MINERÍA DE TEXTOS PARA BANCOS CENTRALES

David Bholat
Stephen Hansen
Pedro Santos
Cheryl Schonhardt-Bailey

INTRODUCCIÓN

La *minería de textos* (también llamada *procesamiento de lenguaje natural*[1] o *lingüística computacional*) es un término amplio para una gama de herramientas de cómputo y técnicas estadísticas que cuantifican texto.[2] La minería de textos se parece a la lectura porque ambas actividades implican extraer significado de series de letras. Sin embargo, el análisis computacional y estadístico del texto difiere de la lectura en dos aspectos importantes. Primero, los enfoques permitidos por la computadora pueden procesar y resumir mucho más texto en comparación con el tiempo que una persona tiene para leer. Y segundo, tales técnicas pueden ser capaces de extraer significado de texto que escapa a los

1 El procesamiento de lenguaje natural es el procesamiento y análisis por computadora de los lenguajes naturales en el ser humano, en contraste con los lenguajes de programación, como Java.

2 También hay métodos asistidos por computadora para el análisis cualitativo de textos. Sin embargo, estos se salen del ámbito de aplicación de este manual. Sin embargo, en este vínculo se encuentra una introducción y contraste de algunas de las herramientas cualitativas de la minería de textos: <http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/support/choo>. Ver también Upshall (2014).

Traduce y publica el CEMLA con la debida autorización el Handbook, núm. 33, del Centre for Central Banking Studies, Banco de Inglaterra, 2015. Agradecemos a Ayeh Bandeh-Ahmadi, Aude Bicquelet, David Bradnum, Peter Eckley, Jo Gill, David Gregory, Sujit Kapadia, Tom Khabaza, Christopher Lovell, Rickard Nyman, Paul Ormerod, Paul Robinson, Robert Smith, David Tuckett, Iulian Udrea y Derek Vallès por sus comentarios. Las opiniones expresadas en este manual son las del autor y no necesariamente las de nuestros empleadores. <david.bholat@bankofengland.co.uk>, <stephen.hansen@upf.edu>, <pedro.santos@bankofengland.co.uk> y <c.m.schonhardt-bailey@lse.ac.uk>.

lectores humanos, quienes pasan por alto ciertos patrones porque no se apegan a sus creencias y expectativas previas.

Aunque son muchas sus aplicaciones en campos como las ciencias políticas y la mercadotecnia, la minería de textos históricamente ha sido menos utilizada como técnica en la economía. En particular, este es el caso de la investigación que realizan los bancos centrales internamente. Esto ha sido así por un par de razones al menos. La primera es que tal vez no sea obvio que el texto puede describirse y analizarse como datos cuantitativos.[3] Por tal motivo, los bancos centrales no están familiarizados con las herramientas y técnicas estadísticas que hacen posible lo anterior. La segunda es que, incluso si los banqueros centrales han oído hablar de la minería de textos, ya tienen acceso a otros datos cuantitativos a su disposición. Tal vez piensan que el costo de oportunidad y otros tipos de costos asociados con la transformación de textos en datos cuantitativos, así como tener que aprender nuevas herramientas y técnicas para analizarlos, superan los beneficios esperados.

Sin embargo, la minería de textos bien pudiera valer la inversión de los bancos centrales porque vuelve manejable una variedad de fuentes de datos importantes para evaluar la estabilidad monetaria y financiera, y que no pueden analizarse cuantitativamente de otras maneras. Los textos fundamentales para los bancos centrales son, entre otros, artículos de prensa, contratos financieros, redes sociales, inteligencia de supervisión y de mercado, así como informes escritos de distinta índole. Con las técnicas de minería de textos, podemos analizar un documento o una serie de documentos (un corpus). Un documento puede ser cierto discurso de un miembro del Comité de Política Monetaria (CPM) del Banco de Inglaterra (en lo sucesivo "el Banco"), una nota del personal o un informe de campo presentado por un agente.[4] El corpus correspondiente sería la totalidad de discursos de miembros del CPM, de notas del personal y de informes de campo.

