. Desde su primera publicación, la prensa especializada se ha hecho eco de este Informe, incluyendo las citas textuales que consideraba más interesantes y tratando de extraer los mensajes clave, con el objetivo de resumir su contenido dentro de las limitaciones naturales de espacio impuestas por el medio.

Aunque en el ámbito de las publicaciones de bancos centrales se han realizado diversos estudios [entre otros, Apel y Grimaldi (2012), Born *et al.* (2014), Correa *et al.* (2018) y Apergis y Pragidis (2019)] para analizar el contenido de diferentes tipos de comunicaciones, no hemos encontrado referencias de análisis en español, ni tampoco relativas al impacto del IEF en la prensa escrita.

Tradicionalmente, el análisis de contenido ha sido una disciplina del ámbito de las ciencias sociales, con una perspectiva cualitativa en la que el peso principal del investigador recaía en su labor de clasificar textos [Aureli (2017)]. Con la aparición de los ordenadores, surgen iniciativas para automatizar el análisis de contenido con una perspectiva más cuantitativa, originando lo que posteriormente se denominaría «minería de textos» como disciplina separada del análisis de contenido. Uno de los primeros trabajos relacionados con el análisis de sentimiento basado en la búsqueda de palabras en diccionarios de categorías usando ordenadores tenía como subtítulo «Un enfoque computarizado al análisis de contenido» [Stone *et al.* (1966)]. En él se describía el programa *The General Inquirer,* que procesaba un texto y buscaba cada una de sus palabras asignándole una categoría dentro de un diccionario de categorías [Stone *et al.* (1962)] [1]. Desde

[1] Estaba programado con tarjetas perforadas para ser ejecutado inicialmente sobre ordenadores de IBM de la serie 709 y posteriormente de la serie 7090-7094. El IBM 7090 costaba alrededor de 3 millones de dólares y pesaba 8 toneladas [Weik (1961)].

entonces han aparecido diferentes aproximaciones a este tipo de análisis, y el aumento de la capacidad de cómputo ha facilitado el uso de técnicas de aprendizaje automático basadas en redes neuronales.

Las distintas herramientas y tecnologías que se engloban dentro de lo que se conoce como «minería de textos» permiten extraer información estructurada y cuantitativa a partir de la información no estructural presente en los textos de un conjunto de fuentes (p. ej., informes, noticias, páginas web, blogs o mensajes en redes sociales). Dentro de la minería de textos, el análisis de sentimiento es una herramienta de clasificación de documentos[2] que trata de determinar el grado de polaridad de un texto entre dos extremos o afectos, como positivo-negativo o fuerte-débil. En el caso de una polaridad positiva-negativa, es frecuente usar el término «tono» para referirse al sentimiento de un texto [Kearney y Liu (2014)]. Como resultado, se obtiene un índice o métrica de los documentos analizados, que tendrá dos polaridades y que representa el tono del documento.

Empleando técnicas de minería de textos, este documento propone un análisis estructurado y cuantitativo de textos financieros en castellano, mediante la creación de un índice de sentimiento que mida la polaridad positiva-negativa de estos documentos (es decir, el grado de optimismo o pesimismo que reflejan). El conjunto de textos que se analizan está formado por todos los números del IEF publicados por el Banco de España desde 2002 hasta 2019 en su versión en español. El ejercicio se completa aplicando el mismo procedimiento a un conjunto de artículos de prensa asociados a estos informes durante el mismo período.

Para calcular los índices de sentimiento, en este trabajo se ha definido el primer diccionario en castellano de palabras con tonalidad (positiva, negativa o neutra) dentro del ámbito de la

[2] Según la clasificación de Miner *et al.* (2012), la minería de textos es un campo de conocimiento que tiene siete esferas de actividad: clasificación de documentos, extracción de información, procesamiento del lenguaje natural (PLN), extracción de conceptos, minería web, recuperación de información y agrupación de documentos. Aunque el análisis de sentimiento se considera un campo dentro la clasificación de documentos, puede hacer uso también de técnicas propias de la minería web o del PLN.

estabilidad financiera. Como se menciona más adelante, la lengua española presenta particularidades que la hacen más compleja a la hora de implementar estas técnicas de minería de texto. Además, el documento contribuye a la metodología de creación de diccionarios de sentimiento por dos vías: primero, al incorporar técnicas de análisis del grado de acuerdo en el proceso de anotación o asignación de tonalidad; y, segundo, al desarrollar una herramienta específica para facilitar este proceso. Para ver la aplicabilidad del índice y del diccionario de referencia, se realiza un análisis comparado del sentimiento de las noticias de prensa relacionadas con el IEF.

Frente a técnicas más complejas, la usada en este documento para determinar el sentimiento mediante la búsqueda de palabras en un diccionario de tonalidades es conceptualmente intuitiva, suele tener buenos resultados y estos son fáciles de interpretar. Además, no está tan restringida por el tamaño del *corpus* disponible. El reducido número de textos con los que se cuenta sobre estabilidad financiera dificulta la aplicación de métodos de aprendizaje automático, que requieren un conjunto significativo de muestras de entrenamiento. Así, disponer de un análisis inicial como el presentado aquí puede facilitar el trabajo de creación de conjuntos de entrenamiento requeridos por este tipo de técnicas y servir de referencia para futuros estudios que quieran evaluar el desempeño del uso de técnicas basadas en aprendizaje automático. Por otro lado, la dificultad principal de la técnica adoptada radica en la necesidad de disponer de un diccionario de categorías. En este caso, de un diccionario que clasifique las palabras en positivas, negativas y neutras, dentro del contexto de la estabilidad financiera.

Entre los trabajos relacionados con el análisis textual y la comunicación de banca central en el ámbito de la estabilidad financiera, cabe mencionar el de Born *et al.* (2014), en el que realizan un análisis de sentimiento del primer capítulo de los informes de estabilidad financiera de 37 países entre 1996 y 2009 en su versión inglesa, haciendo uso de la aplicación DICTION 5.0[3]. Correa *et*

[3] Un producto comercial de análisis de textos asistido por ordenador [Digitext, Inc. (s.f.)].

al. (2018) estudian para el período 2000-2017 los informes de estabilidad de 64 países en su versión inglesa, más los publicados por el Banco Central Europeo y el Fondo Monetario Internacional. En este caso, se basan en un diccionario de tonalidades específico creado por los autores [Correa *et al.* (2017)], con 96 palabras positivas y 295 negativas.

La aplicación de diccionarios de uso general (p. ej., DICTION) para textos financieros da resultados menos precisos que la de diccionarios más específicos, como muestran los análisis de Loughran y McDonald (2011) y de Henry y Leone (2016). Al observar que tres cuartas partes de las palabras negativas del diccionario de tonalidades genérico Harvard-IV-4 no tenían un sentido negativo en el contexto de los informes anuales de empresas, Loughran y McDonald (2011) crearon un diccionario específico para informes financieros. Además, Correa *et al.* (2018) consideraron que el contexto de estabilidad financiera tenía diferencias suficientes respecto al de los informes financieros como para justificar también el uso de un diccionario específico, diferente al de Loughran y McDonald (2011). En particular, los textos de estabilidad financiera emplean una serie de palabras que no necesariamente tienen una connotación negativa, pero que los diccionarios generalistas de sentimiento las clasificarían como negativas. La palabra *confined* en inglés, por ejemplo, que normalmente tiene un tono negativo, en estabilidad financiera tiene un tono positivo, ya que se usa en relación con *limiting negative spillovers* [Correa *et al.* (2018)]. De la misma manera que no existen estudios que realicen análisis de sentimiento en español relacionados con la estabilidad financiera, tampoco existen diccionarios de tonalidad públicos con términos específicos del contexto financiero. La traducción directa de diccionarios en inglés, como el diccionario de Correa *et al.* (2017), no es la mejor alternativa, dado que la connotación positiva o negativa de una palabra no siempre se traslada en la traducción. Por ejemplo, palabras que en un idioma son polisémicas y no adecuadas para ser incluidas en el diccionario pueden no serlo en otra lengua. Por esta razón, se ha optado por la construcción de un diccionario de tonalidad propio y específico en español para el contexto de estabilidad financiera.

En la siguiente sección se mostrará la metodología empleada para el procesamiento de los textos y la creación del diccionario. Seguidamente, se definirán las reglas para el cálculo de los distintos índices propuestos para analizar el sentimiento. En la sección de resultados se analizará la robustez del índice de sentimiento, se examinará la consistencia del diccionario y se calculará el índice para distintas secciones del Informe. Finalmente, se calculará el índice para los artículos de prensa relacionados con aquel.

2. Creación de un diccionario en español en el contexto de estabilidad financiera

El diccionario se crea asignando tonalidades a las palabras utilizadas en el IEF del Banco de España. La muestra de textos analizada incluye los distintos informes en formato PDF desde 2002 hasta noviembre de 2019, que constituyen un total de 35[4].

Para poder realizar el análisis de las palabras y de las frases es necesario procesar estos documentos. El diagrama completo del proceso puede observarse en la Figura 1. Con el fin de preservar el flujo del texto y facilitar la extracción, se usa la herramienta comercial *Nuance Power PDF Advanced*, que permite convertir documentos PDF a documentos Word, extrayendo el texto incluso cuando está en imágenes. Esta primera conversión posibilita mantener la estructura de los documentos originales en un formato editable. Dado que se pretende realizar un análisis independiente de la introducción y del cuerpo para captar las posibles diferencias y solo se trata de 35 informes, se seleccionan de manera manual los apartados correspondientes y se almacena en ficheros separados el contenido textual, exceptuando los pies de página, cabeceras, índice y página del título. De manera programática se realizan actuaciones adicionales sobre el texto para tratar de reducir los errores de procesamiento. Entre otras, se corrigen guiones mal extraídos y se juntan palabras divididas por un guion al final de línea. Finalmente, se pasa el texto a minúsculas para

[4] Todos los informes pueden descargarse en formato PDF en la página web del Banco de España https://www.bde.es/bde/es/secciones/informes/boletines/Informe_de_Estab/.

agilizar el análisis. Tradicionalmente, uno de los problemas asociados al procesamiento de textos en idiomas diferentes al inglés ha sido la existencia de caracteres no presentes en la lengua inglesa. En el caso del español, tenemos la «ñ», las vocales acentuadas o la «u» con diéresis. Muchas herramientas y librerías de programación están orientadas al mercado anglosajón y tienen problemas con estos caracteres. Para solventar estas limitaciones, eventualmente se opta por eliminar los acentos y las diéresis, y por convertir la «ñ» a «n», o incluso a la combinación «ny». En nuestro caso, y debido a que las herramientas usadas carecen de esa limitación, no se realiza este tipo de transformación[5].

Una vez extraídos los textos, el IEF se divide en dos subconjuntos, uno con la introducción y otro con el resto del cuerpo. El motivo de esta separación radica en que, generalmente, los procesos de redacción y discusión entre ambas partes suelen ser distintos; y, en particular, la introducción habitualmente está sujeta a una revisión más intensa. El Informe, una vez procesado, tiene una media de 27.554 palabras, en tanto que la media de las introducciones es de 1.820 y la de los cuerpos de 25.734. Agrupadas en frases, las introducciones tienen 47 frases de media, y los cuerpos, 677 [6].

Antes de realizar la revisión de palabras que se han de incluir en el diccionario, se extraen todas las palabras individuales y se eliminan las palabras *vacías*, es decir, las muy comunes y que no aportan valor al análisis, como artículos, pronombres y preposiciones (p. ej., «un», «el», «la», «le», «se», «les», «con», «bajo», «mediante», etc.). El número total de términos obtenidos es de 14.736, correspondientes a 6.111 raíces. Palabras que solo se diferencian en género, número o forma verbal

[5] Para programar el proceso, se ha usado Python 3.7, que trabaja nativamente en Unicode y permite codificar sin problemas los caracteres españoles. Además, se han empleado dos librerías específicas para el procesamiento de textos que tampoco tienen este tipo de limitación: NLTK y spaCy.

[6] En análisis de textos se denomina *token* al elemento básico de proceso, pudiendo ser cualquier conjunto de caracteres alfanuméricos limitados por espacios o símbolos de puntación. A efectos prácticos, es equivalente a «palabra», y en este documento se ha optado por usar «palabra» en el sentido de *token*. De la misma manera, una frase comprende el texto contenido entre dos puntos seguidos, o salto de línea en su defecto.

pueden tener polaridades diferentes, por lo que en nuestro análisis la tonalidad no se define a nivel de lexema o raíz, sino a nivel de lema o palabra, con la flexión correspondiente. El español es una lengua mucho más flexiva que el inglés. Los adjetivos y sustantivos tienen morfemas tanto de género como de número, mientras que en inglés los adjetivos no tienen morfemas de género y solo los sustantivos tienen morfemas de número. La complejidad de las formas verbales también es superior en español. Esto hace que para cada lexema o raíz se multiplique el número de palabras o lemas que pueden formarse y de las que deba revisarse su tonalidad. Determinadas palabras pueden tener una connotación positiva o negativa únicamente en alguna de sus formas, y no en otras. Siguiendo las recomendaciones de Loughran y McDonald (2011), la connotación de cada palabra se define de manera deductiva, es decir, dentro del contexto en el que se usa; en este caso, el del IEF.

Para facilitar la revisión se ha creado una herramienta que permite el análisis individual de las palabras según su connotación dentro de las frases en las que aparecen (véase el esquema de la Figura 2). En la herramienta, al seleccionar una palabra se muestran las frases en las que esta aparece en los diferentes informes de estabilidad, y el anotador decide por juicio experto si la palabra tiene una tonalidad positiva, negativa o neutra. Las palabras en las que pudiera existir alguna duda o desacuerdo pueden también marcarse en la herramienta. Del total de términos, en una primera revisión por juicio experto se identifican 3.706 palabras susceptibles de tener tonalidad. A esas palabras se les asigna una connotación (positiva, negativa o neutra) según el contexto y se crea la primera anotación o anotación de referencia. Posteriormente se realizan tres particiones del listado de palabras y se distribuyen entre otros seis anotadores con formación académica universitaria en economía y conocimiento nativo del español, de manera que cada palabra del listado de 3.706 es revisada por dos de esos seis anotadores[7].

[7] Los anotadores no tienen experiencia previa en codificación de textos ni formación específica en lingüística. Se les realizó una primera sesión formativa en la herramienta y en guías básicas para determinar la polaridad de las palabras.

El objetivo del proceso de anotación es crear un *estándar de oro* o diccionario base que nos permita calcular el sentimiento para los textos de estabilidad financiera en castellano con un grado alto de fiabilidad. En este caso, a partir del conjunto de anotaciones, se establecen tres fases para resolver los desacuerdos y fijar el estándar de oro. Primero, se realiza una revisión individual de las anotaciones de manera comparada con la otra anotación de la misma partición, haciendo uso de la herramienta (fase 1). A continuación, los anotadores de las otras particiones revisan las palabras para las que continúa habiendo discrepancias, y se obtiene así la opinión de todos los anotadores para cada una de estas palabras (fase 2). Finalmente, se asigna la tonalidad por mayoría o *sabiduría del grupo*, teniendo en cuenta también la anotación inicial de referencia (fase 3).

Las diferentes fases del proceso se muestran, a modo de esquema, en la Figura 2. Nótese que las palabras en las que hay acuerdo tras la fase 1 no son analizadas por el resto de los anotadores, con lo que su tonalidad no tiene un apoyo explícito de una mayoría de anotadores. El hecho de que tengan una doble revisión y coincidan con la referencia sin conocerla reduce su grado de incertidumbre, constituyendo así un compromiso aceptable entre precisión y esfuerzo de anotación. El diccionario resultante, resolviendo los desacuerdos por sabiduría del grupo, contiene 376 palabras negativas y 189 positivas. El resto de las palabras, que son la mayoría, tienen una connotación neutra. En la *Figura 3* se muestran en forma de nube, y diferenciando por su tonalidad, las palabras del diccionario resultante o estándar de oro.