Aunque los bancos centrales no han hecho mucho uso intencional de las técnicas de minería de textos, los banqueros centrales sí cosechan todos los días los beneficios de las aplicaciones de la minería de textos. Considérese, por ejemplo, la frecuencia con que los banqueros centrales (o cualquiera) usan Google para buscar información o utilizan el corrector de ortografía antes de publicar un documento. Considérense también los cortafuegos para aislar a los bancos centrales de los ataques cibernéticos, o la funcionalidad de búsqueda en las bases de datos de citas para recuperar las publicaciones académicas sobre cualquier tema. En este y otros casos, las técnicas de minería de textos funcionan en un segundo plano para ayudar a los bancos centrales a realizar su trabajo más eficientemente.

Otra finalidad de este documento es, entonces, demostrar el valor que los bancos centrales podrían obtener de una aplicación más concienzuda de las técnicas de minería de textos y explicar algunas de ellas utilizando ejemplos relevantes para esas instituciones. Este documento consta de dos partes principales. En la primera se explica cómo puede aplicarse la minería de textos a la investigación y la formulación de políticas que realiza el banco central, con base en ejemplos tomados de publicaciones. En la segunda parte del documento se proporcionan los primeros pasos para la minería de textos. Empezaremos con una explicación de cómo preparar los textos para su análisis. Luego comentamos varias técnicas de minería de textos, empezando por algunos métodos intuitivos, como las técnicas booleanas y las basadas en diccionario, antes de proceder a explicar los más

[3] Al texto a veces se le denomina *datos no estructurados,* en contraste con los datos estructurados (los números). Sin embargo, calificar a un texto como no estructurado es un poco engañoso. Un texto sí tiene estructura, siendo la gramática la más obvia, pero también patrones estructurales de distinto tipo que se extraen mediante las técnicas de minería de textos.

[4] Los agentes son empleados del Banco en todo el Reino Unido que proporcionan inteligencia sobre las condiciones de la economía local.

complejos, como el análisis semántico latente, la asignación de Dirichlet latente y la clasificación jerárquica descendente.

La minería de textos booleana y la basada en diccionarios, por un lado, y el análisis semántico latente, la asignación de Dirichlet latente y la clasificación jerárquica descendente, por el otro, son técnicas que se sirven de epistemologías diferentes, es decir, aproximaciones diferentes de la generación de conocimiento: la deducción y la inducción.[5] La deducción parte de una teoría general y luego comprueba si es válida utilizando ciertos conjuntos de datos. En contraste, la abducción se basa en algunos datos para intentar inferir la mejor explicación de cierto suceso, pero sin el propósito de obtener una explicación generalizable a otros casos.[6] La minería de textos booleana y la basada en diccionarios son técnicas deductivas porque parten de una lista predefinida de palabras, motivada por una teoría general respecto al porqué de su importancia. Las fortalezas de esta técnica son su sencillez y su escalabilidad. El código para su instrumentación por lo general tiene pocas líneas de longitud y puede aplicarse fácilmente a archivos de texto enormes. La debilidad de esta técnica es que se centra sólo en las palabras que el investigador consideró previamente como informativas; todas las demás son ignoradas. Por otro lado, el análisis semántico latente, la asignación de Dirichlet latente y la clasificación jerárquica descendente infieren patrones temáticos en cierto corpus sin afirmar que tales patrones se presentan en otros documentos. La principal fortaleza de estas técnicas es que analizan todas las palabras dentro de la muestra y arrojan resultados estadísticos más complejos. Su principal desventaja es la complejidad de la programación.

La minería de textos es un tema vasto. Por necesidad, hemos tenido que ser selectivos con las técnicas que tratamos en este documento. Principalmente nos concentramos en las técnicas de aprendizaje automático sin supervisión. *El aprendizaje automático sin supervisión implica* tomar observaciones *no clasificadas* y sacar a la luz los patrones ocultos que las estructuran en cierta manera significativa.[7] Estas técnicas pueden contrastarse con el *aprendizaje automático supervisado.* El aprendizaje automático supervisado inicia con las observaciones que *clasifica* el investigador para *entrenar* un algoritmo con *supervisión* humana, con el fin de que *aprenda* la correlación entre las clasificaciones asignadas por el investigador y las palabras características de los documentos en esas clasificaciones (Grimmer y Stewart, 2013). Aunque en la conclusión tratamos brevemente el aprendizaje automático supervisado, el punto focal de este documento son las técnicas de aprendizaje automático sin supervisión porque guardan relación con las prácticas cambiantes del Banco respecto a los *grandes volúmenes de datos* (Bholat, 2015; Haldane, 2015). En el documento, pondremos en cursivas los términos donde aparezcan definidos, tal como lo hemos hecho en esta introducción.

[5] Por supuesto, la deducción y la abducción son tipos ideales. En realidad, todos los métodos explicativos están mezclados. No obstante, creemos que esta clasificación ayuda a situar las distintas técnicas de minería de textos en términos de sus similitudes y sus diferencias.

[6] La *inducción* es una tercera epistemología que, como la abducción, parte de datos sin creencia previa; sin embargo, al igual que la deducción, busca producir afirmaciones teóricas generales.

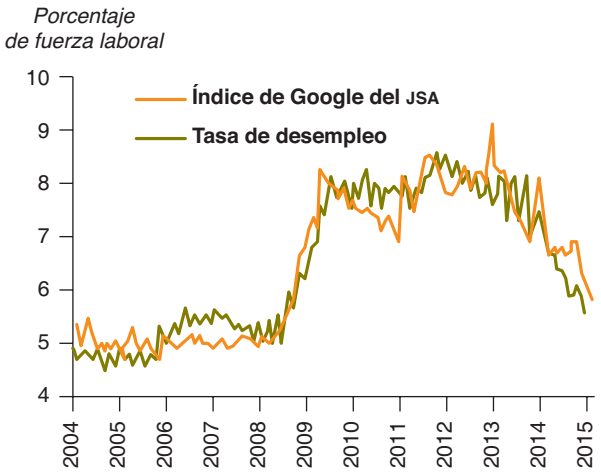
[7] Los resultados de algoritmos para el aprendizaje automático no supervisado pueden ingresarse en modelos econométricos para predecir alguna variable de interés, pero este es un método distinto a elegir intencionalmente las dimensiones del contenido con base en su capacidad predictiva.

1. EL TEXTO COMO DATOS PARA LA INVESTIGACIÓN QUE REALIZA EL BANCO CENTRAL

Con el fin de estimular un análisis de la minería de textos, empezaremos considerando los temas de investigación medulares que son temas de interés para los bancos centrales, utilizando como variable sustituta el estudio analítico One Bank Research Agenda (OBRA, Banco de Inglaterra, 2015). De hecho, el documento para discusión de OBRA identifica texto como fuente de datos cuyo potencial analítico no ha sido aprovechado del todo. Esto se debe al raudal de nuevos textos con datos disponibles mediante los registros de búsqueda en internet y las redes sociales. McLaren y Shanbhogue (2011) proporcionan un buen ejemplo de lo que puede hacerse. Utilizando datos de Google sobre volúmenes de búsquedas, descubrieron que tales datos proporcionan un seguimiento más oportuno de las variables económicas clave que las estadísticas oficiales. Por ejemplo, la gráfica 1 muestra que las búsquedas en Google del Jobseeker's Allowance (JSA)[8] prácticamente replican la cifra oficial de desempleo.

Gráfica 1

***GOOGLEANDO* EL MERCADO LABORAL**



Fuente: McLaren and Shanbhogue (2011).