Al igual que ocurre en Correa *et al.* (2017), se detectan palabras con connotaciones diferentes en estabilidad financiera respecto a la presente en diccionarios de tonalidades genéricos o incluso en diccionarios para contextos financieros. Por ejemplo, la palabra «morosidad» podría interpretarse como un término con connotación negativa, pero en general en el IEF va asociado a «tasa de morosidad», que no conlleva sentimiento por sí misma. Lo mismo ocurre con «fallidas» o «dudosas», que suelen ir asociadas a la métrica «número de operaciones».

En relación con particularidades del español, encontramos casos de polisemia que no se dan en inglés. Por ejemplo, la palabra «bien», que podría considerarse positiva, aparece en general asociada a la expresión «si bien», que en sí misma tampoco tiene connotación. Aunque más frecuentemente en plural, también puede asociarse al concepto de propiedad (p. ej., en «bienes inmuebles»), por lo que en el contexto del IEF en español es de polaridad neutra. Por otro lado, al definir las tonalidades por lexemas, se observa que en ocasiones una palabra en una determinada forma tiene una connotación en el diccionario final, mientras que otras formas de la misma palabra tienen otra connotación. Por ejemplo, «absorbida» o «absorbidas», que generalmente se usan refiriéndose a «pérdida/s», tienen asignadas connotaciones positivas; y, sin embargo, «absorbido» y «absorbidos» se asocian con conceptos variados, lo que les confiere una connotación neutra.

3. Análisis de la concordancia entre los anotadores

Para evaluar la calidad del diccionario, en esta sección se muestra el nivel de acuerdo observado durante el proceso de obtención del estándar de oro. La calidad de un proceso de anotación suele medirse en términos de exactitud y concordancia. La exactitud hace referencia al grado de cumplimiento con las especificaciones, mientras que la concordancia se refiere al nivel de acuerdo de los anotadores entre ellos. Se considera que estas dos medidas están correlacionadas y que una medida de la concordancia suele ser indicativa de la calidad de la anotación [Wong y Lee (2013)].

La medida de la concordancia de un proceso de anotación se realiza en términos del acuerdo entre anotadores o *fiabilidad interjuez*. Para medir el grado de acuerdo en la creación del diccionario base se emplean estadísticos no paramétricos de tipo *Chance-Corrected Agreement Coefficients* (CAC), que ajustan el acuerdo observado por la probabilidad de concordancia debida al azar. Además, como medida de distancia para ponderar las discrepancias empleamos la distancia

euclídea, es decir, la diferencia absoluta entre los valores anotados [véanse Antoine *et al.* (2014) y Mielke *et al.* (2011)][8].

Zhao *et al.* (2013) evidenciaron las limitaciones de los diferentes estadísticos de tipo CAC existentes y recomendaron distintos estadísticos en función de cada situación, no encontrando ningún índice apropiado para los casos en los que los anotadores tratan de ser precisos pero involuntariamente realizan asignaciones aleatorias. Por esta razón, en nuestro estudio se calculan tres de estos estadísticos de tipo CAC en cada fase del proceso de anotación. En particular, los conocidos como *kappa de Cohen* (κ), *alfa de Krippendorff* (α) y AC_2 *de Gwet*. Estos estadísticos se emplean normalmente en estudios médicos y en psicología, y son aplicables también a procesos de anotación como el nuestro, según afirman Gwet (2008), Zhao *et al.* (2013) y Wongpakaran *et al.* (2013). La evolución de los estadísticos para las diferentes fases de la anotación se muestra en la Tabla 1. En el apéndice B se muestran unos ejemplos simplificados de cálculo de estos estadísticos. Nótese que la κ de Cohen solo mide el acuerdo entre dos evaluadores; por tanto, tenemos un valor para cada combinación de dos anotadores, mientras que los estadísticos α y AC_2 permiten medir acuerdos entre dos o más anotadores. Para los cálculos de estos estadísticos, teniendo en cuenta múltiples anotadores, se consideran todas las palabras. A aquellas palabras para las que no hay discrepancia en la fase 1 (es decir, tienen tres anotaciones iguales: las de los dos anotadores más la inicial de referencia) se les asignan valores vacíos para el resto de los anotadores que no las han evaluado, ya que el cálculo de los coeficientes α y AC_2 permite tener valores vacíos para alguno de los anotadores[9].

[8] En una categorización con tres niveles de polaridad (-1, 0 y 1), la distancia binaria (1 si son distintos y 0 si son iguales) y la distancia euclídea coinciden en la mayoría de las ocasiones, ya que las diferencias entre anotadores suelen ser entre neutro y cualquiera de las otras polaridades, y raramente se produce una discrepancia entre negativo y positivo.

[9] Todos los estadísticos se calcularon usando la librería irrCAC [Gwet (2019)] para el lenguaje de programación R.

La interpretación más común de los valores de κ y α es que la concordancia es alta por encima de 0,8, mientras que es moderada entre 0,67 y 0,8 [véanse Krippendorff (2004) y Antoine *et al.* (2014)]. En el caso de AC_2, a falta de una interpretación recomendada debido a su reciente creación, es frecuente usar el mismo criterio que para el resto de los coeficientes de tipo CAC. Los valores de κ obtenidos en nuestro proceso se encuentran en el intervalo [0,169, 0,794]. Tras finalizar el proceso de anotación, y teniendo en cuenta las siete anotaciones, se obtiene un valor de α igual a 0,454 y un coeficiente de acuerdo AC_2 igual a 0,905. En general, la categoría neutra tiene una mayoría de asignaciones, resultando el número de palabras significativamente superior a las palabras de las categorías positiva y negativa. En la literatura sobre anotaciones relacionadas con la evaluación del sentimiento es habitual tener valores más bajos de lo que cabría esperar para los coeficientes de κ y de α [Antoine *et al.* (2014)]. Esta situación se conoce como «primera paradoja de kappa», por la que distribuciones simétricas tienden a tener *kappas* más elevadas que distribuciones en las que una categoría prevalece [véanse Cicchetti y Feinstein (1990) y Callejas y López-Cózar (2008)]. En estos casos, es recomendable usar estadísticos alternativos, como el coeficiente AC_2, que es menos sensible a distribuciones asimétricas, es decir, cuando una de las categorías prevalece sobre el resto [Gwet (2008)]. Aun así, dado que Zhao *et al.* (2013) consideran que AC_2 tiende a dar resultados altos, vemos interesante mostrar los valores de κ y de α como referencia adicional.

Estas diferencias entre los distintos coeficientes analizados indican también las dificultades que surgen al establecer un consenso en la polaridad de ciertas palabras. En particular, para la partición A los coeficientes de κ y de α fueron especialmente bajos, debido probablemente a la existencia de un sesgo de anotador, en este caso originado por la tendencia de uno de los anotadores a polarizar la anotación para contextos en los que otros anotadores optarían por una clasificación neutra. El proceso de revisión individual comparada da como resultado unos coeficientes superiores, y esto es indicativo de que este método es un mecanismo efectivo y relativamente poco

costoso para mejorar el acuerdo cuando existen dos anotadores. Como se ha indicado anteriormente, en el caso de las palabras evaluadas por los siete anotadores optamos por una resolución de desacuerdos mediante *sabiduría del grupo* (es decir, por mayoría simple), minimizando así el posible efecto del sesgo de anotador.

Cabe destacar que la mayoría de las discrepancias se dan entre anotaciones neutras y positivas o entre neutras y negativas. En la *Figura 4* se muestra la distribución de distancias en las anotaciones por número de palabras en relación con la primera anotación de referencia (R) y en relación con los otros anotadores (i). Una distancia de 2 indica una discrepancia entre polaridades positiva y negativa, mientras que una distancia de 1 indica una discrepancia entre polaridad positiva o negativa y neutra. Una distancia de 0 representa acuerdo entre anotadores. Así, la barra $A(i) - A(R)$ representa la distribución de diferencias en asignación de polaridad de las palabras de la partición A entre el anotador i y la anotación inicial de referencia R. Se observa que el anotador $i = 2$ de la partición A tiene un criterio más flexible a la hora de asignar polaridades. En cualquier caso, no se producen casi discrepancias con distancia igual a 2, lo que también reduce el impacto de un posible sesgo de anotador.

4. Cálculo del índice de sentimiento para el *Informe de Estabilidad Financiera*

Una vez creado el diccionario y analizado el grado de acuerdo en su creación, podemos pasar a analizar el sentimiento de un determinado texto en español dentro del contexto de la estabilidad financiera. En nuestro caso, el sentimiento de un texto se medirá como una función del número de palabras con tonalidad que aparecen en él. Para ello necesitamos: primero, identificar las palabras del texto que aparecen en el diccionario estándar de oro, que determina qué palabras tienen tonalidad; luego, definir unas reglas para contar el número de palabras dentro de cada categoría; y, finalmente, una función que nos lleve del número de palabras categorizadas a una

medida o índice de sentimiento. Por ejemplo, una medida de sentimiento consistiría en sumar el total de las palabras del texto con una polaridad determinada (positiva o negativa).

Se puede intentar crear reglas que traten de identificar el sentimiento de una frase considerando las múltiples posibilidades que permite la construcción gramatical de un idioma. Estas reglas serían, en general, muy complejas y costosas de implementar en su totalidad. En este análisis se ha optado por simplificar en la medida de lo posible el esquema de reglas. De este modo, la regla de cómputo se reduce a contar directamente todas las palabras del texto identificadas con una determinada tonalidad, pero teniendo en cuenta si están negadas dentro de la frase. En caso de que la palabra aparezca negada, si tiene tonalidad positiva se considera que pasa a tener una connotación negativa, y si tiene una tonalidad negativa (doble negación) pasa a ser considerada neutra. Esta regla recoge la idea de que algo que «no es bueno» tiene una tonalidad negativa, pero algo que «no es malo» no necesariamente tiene una tonalidad positiva [véanse Correa *et al.* (2018) y Loughran y McDonald (2011)]. Para determinar si una palabra está negada, buscamos las siguientes palabras en las tres posiciones anteriores a la palabra que se analice: «menos», «no», «nunca», «sin», «pérdida» y «disminución». En el recuento del número de palabras con tonalidad no se ha considerado la presencia de intensificadores del tipo «muy», «gran», «mucho», «enormemente», que podrían modificar la carga de sentimiento de una palabra, sino que se les ha asignado el mismo peso a todas las palabras con polaridad. Aunque la regla implementada para el cómputo de palabras sea sencilla, hay que tener en cuenta que durante el proceso de anotación se ha asignado la tonalidad considerando el contexto y, por tanto, ya se ha tenido presente implícitamente la estructura de la frase durante la fase de creación del diccionario[10].

[10] Por ejemplo, en el caso del término «adverso», que tiene tonalidad negativa en el diccionario final, los anotadores habrán ponderado todos los casos de uso para decidir que esta palabra tiene una tonalidad mayoritariamente negativa, considerando también los casos en los que esta palabra se usa con tonalidad neutra (por ejemplo, cuando se emplea en el bigrama «escenario adverso», en referencia a las pruebas de esfuerzo a la banca).

Una vez implementada la regla para identificar y contar las palabras con connotación dentro de un texto, el siguiente paso es definir las funciones para medir el sentimiento o tonalidad global de ese texto. Así, para cada Informe de la muestra, identificado por su fecha de publicación t, definimos la *Negatividad* como el número de palabras negativas sobre el total de las palabras del texto:

$$Negatividad_t = \left(\frac{\#\, negativas}{\#\, palabras\ totales}\right)_t; \qquad [1]$$

y la *Positividad* como el número de palabras positivas sobre el total:

$$Positividad_t = \left(\frac{\#\, positivas}{\#\, palabras\ totales}\right)_t. \qquad [2]$$

Finalmente, la *Negatividad neta* se define como la diferencia entre la *Negatividad* y la *Positividad* para la fecha t:

$$Negatividad\ neta_t = Negatividad_t - Positividad_t = \left(\frac{\#\, negativas - \#\, positivas}{\#\, palabras\ totales}\right)_t \qquad [3]$$

Por ejemplo, para calcular sus índices de sentimiento, Feldman *et al.* (2010) y Correa *et al.* (2018) emplean algunas de las medidas de las ecuaciones [1]-[3]. Otros análisis de contenido [p. ej., Apel y Grimaldi (2012) y Apergis y Pragidis (2019)] definen medidas relativas, es decir, en vez de emplear el número de palabras totales (con y sin connotación), se mide el diferencial entre palabras negativas y positivas con respecto al total de palabras con tonalidad (es decir, palabras positivas más negativas), obteniendo así una medida de *Negatividad neta relativa*, escalada entre -1 y 1. En nuestro análisis se opta por utilizar esta medida relativa para definir el índice de sentimiento del *Informe de Estabilidad Financiera (ISEF);* de este modo, para la fecha t:

$$ISEF_t \equiv Negatividad\ neta\ relativa_t = \left(\frac{\#\, negativas - \#\, positivas}{\#\, negativas + \#\, positivas}\right)_t \qquad [4]$$

Nótese que las funciones [3] y [4] son crecientes con el aumento de la negatividad; es decir, cuanto mayor es el valor del índice $ISEF_t$, menos optimista es el sentimiento transmitido, coincidiendo así con el criterio de signos empleado por Loughran y McDonald (2011) y por Correa *et al.* (2018). En general, los resultados usando cualquiera de las dos medidas deberían ser comparables, ya que la longitud del documento normalmente es proporcional a la suma de palabras negativas y positivas.

Las posibles divergencias que quepa encontrar entre las medidas [3] y [4] pueden deberse a las palabras neutras aparecidas en secciones del documento más académicas y que podrían en cualquier caso descartarse. Correa *et al.* (2018), a pesar de usar el número de palabras totales en el denominador, descartan en su análisis aquellas secciones que consideran más teóricas y no relacionadas con la estabilidad financiera, reduciendo así el número de palabras.

5. Resultados para la introducción y el cuerpo del Informe

El cálculo de las distintas medidas mostradas en las ecuaciones [1] a [4] nos permite analizar desde distintas perspectivas las variaciones de tono del Informe a lo largo del tiempo. En la *Figura 5* se muestra el resultado de calcular estas medidas para los textos correspondientes a las introducciones del Informe, usando el diccionario de referencia y las reglas descritas. Para ayudar a situar el Informe y el valor del índice asociado en su contexto macrofinanciero, se muestran en los gráficos los eventos económicos más relevantes (a escala nacional o internacional) ocurridos a lo largo del período de estudio (2002-2019). Se observa que el número de eventos relevantes no se distribuye de manera uniforme por toda la serie, sino que se concentra a partir del inicio de la crisis financiera global, es decir, de 2008-2009 en adelante. Para el período 2003-2007, la ausencia de eventos coincide con una fase expansiva del ciclo económico. A su vez, estos eventos pueden agruparse en varios subconjuntos: los eventos relacionados con la crisis financiera global y con la crisis de la deuda soberana europea, los más específicos del sistema financiero español y de la situación española, y, en los últimos años, los relacionados con eventos geopolíticos. Por completitud, también se resaltan los períodos correspondientes a los mandatos de los diferentes gobernadores[11]. Aunque no es el objetivo de este análisis, los cambios de gobernanza podrían potencialmente influir en el tono de los textos (p. ej., cambios en el comité de redacción y en la Comisión Ejecutiva que aprueba el Informe).

[11] J. Caruana (julio de 2000-julio de 2006), M. Á. Fernández Ordóñez (julio de 2006-junio de 2012), L. M. Linde (junio de 2012-junio de 2018) y P. Hernández de Cos (junio de 2018-fecha del análisis).