Un tema que interesa a los bancos centrales es la medición del riesgo y de la incertidumbre en la economía y el sistema financiero. La investigación

[8] Subsidio por desempleo que paga el gobierno del Reino Unido.

de Nyman *et al.* (2015) fue una contribución reciente en este sentido. Nyman y sus coautores parten de una teoría general: la hipótesis de las finanzas emocionales. Conforme a esta, las personas se convencen de tomar posiciones en los mercados financieros creando narraciones respecto a los posibles resultados de sus acciones. Estas narraciones de convencimiento incorporan emociones como excitación por las ganancias esperadas y ansiedad por las posibles pérdidas. De acuerdo con Nyman y sus coautores, estas narraciones no las fabrican las personas aisladamente. Más bien, se construyen socialmente mediante interacciones como son las conversaciones entre personas. Es por medio de estas interacciones sociales que las narraciones se crean y se difunden, pudiendo así repercutir en el precio de los activos.

Para probar su hipótesis, analizaron tres fuentes de datos de texto: el comentario del mercado que proporciona diariamente el Banco (2000-2010), los informes de investigación de inversionistas (2010-2013) y el archivo de noticias de Reuters (1996-2014). El sentimiento se mide mediante el coeficiente de sentimiento construido en la ecuación 1 (ver Nyman *et al.,* 2015).

$$SI[T] = \frac{\left(|Excitment| - |Anxiety|\right)}{|T|}, \quad (1)$$

donde *SI[T]* es el coeficiente de sentimiento del documento *T*; |Excitement| es el número de palabras que expresan "excitación"; |Anxiety| es el número de palabras que expresan "ansiedad"; y |*T*| es el número total de palabras en el documento *T*.

El signo del coeficiente indica el sentimiento del mercado: alcista, si es positivo; bajista, si es negativo. El coeficiente luego se compara con los acontecimientos históricos y otros indicadores financieros.

Los autores también miden el consenso de la narración. En particular, su método consiste en agrupar artículos en grupos temáticos.[9] La incertidumbre en la distribución de los temas entonces actúa como variable sustituta de la incertidumbre. En otras palabras, la entropía reducida en la distribución de temas se utiliza como indicación de la concentración de los temas o consenso. En la gráfica 2 se muestra la serie de tiempo del índice de sentimiento y la medida del consenso. Los autores encontraron evidencia de comportamiento de manada (entropía disminuida) y mayor excitación antes de la crisis financiera reciente.

Una vez que se mide la incertidumbre en la economía, el objetivo de los bancos centrales es manejarla. Esta es una de las motivaciones principales de la reciente política del Banco de guía prospectiva (Carney, 2013), por la cual el Banco dirige las expectativas respecto a la orientación futura de la política mediante la comunicación de sus intenciones futuras y sus pronósticos oficiales. La minería de textos puede ayudar en este caso y otros parecidos al medir en qué grado los funcionarios del Banco están comunicando un mensaje congruente al resto del mundo.[10] Y la evaluación de la eficacia de las comunicaciones del Banco es un área de investigación identificada mediante el documento de discusión OBRA.

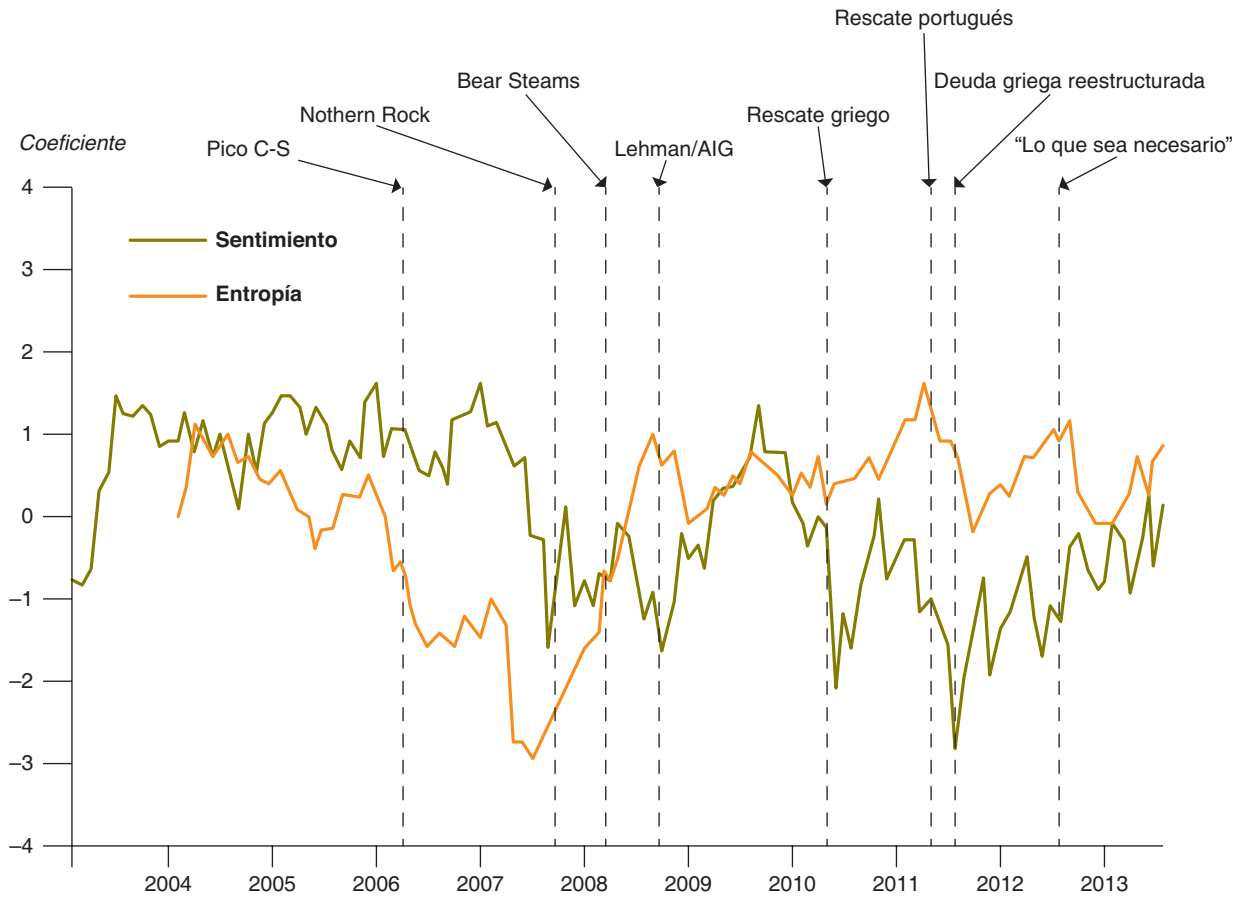
La gráfica 3 de Vallès y Schonhardt-Bailey (2015) ejemplifica el tipo de investigación que puede efectuarse. En la gráfica se muestra el contenido temático de los discursos del CPM y las minutas durante

[9] En particular, los autores utilizan un algoritmo de agrupamiento de X medias, que emplea criterios de información bayesianos para determinar el número óptimo de agrupamientos. Luego, usan la entropía de Shannon como medida de la distribución de temas. El consenso incrementado se mide mediante *1)* una reducción en el número de agrupamientos de temas, cuando el tamaño real de cada agrupamiento no cambia y *2)* el crecimiento relativo de un tema determinado, para un número fijo de agrupamientos de temas.