En el panel a) de la *Figura 5* se muestran la Negatividad, la Positividad y la Negatividad neta (ecuaciones [1], [2] y [3], respectivamente) para los textos del IEF correspondientes a la introducción, desde otoño de 2002 hasta otoño de 2019. En general, cuando la *Positividad* sube, la *Negatividad* baja, y viceversa, estando así estos índices negativamente correlacionados entre sí, mientras que la *Negatividad* está positivamente correlacionada con la *Negatividad neta* (véase tabla 2, con coeficientes de correlación entre índices). Como principal ventaja, las medidas de *Positividad* y *Negatividad* (ecuaciones [1] y [2]) facilitan la identificación de situaciones en las que la *Negatividad neta* y la *Negatividad* divergen, debido al efecto de la *Positividad;* es decir, aquellas situaciones en las que la proporción de palabras positivas es significativamente diferente de un período a otro, invirtiendo la tendencia del índice de *Negatividad* que lo tiene en cuenta [2]. Un ejemplo de una situación aparentemente contradictoria ocurre en otoño de 2006, cuando se observa que, con respecto al Informe de primavera de 2006, la *Negatividad* (línea roja) aumenta al mismo tiempo que la *Positividad* (línea verde), dando como resultado una ligera caída en la *Negatividad neta* (línea azul). El índice neto disminuye por el aumento de palabras positivas, pero sucede al mismo tiempo que un incremento en el número de palabras con tonalidad negativa.

Para la introducción, se observa que la *Negatividad neta* [línea azul del panel a)] y el *ISEF* [línea marrón del panel b) de la figura 5] siguen evoluciones muy similares y están fuertemente correlacionados (por encima del 90%) (véase tabla 2). El máximo histórico para el *ISEF* se corresponde con el Informe de otoño de 2011, meses antes del segundo rescate a la economía griega. En este Informe casi no aparecen palabras con tonalidad positiva. Por el contrario, uno de los valores más bajos de la serie (es decir, uno de los valores con sentimiento más positivo) se produce en el Informe de primavera de 2015, después del comienzo del «plan Draghi»[12].

Los índices no solo nos permiten analizar posibles discrepancias entre informes, sino que también ayudan a analizar las discrepancias dentro de un mismo Informe. En nuestro caso, al separar la

[12] En referencia a la aplicación de la política monetaria denominada «expansión cuantitativa», que el BCE inició en 2015 mediante un plan de compra de activos de bancos comerciales.

introducción del resto del cuerpo del IEF, podemos estudiar posibles diferencias entre los textos de estas dos partes. La introducción funciona a modo de resumen ejecutivo y muestra normalmente los puntos más destacados del Informe. El resto del cuerpo hace mención no solo a la situación económica y del sector bancario, sino también a otros aspectos que pudieran ser más académicos o divulgativos.

En el panel b) de la figura 5 se muestra el índice de sentimiento o *ISEF*, definido como la *Negatividad neta relativa* (ecuación [4]), calculado para la introducción y el cuerpo del IEF de la muestra. Además, en la fila 1 de la tabla 3 se muestran los coeficientes de correlación entre los distintos índices calculados usando los textos de la introducción y del cuerpo del Informe. En particular, comparando el *ISEF* del cuerpo y la introducción en el panel b) de la figura 5, se observa una correlación alta y positiva (0,90), como cabría esperar, considerando que la introducción representa un resumen del resto del Informe. Estos resultados aportan confianza en la fiabilidad del índice y del diccionario utilizado. El gráfico también muestra que existe más variabilidad para el índice calculado con la introducción que para el índice calculado con el cuerpo. En la introducción, al ser un texto más corto, aparece un menor número de palabras con connotación, y el índice es más sensible a la entrada o salida de alguna de estas palabras. Además, se observa que durante ciertos períodos (especialmente, entre 2003-2007) la introducción y el cuerpo han mostrado niveles del *ISEF* dispares. Aunque por los posibles errores de medida, el nivel no es tan informativo como la evolución de los dos índices. A pesar de la alta correlación en la evolución de ambos, hay algunos puntos en los que el sentido de estos índices diverge. Por ejemplo, en el segundo semestre de 2017 el índice para la introducción muestra un aumento del optimismo (es decir, caída del índice) con respecto al Informe anterior, mientras que el índice para el resto del Informe presenta un aumento del pesimismo (es decir, subida del índice). En concreto, para este Informe la introducción se inicia con «Las tensiones geopolíticas latentes no han evitado que la recuperación económica, a nivel internacional, continúe su senda positiva», para a continuación expandir la información en relación con la senda positiva iniciada. Por su lado, el cuerpo del

Informe destaca en varias ocasiones las tensiones políticas en Cataluña, dedicando incluso un recuadro específicamente a este tema.

A la luz de los resultados anteriores, se muestra la utilidad de obtener una variable cuantitativa para resumir la información de un conjunto de textos. Las distintas medidas mostradas nos han permitido analizar las variaciones y la evolución del sentimiento del Informe, señalando fácilmente los puntos con valores más extremos o los cambios de tendencia. La variación del sentimiento mostrada es coherente con el desarrollo de los eventos económicos más relevantes ocurridos a lo largo de la muestra. Además, se ha podido comparar el sentimiento de distintas secciones dentro de un mismo Informe; en concreto, entre la introducción y el resto del documento. Los resultados mostrados para ambas secciones son coherentes, y el índice permite centrar la atención en el Informe donde existen divergencias más llamativas.

6. Análisis de robustez: completitud del diccionario

Para analizar la robustez del índice hemos realizado dos análisis. Por un lado, de manera similar a Correa *et al.* (2017), hemos determinado el rango de variación del *ISEF* sobre el cuerpo mediante la técnica de eliminación del 5% de las palabras del diccionario de manera aleatoria durante 1.000 iteraciones. En el panel a) de la figura 6 se muestran las bandas de confianza del índice obtenidas en este proceso de eliminación de palabras y recálculo del índice. Además, para determinar no solo el rango sino también la distribución de la tendencia alcista o bajista del *ISEF* en los diferentes diccionarios generados, se calculó igualmente la pendiente en cada punto, de manera que los valores de pendiente positivos indicaban una tendencia alcista y los negativos una tendencia bajista. El panel a) de la figura 6 muestra también las distribuciones de las pendientes en cada punto, representadas en forma de diagramas de caja y bigotes. Se observa que hay numerosos puntos en los que la totalidad de la muestra se sitúa a un lado o a otro del eje, indicando que la pendiente positiva o negativa se mantiene en los 1.000 diccionarios. El cálculo de las

pendientes permite, además, identificar puntos en los que se ha producido un cambio significativo en el sentimiento del texto con respecto a los textos anteriores.

Por otro lado, y aprovechando la información obtenida del proceso de anotación, se realizaron nuevamente 1.000 iteraciones, pero esta vez asignando a cada palabra del diccionario una tonalidad con una probabilidad proporcional al número de anotadores que asignaron esa tonalidad. Así, si una palabra fue categorizada como positiva por seis anotadores y un anotador la consideró neutra, se le asigna una probabilidad de 0,14 de ser neutra y 0,86 de ser positiva. El resultado se muestra en el panel b) de la *Figura 6*. En este caso, las franjas se amplían, pero los rangos de valores de pendientes se mantienen, en general, a un lado o a otro del eje cuando las pendientes son pronunciadas. Los colores representan los percentiles habituales de 0-25, 25-50, 50-75 y 75-100. La ventaja de esta aproximación es que permite aprovechar la información del proceso de anotación sin reducirlo a un listado final de palabras clasificadas a tres niveles. Según defienden Wong y Lee (2013), existen situaciones en las que la ambigüedad intrínseca de una clasificación es lo que causa el desacuerdo, y el hecho de que la clasificación sea ambigua es de por sí una información importante. Tratar de forzar un acuerdo artificial en estas situaciones puede no ser la mejor opción, con lo que vemos esta aproximación basada en probabilidades como una opción interesante para preservar el desacuerdo legítimo. Un caso significativo es el de la palabra «crisis». Según el diccionario de Correa *et al.* (2017), esta palabra tiene una tonalidad neutra, pero en nuestro diccionario recibió esa clasificación solo por dos anotadores, mientras que cinco anotadores la consideraron negativa. Esta discrepancia posiblemente sea debida a que en un buen número de casos la palabra se usa como referencia histórica a la crisis financiera global, aunque finalmente hubo una mayoría de anotadores que consideraron que, en general, su uso daba un tono negativo al texto. Al asignar una clasificación final negativa a la palabra, se pierde el hecho de que hubo un desacuerdo legítimo de dos anotadores, que consideraron que no estaba justificado asignar una tonalidad negativa a esa palabra. Mostrar rangos de confianza basados en el número de

anotadores que optaron por cada clasificación permite exponer una visión más fiel al sentimiento percibido por los anotadores.

7. Comparativa con indicadores financieros

Además de analizar la consistencia interna y la robustez del diccionario, el hecho de tener una medida cuantitativa del sentimiento nos permite poder comparar el índice con indicadores financieros cuantitativos, por un lado, para observar la coherencia del índice de sentimiento y, por otro, para ver cuánta información de los indicadores relacionados con la estabilidad financiera está incorporada en él [Correa *et al.* (2018)]. El número reducido de observaciones hace que este análisis tenga ciertas limitaciones y que los resultados sean más dependientes de eventos específicos y de pequeñas variaciones en la muestra. Conviene destacar que el objetivo en este punto no es medir la capacidad del índice de sentimiento para establecer predicciones sobre los indicadores financieros.

En particular, se analizan las siguientes regresiones para ver cómo los indicadores financieros (variables cuantitativas) explican la variación temporal del índice de sentimiento:

$$ISEF_t = \mu_i + \beta_i X_{i,t-h} + \varepsilon_{i,t} \,, \qquad [5]$$

donde $X_{i,t-h}$ representa la variable económica i analizada en cada caso con un retardo igual a h períodos. Los resultados para los coeficientes β_i y sus errores estándar se muestran en la tabla 4 para la regresión contemporánea ($h = 0$) y para la regresión con un retardo de dos semestres en el indicador ($h = 2$). Esta última permite ver si la información financiera cuantitativa se incorpora de manera coherente en los textos del Informe, medidos mediante el índice de sentimiento. Las variables $X_{i,t}$ consideradas son una muestra representativa del tipo de variables, relacionadas con el ciclo de la estabilidad financiera, de las que se hace seguimiento en el Informe. En concreto, siguiendo los indicadores empleados en Correa *et al.* (2018), se han considerado los siguientes tipos de variables: variables macroeconómicas y de política monetaria, como la variación trimestral del PIB para España, la tasa de desempleo, el tipo de interés interbancario a corto plazo

y el tipo de interés virtual a corto plazo; variables relacionadas con el ciclo de crédito, como la brecha de crédito sobre PIB y la ratio del servicio de la deuda privada no financiera; variables que miden la evolución de las valoraciones, como la rentabilidad por dividendo del índice bursátil IBEX-35, el valor de mercado sobre valor contable de los bancos y las variaciones en los precios reales de la vivienda en España; y, finalmente, variables de riesgos financieros, como la prima de riesgo de crédito del sector bancario español, la volatilidad de la divisa y la volatilidad del IBEX-35.

Los resultados de las regresiones se calculan tanto para las variables contemporáneas como para las variables retardadas dos períodos (un año). Aunque de alguna manera la confiabilidad de los estimadores se ve afectada por el tamaño reducido de la muestra (36 observaciones), los resultados señalan que los índices de sentimiento están correlacionados contemporáneamente ($h = 0$) con varios de los indicadores del ciclo de la estabilidad financiera analizados. En particular, los coeficientes son significativos para aquellas variables que miden riesgos, como, por ejemplo, la volatilidad del IBEX-35 o la prima de riesgo del sector bancario recogida en la cotización del diferencial del CDS de uno de sus bancos más representativos. De este modo, un aumento en la volatilidad y en la prima de riesgo de crédito está acompañado por una subida en el índice de sentimiento, es decir, un aumento del pesimismo medido en el Informe. La correlación entre el índice y la variación del PIB de España es también alta y significativa. En este caso, el coeficiente es de signo negativo: un empeoramiento o decrecimiento de la economía supone una subida en el índice de sentimiento. Además, también se observan coeficientes significativos para los indicadores relacionados con la evolución de las valoraciones. De este modo, un deterioro en los precios reales de la vivienda o en el valor de mercado de las acciones de los bancos frente al valor contable está relacionado con un empeoramiento en el índice de sentimiento. Finalmente, cuando las regresiones se realizan con las variables retardadas un año, los coeficientes para estas variables mantienen su signo, aunque su nivel de significatividad disminuye, pasando algunos a no ser significativos.

8. Otras reglas para el cálculo del índice: metodología TF-IDF

En algunos análisis de sentimiento es habitual ponderar las palabras con sentimiento, de manera que no se les asigne la misma importancia a todas. Una ponderación habitual es la denominada TF-IDF, que hace referencia a la expresión en inglés *Term Frequency-Inverse Document Frequency* («frecuencia de término-frecuencia inversa de documento») [Manning *et al.* (2009)]. Este concepto tiene sus orígenes en los algoritmos para la recuperación de información a partir de una cadena de búsqueda, y se usa para calcular la importancia de un término dentro de un documento. Para ello se pondera cada término teniendo en cuenta dos aspectos: por un lado, considerando la frecuencia de término, esto es, el número de veces que aparece el término en el documento analizado, de manera que un determinado término es más significativo cuantas más veces aparece; por otro, incluyendo la frecuencia inversa de documento, teniendo en cuenta el número de documentos en el que aparece ese término dentro del conjunto de documentos analizados. Se considera que un término será poco relevante en un determinado documento si el mismo término aparece en otros muchos documentos. Por ejemplo, en el IEF, la palabra «riesgo» aparece en todos los informes, con lo que su relevancia en un determinado Informe será menor que otras palabras que sean específicas de ese Informe.

Loughran y McDonald (2011) recomiendan una ponderación TF-IDF como resultado de su análisis, razonando que permite minimizar los errores causados por la inclusión en el diccionario de palabras que podrían considerarse vacías o poco relevantes por su alta frecuencia[13]. Por otro lado, Jegadeesh y Wu (2013) argumentan que no hay ninguna razón específica para que la frecuencia de ocurrencia de una palabra en un documento esté relacionada con el sentimiento percibido, sugiriendo una ponderación alternativa basada en la reacción percibida. En el caso de su análisis, enfocado a informes anuales, se basaron en la reacción de los mercados. En nuestro caso, no existe

[13] Loughran y McDonald (2011) realizan su análisis de acuerdo con palabras negativas únicamente, ya que observan frecuentemente el uso de palabras positivas incluso para describir situaciones negativas, haciendo muy complicado un procesado automático que tuviera en cuenta el verdadero sentido de las palabras positivas.

ningún valor cuantitativo que pueda tomarse como medida directa de la reacción percibida a un IEF. Shapiro *et al*. (2019) realizan una comparativa empleando artículos de periódico y concluyen que la diferencia entre un mecanismo proporcional y uno ponderado por TF-IDF no es significativa. Por otra parte, Correa *et al.* (2017) no aplican ningún mecanismo de ponderación, dado que consideran que son más útiles en muestras grandes de pequeños documentos.

Al usar lemas y no lexemas en la definición del diccionario, es posible que dos lemas que compartan raíz tengan ponderación diferente según TF-IDF solo por su diferente frecuencia en el documento. Como hemos comentado, el español es una lengua muy flexiva, y no parece razonable pensar que una diferencia de frecuencia en dos palabras que únicamente se distinguen en su flexión puedan tener un nivel de polaridad diferente. Por ello, en nuestro caso, y a modo comparativo con el índice *ISEF* no ponderado (ecuación [4]), realizamos la ponderación usando la frecuencia del lema base, en lugar del lema específico. Así, calculamos el índice ponderado por TF-IDF mediante la diferencia:

$$ISEF_w = \sum_{T \in negativas} w_T - \sum_{T \in positivas} w_T \ , \qquad [5]$$

donde w_T representa la ponderación del término T (con connotación) dentro del documento (en el apéndice B se presentan los detalles de cómo se calculan estas ponderaciones). El resultado obtenido no difiere significativamente del índice *ISEF* sin ponderar (véanse las correlaciones en las tablas 2 y 3). Como en el caso de Shapiro *et al.* (2019), este resultado podría ser indicativo de la robustez del diccionario final.