[10] Ver Rosa y Verga (2006), Blinder *et al.* (2008), Jansen y Haan (2010), Bennani y Farvaque (2014) para investigaciones similares sobre la congruencia de la comunicación de los bancos centrales. Sin embargo, no siempre es bueno que la comunicación sea congruente. Por ejemplo, Humpherys *et al.* (2011) crearon modelos para detectar declaraciones financieras fraudulentas en las comunicaciones de la dirección y encontraron evidencia de que las declaraciones fraudulentas tienen más probabilidad de contener menos diversidad léxica.

Gráfica 2

SENTIMIENTO Y ENTROPÍA EN LOS ARCHIVOS DE NOTICIAS REUTERS



Fuente: Nyman *et al.* (2015).

el último año de la gestión de Mervin King y el primero de la de Mark Carney.

Cada gráfica representa espacialmente co-ocurrencias, es decir, la convergencia y la divergencia de personas al hablar sobre ciertos temas. La proximidad espacial sugiere un mayor grado de coocurrencia. Por ejemplo, en ambas gráficas, la categoría temática *economía real* está muy relacionada con el CPM cuando habla como comité en sus minutas.

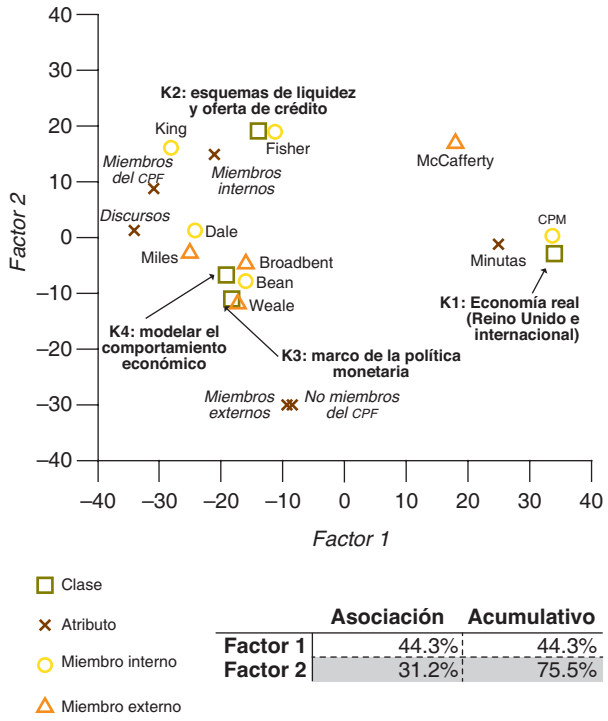
Puede identificarse una diferenciación evidente en los temas tratados por los miembros del CPM en sus discursos externos durante la era de Carney. Algunos miembros utilizaron sus discursos para comentar la guía prospectiva, pero otros no. Así, Vallès y Schonhardt-Bailey arrojaron luz respecto a dónde transmite un mensaje el comité como un todo, mientras que los miembros individuales están mandando mensajes más variados.[11]

[11] Otros estudios recientes que emplean la minería de textos para entender las comunicaciones de los bancos centrales son el análisis de Bulir *et al.* (2014) respecto a los informes de inflación de los bancos centrales y el que hace Siklos (2013) de las minutas de cinco bancos centrales, para demostrar la manera como cambió su tono después de la crisis financiera. También Nergues *et al.* (2014) obtienen mediciones de red para investigar cambios en el discurso del Banco

Gráfica 3

CONTENIDO TEMÁTICO DE LAS MINUTAS DEL CPM[a]

ANÁLISIS CORRESPONDIENTE DE VARIABLES DE CLASES Y PASIVOS. MINUTOS Y DISCURSOS DEL CPM, GOBERNACIÓN DE KING (JULIO 2012-JUNIO 2013)

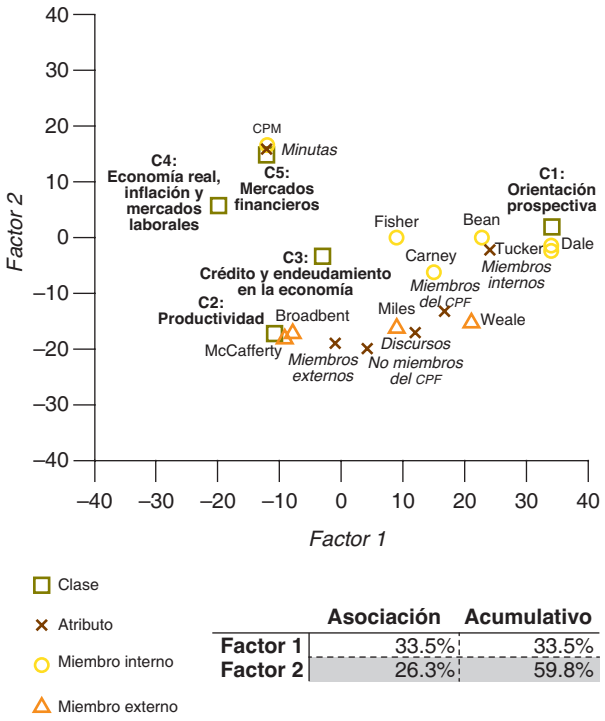


Clase
Atributo
Miembro interno
Miembro externo

	Asociación	Acumulativo
Factor 1	44.3%	44.3%
Factor 2	31.2%	75.5%

Clase 1 (40%) — Economía real (RU e internacional)
Clase 2 (26%) — Esquemas de liquidez y oferta de crédito
Clase 3 (18%) — Marco de la política monetaria
Clase 4 (16%) — Modelar el comportamiento económico

ANÁLISIS CORRESPONDIENTE DE VARIABLES DE TEMAS Y PASIVOS. MINUTOS Y DISCURSOS DEL CPM, GOBERNACIÓN DE CARNEY (JULIO 2013-JUNIO 2014)



Clase
Atributo
Miembro interno
Miembro externo

	Asociación	Acumulativo
Factor 1	33.5%	33.5%
Factor 2	26.3%	59.8%

Clase 1 (27%) — Orientación prospectiva
Clase 2 (27%) — Productividad
Clase 3 (10%) — Crédito y endeudamiento en la economía
Clase 4 (24%) — Economía real, inflación y mercados laborales
Clase 5 (12%) — Mercados financieros

[a] Estas gráficas representan las correlaciones entre temas y oradores en la gobernación de King y Carney. Las posiciones de los puntos y la distancia entre puntos reflejan el grado de co-ocurrencias. Los ejes identifican la cantidad máxima de asociación por factores, como se explica con más detalle en la seccción 2.
Fuente: Vallès y Schonhardt-Bailey (2015).

Los bancos centrales desean saber si están comunicando un mensaje congruente, pero también si las distintas políticas que proclaman se complementan o están en conflicto. De hecho, el documento de discusión OBRA señala la comprensión de las interacciones de la política monetaria, la macroprudencial y la microprudencial como un importante tema de investigación para el Banco. La minería de textos podría ser útil para entender estas interacciones. Aquí me inspiro en el ensayo reciente de William Li y coautores titulado "Law is Code" (Li *et al.*, 2015), quienes rastrean la complejidad creciente de la jurisprudencia estadounidense con el paso del tiempo mediante un análisis de todo el código legal de Estados Unidos (el US

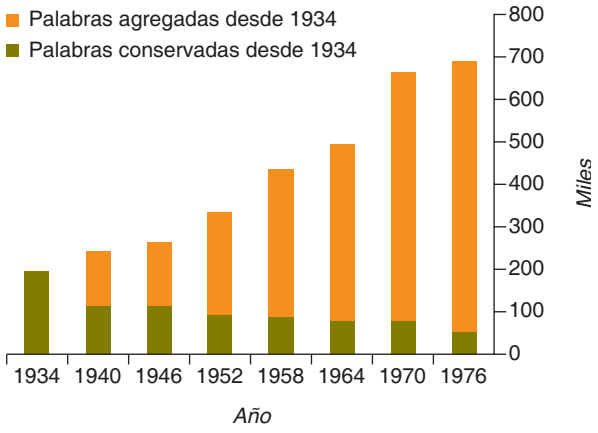
Central Europeo antes de la crisis financiera y después de ella.