Aunque en nuestro análisis de sentimiento no apreciamos beneficios en usar un índice ponderado mediante TF-IDF, sí encontramos útil esta ponderación para identificar las palabras más relevantes en un determinado Informe, como puede observarse en las nubes de palabras de la figura 7, donde el tamaño de estas representa el peso del lema base dentro de cada Informe en función del algoritmo TF-IDF. Darle mayor relevancia a los términos específicos de cada Informe permite ver la evolución de los conceptos más relevantes a lo largo de las diferentes publicaciones.

9. Comparativa con el sentimiento de los artículos de prensa

La creación de un índice de sentimiento permite hacer análisis comparativos de un conjunto de documentos. Esta comparativa puede ser entre distintos períodos o fechas, entre distintas partes de un mismo Informe o entre dos textos de orígenes distintos pero relacionados. En esta sección se analizará este último caso, comparando los índices calculados para los textos del Informe con los índices calculados para artículos de prensa de los días posteriores a la publicación del IEF que hacen referencia a este. Para extraer el texto de los artículos de prensa, se emplean los boletines de prensa internos del Banco de España, en los que se lleva a cabo una selección de las noticias relacionadas con el Banco, el sistema financiero y la coyuntura económica. En particular, se usan los boletines de hasta tres días posteriores a la publicación del Informe y se analizan los artículos clasificados bajo el epígrafe de «noticias sobre el Banco de España» y que además hacen mención del Informe. El proceso de extracción de los textos sigue las mismas pautas que para el Informe. Para el cálculo del número de palabras con connotación se han usado el diccionario ya creado para el Informe y las mismas reglas de cálculo definidas en la sección anterior.

En las filas 2 y 3 de la tabla 3 se muestran las correlaciones entre los índices calculados para las noticias de periódicos con los índices calculados para la introducción y el cuerpo del IEF. Para la *Negatividad*, *Negatividad neta* y el índice *ISEF* (sin ponderar y ponderado), la correlación entre periódicos y textos del Informe es relativamente alta y significativa. En particular, para el *ISEF* es igual a 0,66 y 0,61 con respecto a la introducción y al cuerpo, respectivamente. Por el contrario, las correlaciones calculadas para la *Positividad* no son significativas a un nivel de confianza del 5%. La figura 8 muestra la evolución del *ISEF* para los tres conjuntos de textos considerados. El índice para los periódicos presenta más variabilidad que el resto de los índices, por el número limitado de noticias que hacen referencia al Informe, sobre todo en el caso de los más antiguos. En general, como se muestra en la tabla 3, se observa una correlación ligeramente mayor con la introducción que con el cuerpo del IEF.

Al consolidar información textual no estructurada (IEF y noticias de periódicos) mediante una métrica numérica estructurada (p. ej., *ISEF),* se facilitan la identificación y la comparación de ciertos puntos de interés. Observando la figura 8, resulta de interés resaltar algunas fechas concretas. Por ejemplo, para el Informe de otoño de 2006, la *Positividad* y la *Negatividad* aumentaban con respecto al Informe anterior (véanse figura 5 y sección 2.1). En comparación con el Informe, se aprecia un aumento brusco de negatividad en los periódicos. En la figura 9 se muestran las diferencias entre la nube de palabras para la introducción y la nube para los periódicos. Los periódicos destacaron el crecimiento significativo de los activos dudosos del crédito a construcción y promoción inmobiliaria (empleando palabras como «peligro» «preocupación» y «temor»), dejando de lado otros aspectos más positivos del Informe (p. ej., «dinamismo» de la actividad bancaria y la situación coyuntural «favorable» a pesar de los elementos de «preocupación»).

Otro punto destacable corresponde al IEF de primavera de 2015, donde aparecen los valores más bajos tanto para el índice de la introducción como para el del cuerpo del Informe. En este caso, la tendencia de los índices del Informe y la del índice de los periódicos divergen sustancialmente. En el caso de los índices del IEF, el inicio del «plan Draghi» supone un aumento del optimismo, frente a la mayor volatilidad predominante hasta comienzos de 2015 desde el Informe anterior: presencia de palabras como «recuperación», «mejora», «favorable», frente a las negativas que tienen un menor peso relativo (véase la nube de palabras en la figura 9). Por su parte, el índice de los periódicos recoge una mayor presencia de palabras negativas y, por tanto, un aumento del pesimismo. En este caso, los periódicos se centraron en mensajes relacionados con la recomendación de realizar recortes en los costes, en particular en el número de oficinas bancarias.

Surgen así diferentes cuestiones al analizar estas discrepancias. Por un lado, están en línea con la menor correlación observada entre el índice construido con los textos del Informe y el índice con las noticias de los periódicos (véase tabla 3). Si se tienen en cuenta únicamente las positividades,

esta correlación no es significativa. Conviene señalar en este punto la dificultad de establecer un criterio para saber si estas correlaciones con los artículos de prensa son altas o bajas, o si existen conexiones entre las divergencias temporales entre estos índices y la estrategia de comunicación del Banco. Por ejemplo, en un contexto donde el índice de sentimiento de los artículos de prensa es negativo, pudiera darse el caso de que el comité encargado del Informe prefiera comunicar la fortaleza del sistema a ciertos eventos, dando lugar a una divergencia entre los índices, sin que esto signifique una pérdida de transparencia en la comunicación.

Finalmente, creemos relevante resaltar la utilidad del diccionario en este caso. Aunque en él no aparezcan todas las palabras con tonalidad empleadas en los periódicos, sí se observa que el diccionario funciona correctamente como una aproximación para el cálculo del sentimiento, incluso en textos fuera de la muestra.

10. Conclusiones

En los últimos años ha cobrado especial interés el análisis de contenido de los diferentes tipos de comunicaciones generadas por los bancos centrales. En su mayoría, estos estudios están diseñados para textos en inglés y con enfoques pensados para dicho idioma. En este documento se presenta el primer diccionario de sentimiento en español en el ámbito de la estabilidad financiera. Además de extender este tipo de análisis de contenido y minería de textos al español, se describe en detalle el procedimiento seguido y se analiza la concordancia de las anotaciones de sentimiento empleando distintos estadísticos no paramétricos. De este modo, se contribuye a la literatura de este tipo de procesos de anotación en cualquier idioma. El esfuerzo de anotación normalmente está infravalorado y es importante tener herramientas que ayuden en esta tarea.

Haciendo uso del diccionario creado a partir de los textos extraídos del IEF del Banco de España desde 2002 hasta 2019, el trabajo analiza diversos índices de sentimiento definidos mediante funciones sobre el número de palabras con connotación. Los índices creados sirven para medir de

manera cuantitativa los textos del Informe, es decir, información no estructurada, y se muestran consistentes cuando se calculan para distintas partes de este (la introducción y el cuerpo). Además, son robustos ante variaciones del diccionario (aleatorias o basadas en los resultados de las anotaciones) o ante cambios en la metodología para definir las ponderaciones de cada palabra dentro del índice (equiponderada o ponderada por la frecuencia de ocurrencia del término). Los índices son coherentes respecto al desarrollo de los eventos macroeconómicos más relevantes a lo largo de la muestra y, además, contienen información relacionada con otros indicadores financieros cuantitativos (p. ej., la prima de riesgo de crédito del sector bancario).

Finalmente, conviene tener en cuenta que, en este tipo de análisis, los resultados también están sujetos a cambios estructurales en la composición del Informe, ya sean por cambios en el comité de redacción o en la Comisión Ejecutiva que aprueba su publicación, y que podrían haber influido en el tono de los textos.

Disponer de un diccionario abre la posibilidad de realizar análisis de sentimiento de otros textos financieros en español y permite compararlos de manera bastante objetiva. De este modo, en el documento se calculan los índices para las noticias de los periódicos relacionadas con el Informe y se analiza la reacción de estos a su publicación. Los resultados muestran que la lista de palabras recogida en el diccionario de referencia es suficientemente consistente como para poder obtener un estimador fiable del sentimiento de los artículos de prensa. Se observa una mayor correlación entre los índices del Informe y los periódicos cuando se calculan usando solo palabras con tonalidad negativa que cuando se calculan considerando solo palabras positivas.

Creemos también que este análisis sirve de base y referencia para posibles enfoques futuros alternativos: análisis con reglas adicionales, exploración de subíndices temáticos o uso de modelos estadísticos de aprendizaje automático, que tengan en cuenta mejor el contexto de uso de cada palabra con el objetivo de ganar precisión.

Apéndice A. Cálculo de los coeficientes de acuerdo interjuez

La forma general de los diferentes coeficientes de acuerdo de tipo CAC usados en este estudio es la siguiente: [14]

$$\text{Coeficiente tipo CAC} = \frac{P_a - P_e}{1 - P_e}, \qquad [A.1]$$

siendo P_a la probabilidad o porcentaje de acuerdo observado y P_e la probabilidad o porcentaje de acuerdo debido al azar. Para dos anotadores, los coeficientes κ de Cohen y AC_2 de Gwet difieren en la manera de estimar la probabilidad de acuerdo debido al azar P_e, mientras que el coeficiente α de Krippendorff difiere también en la manera de estimar P_a.

Tomemos el siguiente ejemplo con dos anotadores que registran el sentimiento de 10 palabras clasificando su polaridad entre positiva (+), negativa (-) y neutra (○):

	X	Y
1	○	○
2	○	○
3	○	○
4	○	○
5	○	○
6	○	○
7	+	+
8	+	-
9	+	+
10	-	-

Los dos anotadores registran las 10 palabras de la misma manera, a excepción de una palabra, que el anotador X considera que tiene una polaridad positiva, mientras que el anotador Y considera que

[14] De manera equivalente, estos coeficientes también pueden definirse en términos de desacuerdo:

$$\text{Coeficiente tipo CAC} = \frac{P_a - P_e}{1 - P_e} = 1 - \frac{D_o}{D_e},$$

siendo $D_e = 1 - P_e$ el porcentaje ponderado de desacuerdo atribuido al azar y $D_o = 1 - P_a$ el porcentaje ponderado de desacuerdo observado.

la tiene negativa. La información anterior puede representarse de manera resumida en la siguiente tabla de contingencia, C_{XY}:

		Y			
		-	○	+	
X	-	1	0	0	1
	○	0	6	0	6
	+	1	0	2	3
		2	6	2	10

La diagonal representa el número de palabras en las que ambos anotadores están de acuerdo para cada polaridad, mientras que los elementos fuera de la diagonal representan las diferentes combinaciones de desacuerdo.

El cálculo de los coeficientes de acuerdo permite considerar también acuerdos de segundo nivel o acuerdos parciales (es decir, entre negativo o positivo, y neutro). De este modo, pueden calcularse probabilidades ponderadas ajustando por el nivel de acuerdo. En nuestro análisis se emplean ponderaciones lineales, mediante la siguiente matriz de pesos, W:

		Y		
		-	○	+
X	-	1	0,5	0
	○	0,5	1	0,5
	+	0	0,5	1

Cálculo del coeficiente κ de Cohen

El coeficiente κ de Cohen [Gwet (2008)] se define como:

$$\kappa = \frac{P_a - P_{e|\kappa}}{1 - P_{e|\kappa}}, \qquad [A.2]$$

donde la probabilidad de acuerdo P_a es la suma ponderada de acuerdos dividida por el total de anotaciones. Es decir, es la suma de elementos de la matriz de contingencia C_{XY} ponderados por los elementos de la matriz de pesos W. Para el ejemplo, sería: $P_a = \frac{1+6+2}{10} = 0{,}9$.

Por otro lado, $P_{e|\kappa}$ se calcula como la probabilidad ponderada de acuerdo debida al azar. Para ello, se obtienen las probabilidades de que los valores de la matriz de contingencia se hayan dado por azar y asumiendo que las probabilidades de asignar una anotación (+, -, ○) por los anotadores X e Y son independientes, es decir, $P_X \cap P_Y = P_X P_Y$. La probabilidad de que Y asigne una determinada anotación se calcula dividiendo por el total de anotaciones (10) el número de veces que haya asignado esa anotación (los totales de cada columna en la matriz C_{XY}. Ponderando estas probabilidades de acuerdo por azar con los elementos de la matriz de pesos lineales W, se obtiene $P_{e|\kappa}$.

Para nuestro ejemplo, resumido en la matriz de contingencia C_{XY}, la matriz de probabilidades debida al azar sería:

		Y		
		-	○	+
X	-	0,02	0,06	0,02
	○	0,12	0,36	0,12
	+	0,06	0,18	0,06

Multiplicando por los pesos: $P_{e|\kappa} = 0{,}02 + 0{,}36 + 0{,}06 + 0{,}06 \times 0{,}5 + 0{,}12 \times 0{,}5 + 0{,}12 \times 0{,}5 + 0{,}18 \times 0{,}5 = 0{,}68$. Finalmente, el coeficiente de Cohen resultante sería: $\kappa = \frac{0{,}9-0{,}68}{1-0{,}68} = 0{,}6875$.

Cálculo del coeficiente α de Krippendorff

La forma general del coeficiente α de Krippendorff [Gwet (2011)] se define como:

$$\alpha = \frac{P_{a|\alpha} - P_{e|\alpha}}{1 - P_{e|\alpha}} \qquad [A.3]$$

El valor de $P_{a|\alpha}$ se calcula en este caso de manera parecida a como se obtiene para el coeficiente de Cohen, pero resulta siempre un valor ligeramente superior. Cuando no hay anotaciones en blanco, $P_{a|\alpha}$ viene dado por:

$$P_{a|\alpha} = \left(1 - \frac{1}{nr}\right) P_a + \frac{1}{nr}, \qquad [A.4]$$

donde n es el número de palabras que se han de anotar y r el número de anotadores. En nuestro ejemplo: $P_{a|\alpha} = \left(1 - \frac{1}{10\times 2}\right) \times 0{,}9 + \frac{1}{10\times 2} = 0{,}905$.

Para este coeficiente, la probabilidad ponderada debida al azar $P_{e|\alpha}$, a partir de los resultados de la anotación, se calcula obteniendo primero las probabilidades de que cualquier palabra esté dentro de una categoría (+, -, ○). En este caso, resulta útil ver las anotaciones en forma de una tabla de acuerdos, donde cada columna representa una categoría y cada fila una palabra, indicando en cada celda el número de anotadores que clasifican esa palabra en esa categoría. Como en el coeficiente anterior, para obtener $P_{e|\alpha}$ se ponderan las probabilidades por azar usando la matriz de pesos lineales W.

La probabilidad de que a una palabra cualquiera se le asigne una categoría concreta vendría determinada por las veces que esa categoría se ha asignado divididas por el total de asignaciones (20 en nuestro caso). Para nuestro ejemplo, la tabla de acuerdos vendría dada por:

	-	○	+
1	0	2	0
2	0	2	0
3	0	2	0
4	0	2	0
5	0	2	0
6	0	2	0
7	0	0	2
8	0	0	2
9	1	0	1
10	2	0	0
Total	3	12	5

	-	○	+
Probabilidad (π)	0,15	0,60	0,25

A partir de la tabla anterior, se obtiene la matriz de probabilidades debidas al azar usando el mismo valor de probabilidad para X y para Y (π en el ejemplo) y multiplicando:

		Y		
		-	○	+
X	-	0,0225	0,09	0,0375
	○	0,09	0,36	0,15
	+	0,0375	0,15	0,0625

y la probabilidad ponderada $P_{e|\alpha}$ para este coeficiente: $P_{e|\propto} = 0{,}0225 + 0{,}36 + 0{,}0625 + 0{,}09 \times 0{,}5 + 0{,}09 \times 0{,}5 + 0{,}15 \times 0{,}5 + 0{,}15 \times 0{,}5 = 0{,}685$. Finalmente, el coeficiente de Krippendorff resultante sería: $\alpha = \frac{0{,}905-0{,}685}{1-0{,}685} = 0{,}698$.

Cálculo del coeficiente AC_2 de Gwet

La fórmula del coeficiente AC_2 es:[15]

$$AC_2 = \frac{P_a - P_{e|AC}}{1 - P_{e|AC}} \qquad [A.5]$$

El cálculo de P_a en el coeficiente de Gwet se realiza de la misma manera que para el coeficiente de Cohen. Sin embargo, la probabilidad por azar ponderada, $P_{e|AC}$, se define como:

$$P_{e|AC} = \frac{T_w}{q(q-1)} \times \sum_{k=1}^{q} \pi_k(1 - \pi_k), \qquad [A.6]$$

donde T_w es la suma de los pesos de la matriz de ponderación W, q el número de clasificaciones posibles y π_k la probabilidad de que un anotador asigne una clasificación k a una determinada observación, es decir, las calculadas en la tabla de acuerdos usada en el cálculo del coeficiente α de Krippendorff [Gwet (2014)].

Para nuestro ejemplo, definido por una clasificación de dos anotadores, tres niveles (+, -, ○) y una suma de pesos de 5: $P_{e|AC} = \frac{5}{6} \sum_{k\in\{+,\circ,-\}} \pi_k(1 - \pi_k)$, con π_k las probabilidades de la tabla de acuerdos: $\pi = (0{,}15;\ 0{,}60;\ 0{,}25)$. Así: $P_{e|AC} = \frac{5}{6}\big(0{,}15 \cdot (1 - 0{,}15) + 0{,}6 \cdot (1 - 0{,}6) + 0{,}25 \cdot (1 - 0{,}25)\big) = 0{,}4625$. Y el coeficiente de Gwet sería: $AC_2 = \frac{0{,}9-0{,}4625}{1-0{,}4625} = 0{,}81395$.

Como puede observarse en la Figura A.1, la propia fórmula de la forma general de los coeficientes CAC hace que su valor decaiga rápidamente cuando P_e es alta, lo que implica que podamos tener valores significativamente diferentes en función del método adoptado para el cálculo de P_e, y que, a valores de P_e altos, más fácilmente podremos tener valores del CAC bajos.

[15] La forma general de los coeficientes AC_1 y AC_2 de Gwet es la misma; la diferencia radica en que AC_1 usa la matriz identidad como matriz de ponderación, por lo que no tiene en cuenta acuerdos de segundo nivel.

Apéndice B. Ponderación TF-IDF

En una ponderación TF-IDF, la *frecuencia de término* $TF_{T,d}$ se define como el número de veces que aparece el término T en el documento d, de manera que un determinado término es más significativo cuando aparece más veces dentro de un documento. Por otro lado, la *frecuencia de documento* DF_T se define como el número de documentos que contienen el término T. Para ponderar la relevancia de un término, y siendo N el número total de documentos, la *frecuencia inversa de documento* de una palabra T se define como:

$$IDF_T = \log \frac{N}{DF_T} \qquad [B.1]$$

Así, la relevancia TF-IDF se define como el producto de la *frecuencia de término* y la *frecuencia inversa de documento*:

$$TF\text{-}IDF_{T,d} = TF_{T,d} \times IDF_T \qquad [B.2]$$

Dado que no parece razonable que un término que aparezca veinte veces en un documento sea realmente veinte veces más importante que otro que solo aparezca una vez, es habitual ponderar usando el logaritmo de la frecuencia de término mediante:

$$WTF_{T,d} = \begin{cases} 1 + \log TF_{T,d} & si\ TF_{T,d} \geq 1 \\ 0 & en\ otro\ caso \end{cases}, \qquad [B.3]$$

con lo que la relevancia ponderada pasaría a ser:

$$WTF\text{-}IDF_{T,d} = WTF_{T,d} \times IDF_T, \qquad [B.4]$$

consiguiendo de esta manera amortiguar la relevancia de términos que aparecen repetidamente en un documento. Esta ponderación no tiene en cuenta la longitud del documento, ya que documentos más largos cuentan con mayor probabilidad de tener más frecuencia de un término relevante. Existen otros mecanismos de ponderación de la frecuencia de término que no solo usan la escala logarítmica, sino que además tienen en cuenta la longitud del documento a través de la media de frecuencias de término. Son las denominadas «ponderaciones *log ave»,* definidas como:

$$WTF_{T,d} = \begin{cases} \frac{1+\log TF_{T,d}}{1+\log(a_d)} & si\ TF_{T,d} \geq 1 \\ 0 & en\ otro\ caso \end{cases}, \qquad [B.5]$$

siendo $a_d = \text{ave}_{i \in d}(TF_{i,d})$ la media de frecuencias de todos los términos i del documento d (excluyendo las palabras vacías). Alternativamente:

$$a_d \equiv \text{ave}_{i \in d}(TF_{i,d}) = \sum \frac{n_d}{u_d}, \qquad \text{[B.6]}$$

donde n_d es el número de palabras del documento d y u_d el número de palabras eliminando los duplicados. A la luz de los resultados de su análisis, Loughran y McDonald (2011) recomiendan esta ponderación de frecuencia de término *log ave*.

De este modo, en nuestro análisis, además del índice *ISEF* sin ponderar (ecuación [4]), se calcula un índice de sentimiento (ecuación [5]) teniendo en cuenta la relevancia ponderada, w_T, de cada término T incluido en el diccionario final, definida como:

$$w_T \equiv WTF\text{-}IDF_{T,d} = \begin{cases} \frac{(1+\log(TF_{T,d}))}{(1+\log(a_d))} \log \frac{N}{DF_T} & si\ TF_{T,d} \geq 1 \\ 0 & en\ otro\ caso \end{cases}, \qquad \text{[B.7]}$$

siendo N el número total de informes que se han de analizar, DF_T el número de informes con la palabra T, $TF_{T,d}$ el número de veces que aparece la palabra T en el informe d, y a_d la media de frecuencias de término del informe d.

Apéndice C. Consideraciones sobre el proceso de anotación

Para futuros trabajos de anotación, se deja constancia en este apéndice de los diferentes comentarios al proceso realizados por los anotadores que participaron en este estudio. Al ser el proceso de anotación una novedad para todo el equipo, algunos anotadores reconocieron una variación en su criterio a medida que progresaban en las anotaciones. En algunos casos, inicialmente se optaba por inclinarse hacia una tonalidad concreta (positiva o negativa). A medida que avanzaban en el proceso de anotación, esta preferencia pasó a no ser tan significativa y se optó por el neutro cuando la tonalidad no estaba claramente definida. En otros casos se produjo la situación contraria: de preferir el neutro a no ser que se viera una tonalidad muy clara, se pasó a buscar una tendencia en los ejemplos que permitiera la clasificación de un término en positivo o negativo.

Debido a que la selección de las palabras se había realizado por frecuencia de lexema o raíz, a pesar de que la clasificación se llevaba a cabo a nivel de lema, se dio el caso de lemas que solo tenían uno o dos ejemplos. En estos casos, la variabilidad en la asignación de polaridad se veía influida por el criterio final. Si el anotador consideraba que la familia de la palabra tenía un patrón claro, podía optar por asignarle esa polaridad también a esa misma palabra. Otros anotadores tenían una tendencia a asignar en estos casos una polaridad neutra, aunque este criterio también fue variando con el tiempo.

En los casos en los que una determinada palabra siempre apareciera acompañada de otras que hacían que las frases fuesen de una misma polaridad, en general se optó por asignar a esa palabra la polaridad detectada en la frase. Se localizaron numerosos casos en los que la polaridad dependía de los modificadores que acompañaban a la palabra que se había de clasificar (p. ej., en relación con tendencias de subida o bajada). En estos casos se optó por una clasificación neutra, pero se vio la posibilidad de agregar reglas adicionales que permitieran incluir estas palabras teniendo en cuenta los modificadores. Además, a las palabras que en general aparecían en títulos o al pie de gráficos, se optó por asignarles una polaridad neutra.

Apéndice D. Listado de palabras del diccionario

Palabras positivas

absorbidas	capaces	mitiga	resisten
abundancia	capaz	mitigaban	resistido
abundante	cómoda	mitigado	restablecer
acomodaticia	contención	mitigar	restableciendo
acomodaticias	desendeudamiento	mitigaron	restablecimiento
acomodaticio	dinamismo	normalidad	restaurar
afiance	dinamizador	normalizado	revalorizaba
afianzado	disfrutan	normalizados	revalorizaciones
afianzamiento	eficaces	normalizando	revalorizado
afianzando	eficaz	normalizándose	revalorizaron
ágil	eficiente	normalizar	revalorizarse
alcista	eficientes	normalizó	revalorizó
alcistas	equilibrada	oportunidades	revitalización
aliviadas	equilibrado	optimismo	revitalizar
aliviado	excelente	ordenada	robusta
aliviando	excelentes	ordenado	robustas
aliviar	expandió	positiva	robusto
aliviará	expansiva	positivamente	robustos
aliviaron	favorable	positivas	saneada
alivio	favorablemente	positivo	saneado
amortigua	favorables	positivos	saneados
amortiguación	favorece	progreso	sanearon
amortiguador	favorecen	progresos	satisfactoria
amortiguan	favorecido	propicias	satisfactoriamente
amortiguar	favorecieron	propicio	sólida
amortiguarlos	fortaleciéndose	reaccionado	sólidas
amortiguarse	fortalecimiento	reactivación	solidez
apoyada	fortaleció	reactivándose	sólido
asentarse	fortaleza	reafirmando	solvente
atenuación	fortalezas	recuperación	solventes
atenuados	ganancia	recuperado	sostenibles
beneficiándose	ganó	recuperan	suaves
beneficiar	holgada	recuperando	suavizarán
beneficiara	holgadamente	recuperándose	superada
beneficiarán	holgadas	recuperar	sustenta
beneficiarían	holgado	recuperara	tranquilidad
beneficiaron	holgados	recuperaron	vigorosamente
beneficiarse	mejora	recuperarse	vigoroso
beneficien	mejorada	recuperase	
beneficioso	mejorado	recuperen	
benigna	mejoran	recuperó	
benignas	mejorando	reequilibrando	
benigno	mejorándose	reequilibrar	
benignos	mejorar	reforzado	
bienestar	mejoraron	reforzándolo	
buen	mejorase	reforzará	
buenas	mejores	reforzaron	
buenos	mejoría	reforzó	
calma	mejorías	remontado	
calmar	mejoró	renovado	

Palabras negativas

abrupta
abruptas
abrupto
abruptos
abusivo
acentuaban
acentuadas
acusados
adversa
adversas
adverso
adversos
afrontan
afrontarían
agotamiento
agravada
agravado
agravamiento
agravando
agravar
agravará
agravarían
agravó
agudas
agudizado
agudizamiento
agudizara
agudizaran
agudizaron
agudizó
agudo
agudos
altibajos
amenaza
amenazados
amenazan
amenazar
amenazas
anómala
arrastrado
asimétricos
ataque
ataques
atonía
atravesando
atraviesan
bache
brusca
bruscas
brusco
bruscos
colapsados
colapso
complejidades
complejo
complica
complicaciones
complicada
complicadas
complicado
complicados
complicando
complicar
complicarían
contagiadas
contagiado
contagiaron
contagie
contagio
contagió
contracción
contracciones
contractiva
contrae
contraerse
contrajo
contraproducentes
contrayendo
contrayéndose
convulso
costosa
costosas
costoso
costosos
crisis
cruda
dañar
dañaría
daño
débil
débiles
debilidad
debilidades
debilita
debilitada
debilitado
debilitamiento
debilitan
debilitar
debilitó
débilmente
decepcionante
decepcionantes
decepcionaron
deficiencias
deficiente
deficitaria
delicada
depresión
deprimidos
deprimirían
desaceleración
desastres
desconfianza
desencadenamiento
desequilibrada
desequilibrio
desequilibrios
desestabilizadores
desfavorable
desfavorablemente
desfavorables
destrucción
destruyendo
desvaneciendo
deteriora
deterioraban
deteriorada
deterioradas
deteriorado
deteriorando
deteriorándose
deteriorar
deteriorarse
deteriorase
deterioro
deterioró
difícil
difíciles
dificulta
dificultad
dificultada
dificultades
dificultado
dificultando
dificultándose
dificultar
dificultaría
dificultarían
disfunción
disfunciones
drástica
drásticas
drásticos
dudas
empeora
empeorado
empeoramiento
empeoramientos
empeoran
empeorando
empeorar
empeoró
endurecido
endureciéndose
endurecimiento
endurecimientos
erosión
erosionado
erosionar
escalada
escándalos
escasísima
estallar
estallido
estancada
estrangulamiento
estrangulamientos
evaporarse
excesivo
excesivos
falta
fatiga
frágil
frágiles
fragilidad
fragilidades
fragmentación
frenazo
frustró
grave
gravedad
gravemente
graves
guerra
impactará
inadecuados
incapaces
incapaz
incertidumbre
incertidumbres
incierta
inciertas
incierto
inciertos
inconveniente
indefinición
indeseado
ineficiencia
ineficiencias
inestabilidad
inestable
inestables
insostenible
insuficiencia
insuficiente
insuficientes
intervenida
intervenidas
intervenir
invalidar
inviabilidad
inviable
inviables
irregular
lamentablemente
lastrada
lastradas
lastrado
lastrar
lastre
lastró
lenta
lento
mal
mala
malas
merma
miedo
negativa
negativamente
negativas
negativo
negativos
obstáculo
oscilaciones
padece
padecían
pánicos
peligro
peligros
penalizado
peores
pérdida
perjudica
perjudicadas
perjudiciales
perjuicios
persistencia
persistente
persistentes
persistieron
perturbaciones
perversos
pesimismo
pesimista
pobre
precipicio
prematura
preocupación
preocupaciones
preocupado
preocupados
preocupante
preocupantes
presión
presiona
presionaban
presionada
presionadas
presionado
presionados
presionan
presionando
presionar
presionará
presionaría
presionaron
presiones
problemas
problemática
problemáticas
quebrar
quebró
quiebra
quiebras
ralentice
ralentiza
ralentización
ralentizar
ralentizara
ralentizarse
ralentizó
rebaja
rebajadas
rebrote
recaída
recalentamiento
recesión
recesivas
recrudecían
recrudecidos
recrudecieron
recrudecimiento
recrudeció
rémora
rescatadas
rescatar
resentido
resentirse
restaron
restringiendo
resurgido
resurgimiento
retraimiento
retrasa
retroceder
retrocedieron
retrocedió
retroceso
retrocesos
revés
secuelas
sensible
serias
serio
serios
severa
severas
severo
sobrecalentamiento
sombras
súbita
sufren
sufrida
sufridas
sufrido
sufridos
sufriendo
sufrieran
sufrieron
sufrió
sufrirán
suspensión
temor
temores
tensión
tensiona
tensionaban
tensionado
tensionamiento
tensionando
tensionaron
tensiones
titubeante
traumática
truncada
turbulencia
turbulencias
urgencia
virulencia
volátil
vulnerabilidad
vulnerabilidades
vulnerable
vulnerables

Referencias

Antoine, J.-Y., J. Villaneau y A. Lefeuvre (2014). «Weighted Krippendorff's alpha is a more reliable metrics for multicoders ordinal annotations: experimental studies on emotion, opinion and coreference annotation», *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* Gothenburg, Suecia.

Apel, M., y M. B. Grimaldi (2012). *The Information Content of Central Bank Minutes,* Sveriges Riksbank Working Paper Series, 261.

Apergis, N., e I. Pragidis (2019). «Stock Price Reactions to Wire News from the European», *International Advances in Economic Research,* 25(1), pp. 91-112.

Aureli, S. (2017). «A comparison of content analysis usage and text mining in CSR corporate disclosure», *The International Journal of Digital Accounting Research*, pp. 1-32.

Born, B., M. Ehrmann y M. Fratzscher (2014). «Central bank communication on financial stability», *Economic Journal*, pp. 701-734.

Callejas, Z., y R. López-Cózar (2008). «Influence of contextual information in emotion annotation for spoken dialogue systems», *Speech Communication,* 50, pp. 416-433.

Cicchetti, D., y A. Feinstein (1990). «High agreement but low Kappa: II. Resolving the paradoxes», *Journal of Clinical Epidemiology,* 43(6), pp. 551-558.

Correa, R., K. Garud, J. M. Londono y N. Mislang (2017). «Constructing a Dictionary for Financial Stability», *IFDP Notes*.

- (2018). *Sentiment in central banks' financial stability reports,* International Finance Discussion Papers, 1203.

Digitext, Inc. (s.f.). *DICTION is a computer-assisted text-analysis (CATA) program,* recuperado en 2019 de https://www.dictionsoftware.com/.

Feldman, R., S. Govindaraj, J. Livnat y B. Segal (2010). «Management's tone change, post earnings announcement drift and accruals», *Review of Accounting Studies,* 15, pp. 915-953.

Gwet, K. L. (2008). «Computing inter-rater reliability and its variance in the presence of high agreement», *British Journal of Mathematical and Statistical Psychology,* 61, pp. 29-48.

- (2011). «On The Krippendorff's Alpha Coefficient», *Communication Methods and Measures*.

- (2014). *Handbook of inter-rater reliability,* Advanced Analytics, LLC, 4.ª edición.

- (2019). «irrCAC: Computing Chance-Corrected Agreement (CAC)», obtenido de https://CRAN.R-project.org/package=irrCAC.

Henry, E., y A. J. Leone (2016). «Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone», *Accounting Review,* 91, pp. 153-178.

Jegadeesh, N., y D. Wu (2013). «Word Power: A New Approach for Content», *Journal of Financial Economics,* 110(3), pp. 712-729.

Kearney, C., y S. Liu (2014). «Textual sentiment in Finance: A survey of methods and models», *International Review of Financial Analysis,* 33, pp. 171-185.

Krippendorff, K. (2004). «Reliability in Content Analysis: Some Common Misconceptions and Recommendations», *Human Communication Research,* 30(3), pp. 411-433.

Krippner, L. (2015). *Term Structure Modeling at the Zero Lower Bound: A Practitioner's Guide,* Palgrave-Macmillan.

Loughran, T., y B. McDonald (2011). «When is a liability not a liability? Textual analysis», *Journal of Finance,* 66, pp. 35-65.

- (2016). «Textual Analysis in Accounting and Finance: A Survey», *Journal of Accounting Research,* 16, pp. 1-11.

Lowe, W., K. Benoit, S. Mikhaylov y M. Laver (2011). «Scaling Policy Preferences from Coded Political Texts», *Legislative Studies Quarterly,* 36.1, pp. 123-155.

Manning, C. D., P. Raghavan y H. Schütze (2009). *An Introduction to Information Retrieval,* Cambridge University Press.

Mielke Jr., P. W., K. J. Berry y J. E. Johnston (2011). «Robustness without rank order statistics», *Journal of Applied Statistics,* 38(1), pp. 207-214.

Miner, G., J. Elder, A. Fast, T. Hill, R. Nisbet y D. Delen (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications,* Academic Press.

Shapiro, A., M. Hale y D. W. Sudhof (2019). *Measuring News Sentiment,* Federal Reserve Bank of San Francisco Working Paper, 2017-01.

Stone, P. J., D. C. Dunphy, M. S. Smith y D. M Ogilvie (1966). *The General Inquirer: A Computer Approach to Content Analysis,* Cambridge, MIT Press.

Stone, P., R. Bales, J. Namenwirth y D. Ogilvie (1962). «The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information», *Behavioral Science,* 7(4), pp. 484-498.

Weik, M. H. (1961). *A Third Survey of Domestic Electronic Digital Computing Systems,* Maryland, Ballistic Research Laboratories, Aberdeen Proving Ground.

Wong, B. T., y S. Y. Lee (2013). «Annotating Legitimate Disagreement in Corpus Construction», *International Joint Conference on Natural Language Processing,* Nagoya, Japón.

Wongpakaran, N., T. Wongpakaran, D. Wedding y K. L. Gwet (2013). «A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples», *BMC Medical Research Methodology,* 13(61).

Zhao, X., J. S. Liu y K. Deng (2013). «Assumptions behind Intercoder Reliability Indices», *Communication Yearbook,* 36, pp. 419-480.

Tablas

Tabla 1

Análisis de las medidas de concordancia durante el proceso de creación del diccionario

Fase	Anotadores a_1, a_2			Anotadores a_3, a_4			Anotadores a_5, a_6			Anotadores a_1-a_6		Anotadores a_0-a_6	
	α	κ	AC_2	α	κ	AC_2	α	κ	AC_2	α	AC_2	α	AC_2
(1)	.044	.162	.356	.372	.379	.832	.544	.544	.868	.282	.711	.372	.787
(1b)	.278	.322	.599	.621	.625	.894	.794	.794	.940	.527	.824	.500	.840
(2)	.217	.253	.625	.199	.217	.819	.169	.169	.872	.166	.636	.170	.652
(3)	-	-	-	-	-	-	-	-	-	.404	.902	.454	.905

Nota: Los estadísticos o medidas de acuerdo calculadas son: la *kappa* de Cohen (κ), el *alpha* de Krippendorff (α) y el coeficiente AC_2 de Gwet. Los detalles técnicos sobre el cómputo de estos estadísticos no paramétricos se muestran en el apéndice A. Se considera un valor alto de concordancia por encima de 0,8 y moderado entre 0,67 y 0,8 [Krippendorff (2004), Gwet (2008)]. En la fase 1, la lista inicial de palabras (3.706 lexemas anotados por un anotador de referencia, R) se divide en tres particiones de idéntico tamaño (P_A, P_B y P_C), revisadas cada una de ellas por dos anotadores. En caso de desacuerdo, hay una segunda ronda de revisiones entre los dos anotadores de cada partición para resolver discrepancias (fase 1b, que solo tiene en cuenta el conjunto de palabras en las que hay discrepancias, $P_{A'}$, $P_{B'}$ y $P_{C'}$). Se muestran los estadísticos para las fases 1 y 1b en la primera y segunda filas de la tabla. La fase 2 incluye solo el subconjunto de palabras de cada partición para las que hay discrepancias después de la fase 1b ($P_{A'}$, $P_{B'}$ y $P_{C'}$); y son revisadas por los cuatro anotadores de las otras particiones (a_1 y a_2 revisan $P_{B'}$ y $P_{C'}$; a_3 y a_4 revisan $P_{A'}$ y $P_{C'}$, y a_5 y a_6 revisan $P_{A'}$ y $P_{B'}$). En la fase 3 se juntan todas las anotaciones, por eso solo se calcula para el conjunto de todos los anotadores (a_1-a_6) más el de referencia o inicial (a_0).

Tabla 2

Coeficientes de correlación entre los distintos tipos de índices calculados para cada texto

	Introducción	Cuerpo	Periódicos
Positividad frente a *Negatividad*	-0.53**	-0.22	-0.18
$ISEF$ frente a *Negatividad neta*	0.95***	0.95***	0.84***
$ISEF$ frente a $ISEF_W$	0.90***	0.85***	0.63***

Nota: En cada columna se muestran las correlaciones (coeficientes de Pearson) entre dos índices distintos calculados para un mismo texto (introducción, cuerpo y periódicos). Los índices utilizados son: *Positividad*, *Negatividad*, *Negatividad neta*, *ISEF* e $ISEF_W$ (ponderado por el algoritmo TF-IDF), correspondientes con las ecuaciones [1], [2], [3], [4] y [5], respectivamente. Los asteriscos *, ** y *** representan la significatividad del coeficiente de correlación a los niveles de confianza del 5 %, 1 % y 0,1 %, respectivamente.

Tabla 3

Coeficientes de correlación entre los índices calculados con la introducción, el cuerpo y los periódicos

	Positividad	*Negatividad*	*Negatividad neta*	***ISEF***	$ISEF_W$
Introducción frente a cuerpo	0.52***	0.90***	0.91***	0.90***	0.73***
Periódicos frente a introducción	0.26	0.50**	0.62***	0.66***	0.59***
Periódicos frente a cuerpo	0.21	0.43**	0.52**	0.61***	0.44**

Nota: En cada columna se muestran las correlaciones (coeficientes de Pearson) entre un mismo índice calculado para dos textos distintos. Los índices empleados son: *Positividad*, *Negatividad*, *Negatividad neta*, *ISEF* e $ISEF_W$ (ponderado por el algoritmo TF-IDF), correspondientes con las ecuaciones [1], [2], [3], [4] y [5], respectivamente. Los asteriscos *, ** y *** representan la significatividad del coeficiente a los niveles de confianza del 5 %, 1 % y 0,1 %, respectivamente.

Tabla 4

Regresiones entre los índices y distintos indicadores financieros

Indicador	*ISEF Introducción*		*ISEF Cuerpo*	
	Contemporánea	*Retardada (dos semestres)*	*Contemporánea*	*Retardada (dos semestres)*
Variación del PIB	-0.38***	-0.15	-0.30***	-0.14**
	(0.11)	(0.13)	(0.06)	(0.08)
Tasa de desempleo	0.02	0.02	0.02**	0.01
	(0.02)	(0.02)	(0.01)	(0.01)
Tipo de interés a corto plazo	-0.07	-0.04	-0.03	-0.01
	(0.07)	(0.08)	(0.03)	(0.04)
Tipo de interés virtual	-0.05**	-0.04	-0.02	-0.01
	(0.03)	(0.03)	(0.02)	(0.02)
Brecha de crédito sobre PIB	-0.004*	-0.004	-0.002	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)
Servicio de la deuda privada no financiera	0.01	0.04	0.02	0.04**
	(0.03)	(0.04)	(0.01)	(0.02)
Variación precio de la vivienda	-0.02***	-0.02	-0.02***	-0.01**
	(0.01)	(0.01)	(0.00)	(0.01)
Rentabilidad por dividendo	0.14***	0.11*	0.10***	0.08**
	(0.05)	(0.07)	(0.03)	(0.04)
Ratio precio-valor contable	-0.69***	-0.42*	-0.41***	-0.24*
	(0.14)	(0.27)	(0.08)	(0.15)
Prima de riesgo-bancos	0.26***	0.15*	0.19***	0.11**
	(0.06)	(0.10)	(0.02)	(0.05)
Volatilidad USD/EUR	0.03	0.01	0.02**	0.01
	(0.02)	(0.03)	(0.01)	(0.01)
Volatilidad IBEX-35	0.03***	0.01	0.02***	0.01*
	(0.01)	(0.01)	(0.01)	(0.01)

Nota: Se muestran los coeficientes de la regresión por MCO del índice *ISEF*, calculado para las introducciones y los cuerpos, con respecto a distintos indicadores. Las variables consideradas son: la variación trimestral del PIB para España (%); la tasa de desempleo en España (%); el rendimiento del tipo de interés a corto plazo (euríbor a tres meses, %); el tipo de interés a corto plazo virtual (%) calculado según Krippner (2015); la brecha de crédito sobre PIB (%), definida como la diferencia entre el cociente crédito/PIB y su tendencia de largo plazo; la ratio del servicio de la deuda privada no financiera (%), calculada por los costes del servicio de la deuda (pagos por intereses y amortizaciones de principal) como proporción de los ingresos; la variación del precio de la vivienda (%), medida como los cambios logarítmicos en el último año del índice del BIS que mide la evolución de los precios reales de la vivienda para España; la rentabilidad por dividendo (%) para el IBEX-35; la ratio valor en precio sobre valor contable para una entidad bancaria española representativa (Banco Santander); la prima de riesgo de crédito del sector bancario español, aproximada por el logaritmo del diferencial del CDS del Banco Santander; la volatilidad de la divisa (%), medida como la volatilidad implícita de las opciones *at-the-money* a un mes sobre el tipo de cambio USD/EUR; y la volatilidad realizada (%) trimestral (a 90 días) del índice de acciones del IBEX-35. Los errores estándar para cada coeficiente se muestran entre paréntesis y están calculados con el estimador de Newey y West (1987) y con el factor T/df de corrección para muestras pequeñas, donde T es el número de observaciones y df son los grados de libertad. Los asteriscos *, ** y *** representan la significatividad del coeficiente a los niveles de confianza usuales: 10 %, 5 % y 1 %, respectivamente. Las fuentes de los datos son Bloomberg y la página de estadísticas del Banco de Pagos Internacionales.

Figuras

Figura 1. Diagrama del trámite para la obtención de los índices de sentimiento del *Informe de Estabilidad Financiera*. Este se procesa inicialmente con la aplicación Nuance Power PDF Advanced 3.0 para obtener documentos editables en formato Word. A partir de estos documentos se seleccionan manualmente la introducción y el cuerpo, y se copia la información en archivos de texto. Estos procesados mediante herramientas de programación. Por un lado, para extraer ejemplos que han de ser usados en la herramienta de anotación y, por otro, y una vez que se dispone del diccionario de tonalidad, para calcular los índices y realizar las visualizaciones apropiadas.

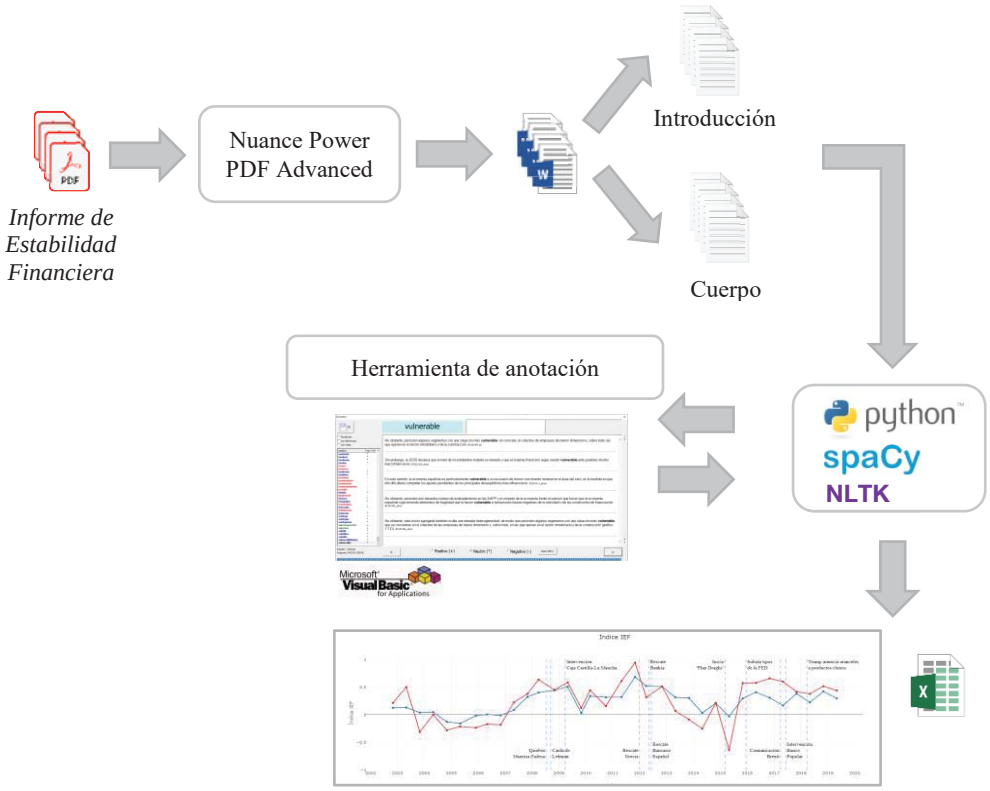


Figura 2. Fases del proceso de anotación. Se parte de un listado de palabras seleccionado y anotado inicialmente por un anotador (anotador de referencia, a_0). En la fase 1, ese listado se divide en tres particiones (P_A, P_B y P_C), que son evaluadas por dos anotadores cada una. En caso de desacuerdos, los anotadores de cada partición revisan las anotaciones no coincidentes y modifican las anotaciones que consideren oportunas (fase 1b). Las palabras con anotaciones discrepantes tras la revisión (bloques sombreados en rojo) son evaluadas por los anotadores del resto de las particiones (fase 2). Con los resultados de la fase 2, en la fase 3 los desacuerdos se resuelven por sabiduría del grupo, considerando todas las anotaciones (siete anotadores distintos en total, a_0-a_6, siendo a_0 el inicial de referencia).

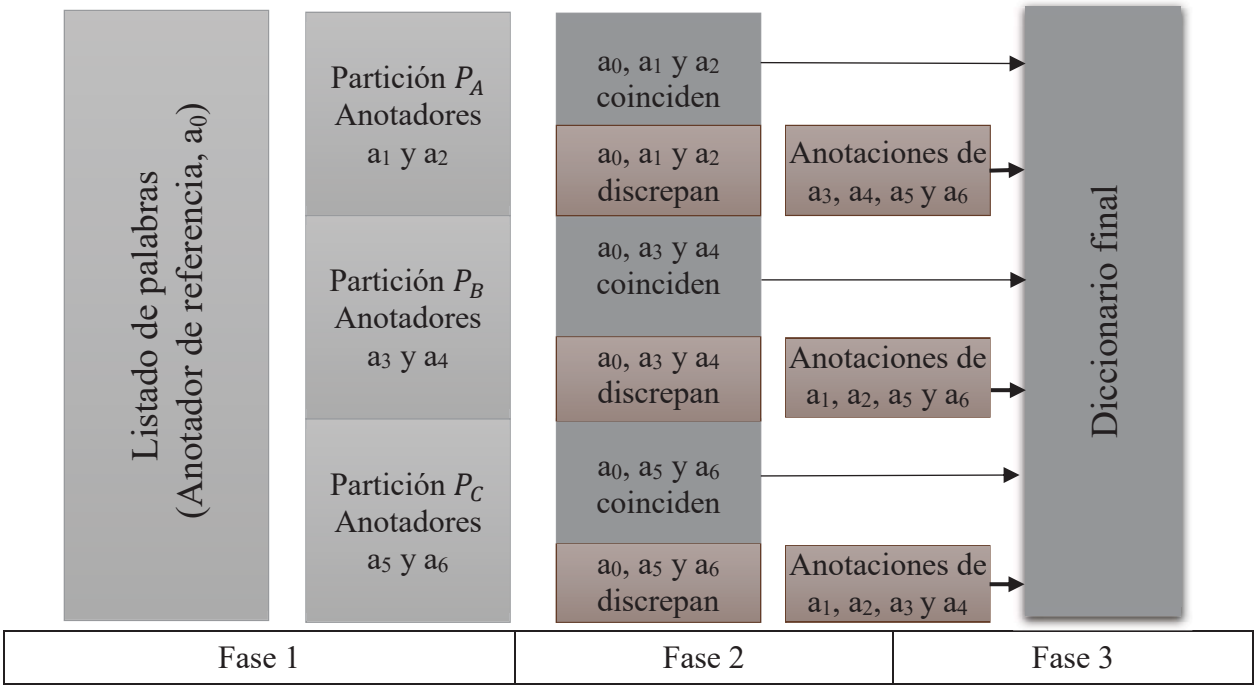


Figura 3. Nubes de palabras incluidas en el diccionario de sentimiento. El tamaño indica la frecuencia de aparición de la palabra en los textos de los informes analizados. El panel a) muestra todas las palabras evaluadas por los siete anotadores (3.706). En el diccionario final, esas palabras se han categorizado en: 189 palabras positivas, 376 negativas, y el resto, neutras (panel b).

Panel a): Total de palabras evaluadas (3.706)

Panel b): Palabras positivas, neutras y negativas

Positivas

Neutras

Negativas

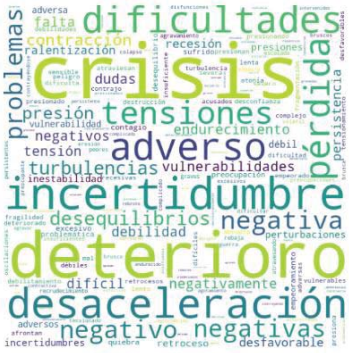


Figura 4. Distribuciones de distancias entre anotadores ($a_i - a_j$; i, j = 0,…,6) tras las diferentes fases del proceso de anotación. La distancia puede tomar los valores 0, 1 y 2. Para cada distribución se muestra entre paréntesis la medida AC_2 de Gwet correspondiente.

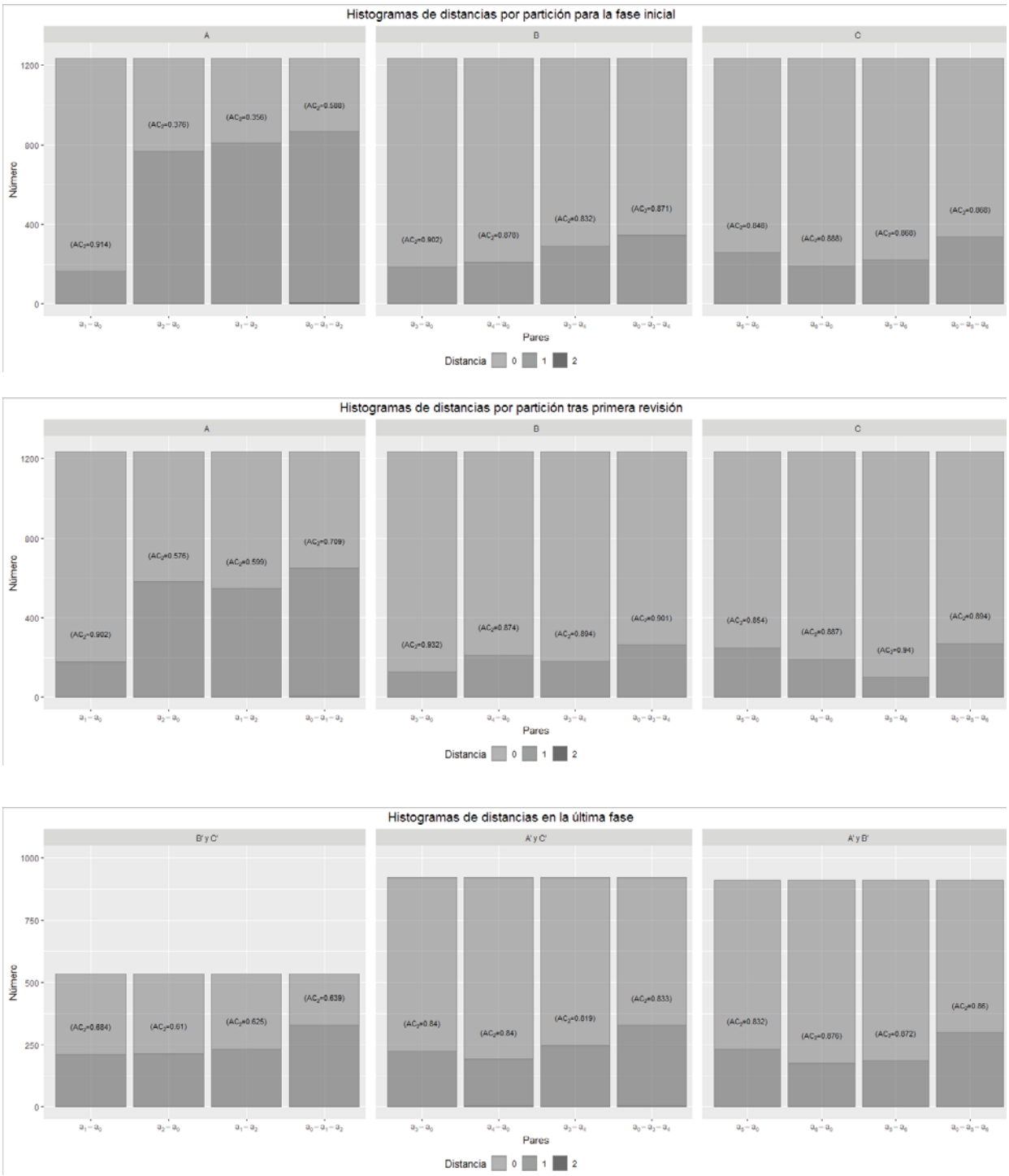
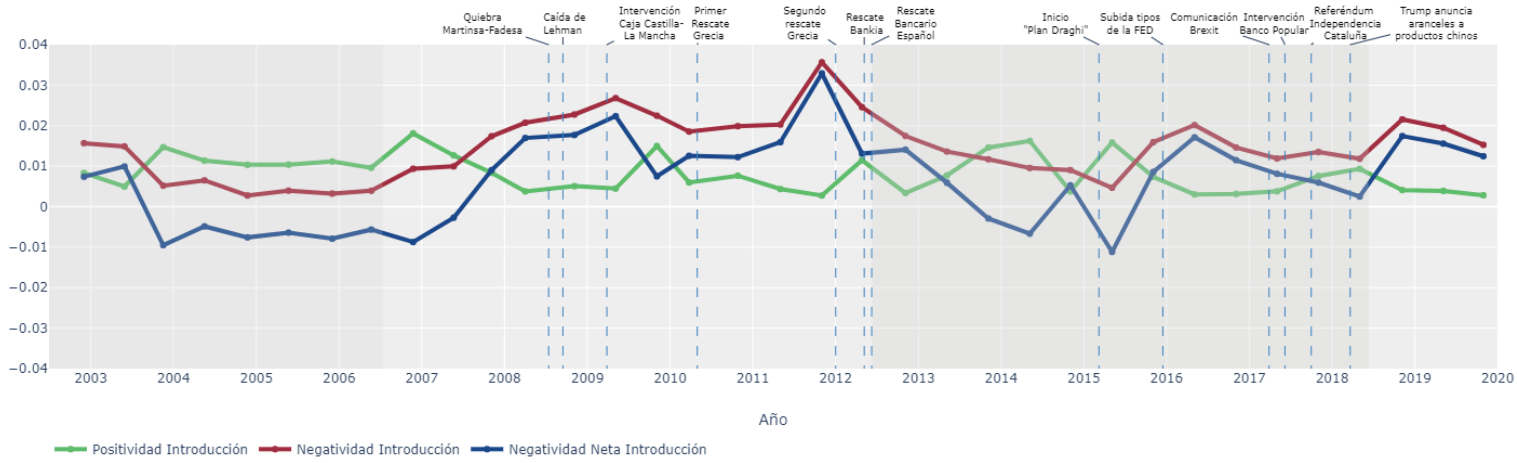


Figura 5. En el panel a) se muestran la *Positividad*, la *Negatividad* y la *Negatividad neta* para la introducción del *Informe de Estabilidad Financiera* desde otoño de 2002 hasta otoño de 2019. En el panel b) se presenta el índice de sentimiento del *Informe de Estabilidad Financiera (ISEF)*, calculado como la *Negatividad neta relativa* (ecuación [4]) para la introducción y el cuerpo de los informes de la muestra. Las franjas sombreadas en gris dividen los períodos de los cuatro gobernadores: J. Caruana (julio de 2000-julio de 2006), M. Á. Fernández Ordóñez (julio de 2006-junio de 2012), L. M. Linde (junio de 2012-junio de 2018) y P. Hernández de Cos (junio de 2018-fecha del análisis).

Panel a)



Panel b)

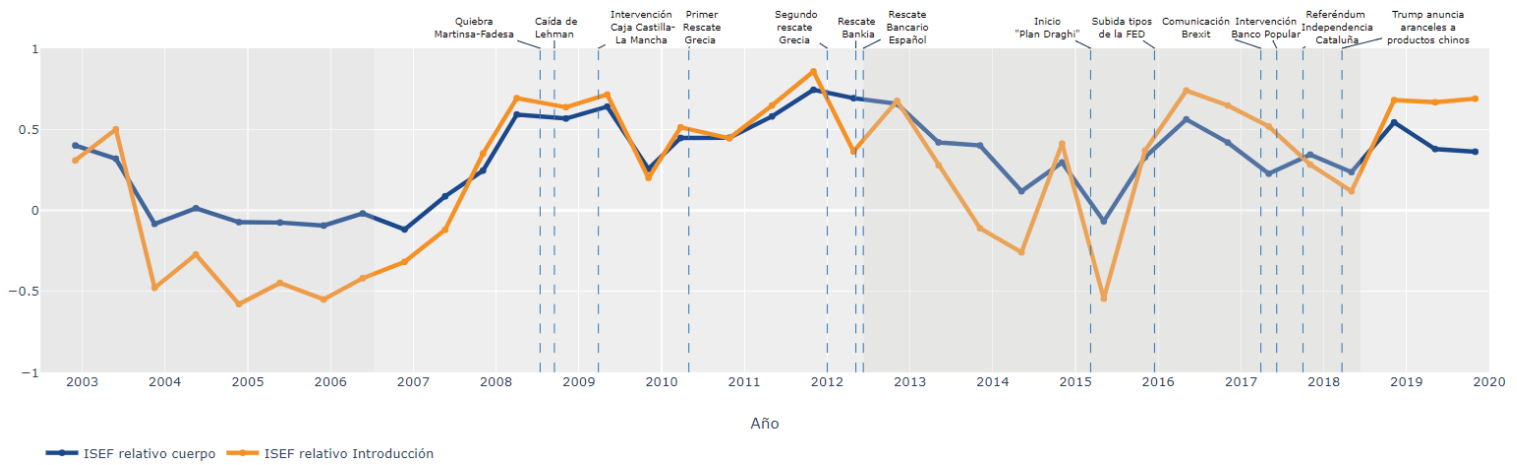
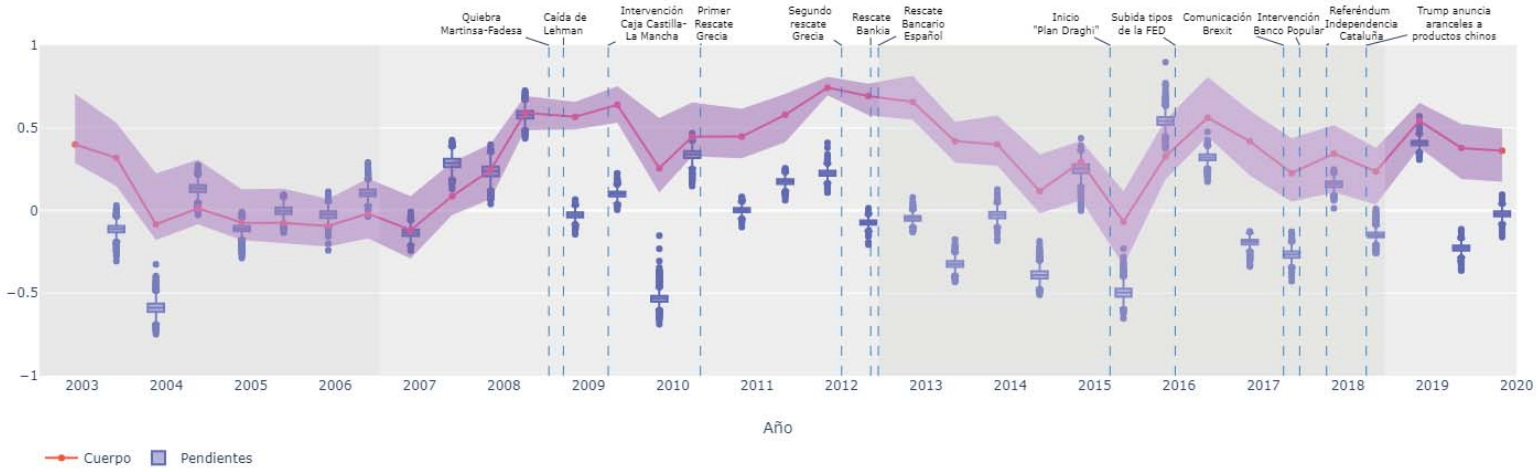


Figura 6. Análisis de consistencia del *ISEF* para el cuerpo de los informes desde otoño de 2002 hasta otoño de 2019. En el panel a) se muestran el índice (línea roja) y las bandas de confianza (sombreado en violeta) calculadas con 1.000 iteraciones, eliminando el 5% de las palabras del diccionario de manera aleatoria en cada iteración. En ambos gráficos se presentan también las distribuciones de la pendiente (variación del índice entre fechas consecutivas) correspondientes a las 1.000 iteraciones generadas. Estas distribuciones se muestran en forma de diagramas de caja (en azul).

Panel a)



Panel b)

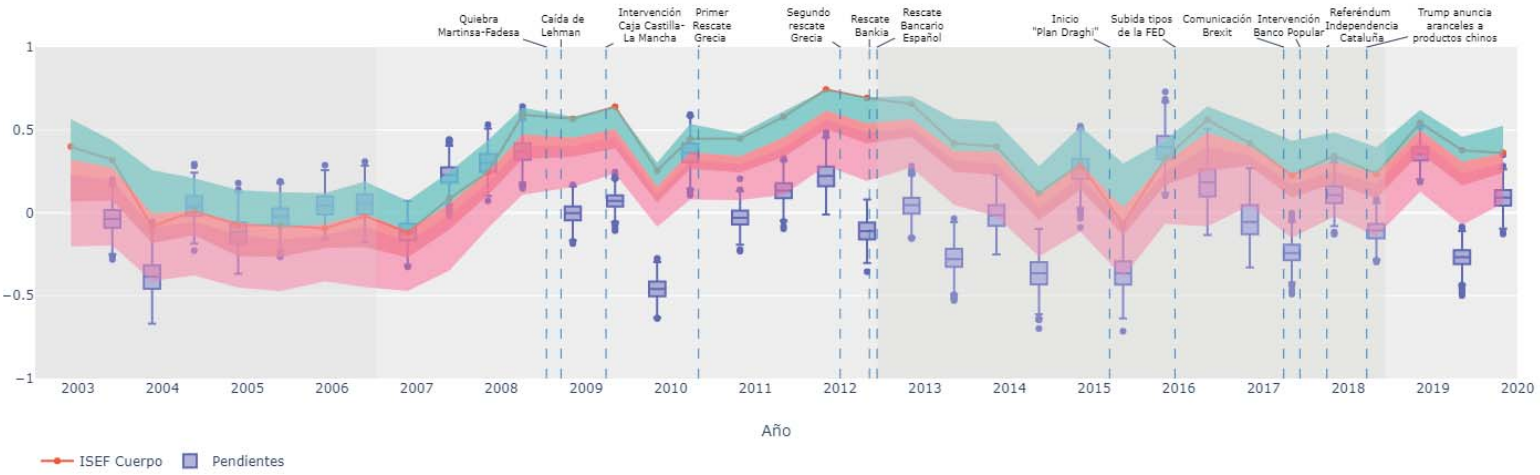


Figura 7. Nubes de palabras para las introducciones. El tamaño de las palabras dentro de la nube representa el peso del lema base dentro de cada Informe en función del algoritmo TF-IDF, lo que permite ver la evolución de los conceptos más específicos de cada Informe.

Figura 8. Cálculo del índice de sentimiento para las noticas de los periódicos relacionadas con los informes publicados desde otoño de 2002 hasta otoño de 2019. Se muestra el índice *ISEF* calculado para las noticias (línea verde) y se compara con la evolución de los índices calculados con los textos del Informe correspondientes a la introducción y al cuerpo (líneas marrón y azul).

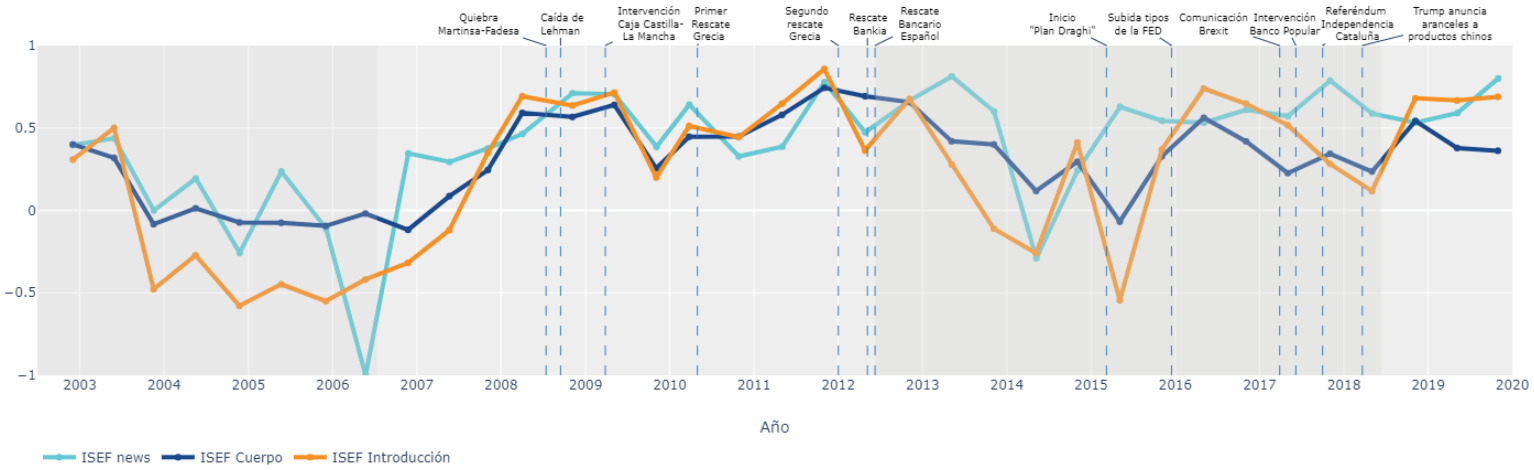
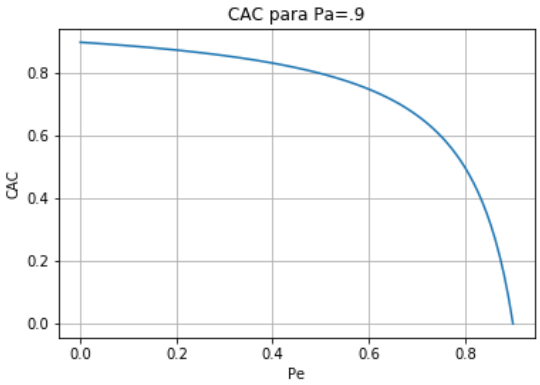


Figura 9. Nubes de palabras por tonalidad. Se muestran tres casos: otoño de 2006, primavera de 2015 y otoño de 2019 para la introducción (panel a) y para las noticias de los periódicos relacionadas con el Informe (panel b). El tamaño de las palabras muestra la frecuencia con la que aparece en el Informe. Las palabras con tonalidad positiva se muestran en verde; las negativas, en rojo.

Figura A.1. Variación de un coeficiente de tipo CAC dada una P_a = 0,9 en función de P_e, siendo $CAC = \frac{P_a - P_e}{1 - P_e}$.



PUBLICACIONES DEL BANCO DE ESPAÑA

DOCUMENTOS DE TRABAJO

1910 JAMES COSTAIN, ANTON NAKOV y BORJA PETIT: Monetary policy implications of state-dependent prices and wages.

1911 JAMES CLOYNE, CLODOMIRO FERREIRA, MAREN FROEMEL y PAOLO SURICO: Monetary policy, corporate finance and investment.

1912 CHRISTIAN CASTRO y JORGE E. GALÁN: Drivers of productivity in the Spanish banking sector: recent evidence.

1913 SUSANA PÁRRAGA RODRÍGUEZ: The effects of pension-related policies on household spending.

1914 MÁXIMO CAMACHO, MARÍA DOLORES GADEA y ANA GÓMEZ LOSCOS: A new approach to dating the reference cycle.

1915 LAURA HOSPIDO, LUC LAEVEN y ANA LAMO: The gender promotion gap: evidence from Central Banking.

1916 PABLO AGUILAR, STEPHAN FAHR, EDDIE GERBA y SAMUEL HURTADO: Quest for robust optimal macroprudential policy.

1917 CARMEN BROTO y MATÍAS LAMAS: Is market liquidity less resilient after the financial crisis? Evidence for US treasuries.

1918 LAURA HOSPIDO y CARLOS SANZ: Gender Gaps in the Evaluation of Research: Evidence from Submissions to Economics Conferences.

1919 SAKI BIGIO, GALO NUÑO y JUAN PASSADORE: A framework for debt-maturity management.

1920 LUIS J. ÁLVAREZ, MARÍA DOLORES GADEA y ANA GÓMEZ-LOSCOS: Inflation interdependence in advanced economies.

1921 DIEGO BODAS, JUAN R. GARCÍA LÓPEZ, JUAN MURILLO ARIAS, MATÍAS J. PACCE, TOMASA RODRIGO LÓPEZ, JUAN DE DIOS ROMERO PALOP, PEP RUIZ DE AGUIRRE, CAMILO A. ULLOA y HERIBERT VALERO LAPAZ: Measuring retail trade using card transactional data.

1922 MARIO ALLOZA y CARLOS SANZ: Jobs multipliers: evidence from a large fiscal stimulus in Spain.

1923 KATARZYNA BUDNIK, MASSIMILIANO AFFINITO, GAIA BARBIC, SAIFFEDINE BEN HADJ, ÉDOUARD CHRÉTIEN, HANS DEWACHTER, CLARA ISABEL GONZÁLEZ, JENNY HU, LAURI JANTUNEN, RAMONA JIMBOREAN, OTSO MANNINEN, RICARDO MARTINHO, JAVIER MENCÍA, ELENA MOUSARRI, LAURYNAS NARUŠEVIČIUS, GIULIO NICOLETTI, MICHAEL O'GRADY, SELCUK OZSAHIN, ANA REGINA PEREIRA, JAIRO RIVERA-ROZO, CONSTANTINOS TRIKOUPIS, FABRIZIO VENDITTI y SOFÍA VELASCO: The benefits and costs of adjusting bank capitalisation: evidence from Euro Area countries.

1924 MIGUEL ALMUNIA y DAVID LÓPEZ-RODRÍGUEZ: The elasticity of taxable income in Spain: 1999-2014.

1925 DANILO LEIVA-LEON y LORENZO DUCTOR: Fluctuations in Global macro volatility.

1926 JEF BOECKX, MAARTEN DOSSCHE, ALESSANDRO GALESI, BORIS HOFMANN y GERT PEERSMAN: Do SVARs with sign restrictions not identify unconventional monetary policy shocks?

1927 DANIEL DEJUÁN and JUAN S. MORA-SANGUINETTI: Quality of enforcement and investment decisions. Firm-level evidence from Spain.

1928 MARIO IZQUIERDO, ENRIQUE MORAL-BENITO and ELVIRA PRADES: Propagation of sector-specific shocks within Spain and other countries.

1929 MIGUEL CASARES, LUCA DEIDDA and JOSÉ E. GALDÓN-SÁNCHEZ: On financial frictions and firm market power.

1930 MICHAEL FUNKE, DANILO LEIVA-LEON y ANDREW TSANG: Mapping China's time-varying house price landscape.

1931 JORGE E. GALÁN y MATÍAS LAMAS: Beyond the LTV ratio: new macroprudential lessons from Spain.

1932 JACOPO TIMINI: Staying dry on Spanish wine: the rejection of the 1905 Spanish-Italian trade agreement.

1933 TERESA SASTRE y LAURA HERAS RECUERO: Domestic and foreign investment in advanced economies. The role of industry integration.

1934 DANILO LEIVA-LEON, JAIME MARTÍNEZ-MARTÍN y EVA ORTEGA: Exchange rate shocks and inflation comovement in the euro area.

1935 FEDERICO TAGLIATI: Child labor under cash and in-kind transfers: evidence from rural Mexico.

1936 ALBERTO FUERTES: External adjustment with a common currency: the case of the euro area.

1937 LAURA HERAS RECUERO y ROBERTO PASCUAL GONZÁLEZ: Economic growth, institutional quality and financial development in middle-income countries.

1938 SILVIA ALBRIZIO, SANGYUP CHOI, DAVIDE FURCERI y CHANSIK YOON: International Bank Lending Channel of Monetary Policy.

1939 MAR DELGADO-TÉLLEZ, ENRIQUE MORAL-BENITO y JAVIER J. PÉREZ: Outsourcing and public expenditure: an aggregate perspective with regional data.

1940 MYROSLAV PIDKUYKO: Heterogeneous spillovers of housing credit policy.
1941 LAURA ÁLVAREZ ROMÁN y MIGUEL GARCÍA-POSADA GÓMEZ: Modelling regional housing prices in Spain.
1942 STÉPHANE DÉES y ALESSANDRO GALESI: The Global Financial Cycle and US monetary policy in an interconnected world.
1943 ANDRÉS EROSA y BEATRIZ GONZÁLEZ: Taxation and the life cycle of firms.
1944 MARIO ALLOZA, JESÚS GONZALO y CARLOS SANZ: Dynamic effects of persistent shocks.
1945 PABLO DE ANDRÉS, RICARDO GIMENO y RUTH MATEOS DE CABO: The gender gap in bank credit access.
1946 IRMA ALONSO y LUIS MOLINA: The SHERLOC: an EWS-based index of vulnerability for emerging economies.
1947 GERGELY GANICS, BARBARA ROSSI y TATEVIK SEKHPOSYAN: From Fixed-event to Fixed-horizon Density Forecasts: Obtaining Measures of Multi-horizon Uncertainty from Survey Density Forecasts.
1948 GERGELY GANICS y FLORENS ODENDAHL: Bayesian VAR Forecasts, survey information and structural change in the Euro Area.
2001 JAVIER ANDRÉS, PABLO BURRIEL y WENYI SHEN: Debt sustainability and fIscal space in a heterogeneous Monetary Union: normal times vs the zero lower bound.
2002 JUAN S. MORA-SANGUINETTI y RICARDO PÉREZ-VALLS: ¿Cómo afecta la complejidad de la regulación a la demografía empresarial? Evidencia para España.
2003 ALEJANDRO BUESA, FRANCISCO JAVIER POBLACIÓN GARCÍA y JAVIER TARANCÓN: Measuring the procyclicality of impairment accounting regimes: a comparison between IFRS 9 and US GAAP.
2004 HENRIQUE S. BASSO y JUAN F. JIMENO: From secular stagnation to robocalypse? Implications of demographic and technological changes.
2005 LEONARDO GAMBACORTA, SERGIO MAYORDOMO y JOSÉ MARÍA SERENA: Dollar borrowing, firm-characteristics, and FX-hedged funding opportunities.
2006 IRMA ALONSO ÁLVAREZ, VIRGINIA DI NINO y FABRIZIO VENDITTI: Strategic interactions and price dynamics in the global oil market.
2007 JORGE E. GALÁN: Uncovering the impact of macroprudential policy on growth-at-risk.
2008 SVEN BLANK, MATHIAS HOFFMANN y MORITZ A. ROTH: Foreign direct investment and the equity home bias puzzle.
2009 AYMAN EL DAHRAWY SÁNCHEZ-ALBORNOZ y JACOPO TIMINI: Trade agreements and Latin American trade (creation and diversion) and welfare.
2010 ALFREDO GARCÍA-HIERNAUX, MARÍA T. GONZÁLEZ-PÉREZ y DAVID E. GUERRERO: Eurozone prices: a tale of convergence and divergence.
2011 ÁNGEL IVÁN MORENO BERNAL y CARLOS GONZÁLEZ PEDRAZ: Análisis de sentimiento del *Informe de Estabilidad Financiera.* (Existe una versión en inglés con el mismo número).

BANCO DE **ESPAÑA**
Eurosistema

Unidad de Servicios Generales I
Alcalá, 48 - 28014 Madrid
E-mail: publicaciones@bde.es
www.bde.es