Code) desde 1926 hasta la fecha. Los autores, que evocan el tono de Haldane (2012) en su crítica a la regulación financiera compleja, argumentan que la complejidad creciente de esa compilación de leyes dificulta su comprensión, tiene consecuencias negativas involuntarias y es un posible lastre para la productividad. Con el fin de consignar la complejidad creciente del Código de Estados Unidos, los autores produjeron varias medidas basadas en textos, entre otras:

1) Las medidas de tamaño y contenido del Código con el tiempo. Los autores interpretan que el aumento del Código, conforme a su número de palabras, significa que cada vez es más gravoso. Señalan que el tamaño bruto del Código y su tasa de aumento han crecido en los últimos decenios. Asimismo, dan seguimiento a los cambios en el contenido de secciones específicas del Código mediante la comparación de las palabras agregadas con las eliminadas con el paso del tiempo. Por ejemplo, la gráfica 4 muestra los cambios en el Título 12 (Bancos y Banca) del Código entre 1934 y 1976.

Gráfica 4

PALABRAS CONSERVADAS Y AGREGADAS AL TÍTULO 12
(número de palabras)



Fuente: Li *et al.* (2015).

2) Medidas de complejidad ciclomática. Para este estudio consideraremos que la complejidad ciclomática se refiere al conteo de los enunciados condicionales en el Código, es decir, las cláusulas que empiezan con la palabra "si", "salvo", "a menos que" y palabras elusivas similares. Los autores argumentan que los enunciados condicionales causan incertidumbre respecto a cuándo se aplican las leyes y, por lo tanto, su presencia indica una mayor complejidad.

3) Medidas de interrelación. Como en el sistema financiero, la interrelación en el derecho puede generar diseminaciones y consecuencias involuntarias (Gai *et al.*, 2011). Li y sus coautores miden la interrelación legal buscando las citas entre distintas partes del Código. Las secciones que tienen muchas referencias y se citan en otras partes son nodos *sistémicamente importantes* en la red jurídica. En el cuadro 1 se enuncian las secciones más importantes del Título 12. Este análisis identifica qué secciones del Código, si se reformaran, repercutirían mucho en otros aspectos de la legislación.

Son varias las aplicaciones obvias del método de Li y sus coautores para los asuntos que competen a los bancos centrales. Por ejemplo, podrían realizarse mediciones de la interrelación textual antes de cualquier cambio propuesto a la regulación o a las formas de informe regulatorio. Esto podría ayudar a cuantificar de antemano las posibles interacciones adversas de los cambios monetarios, macroprudenciales y microprudenciales. En términos más generales, las mediciones que dan seguimiento a los cambios en el tamaño, el contenido, la complejidad ciclomática y la interrelación de la regulación podrían calcularse en el Reino Unido y en otros entornos para someter a prueba la hipótesis de que la normatividad financiera se ha vuelto más compleja con el tiempo.

Cuadro 1

SECCIONES DEL TÍTULO 12 DEL CÓDIGO DE ESTADOS UNIDOS CON MAYOR INTERRELACIÓN

Número de sección	*Nombre*
1841	Definiciones de la Ley de Controladoras Bancarias
101	Abrogada (entrega de billetes en circulación)
1818	Terminación del Estatus como Institución de Depósito Asegurada
1709	Seguro de Hipotecas
1813	Definiciones de la Ley Federal de Seguros de Depósitos

Fuente: W. P. Li, P. Azar, D. Larochelle, P. Hill y A. W. Lo (2015), "Law is Code: A Software Engineering Approach to Analyzing the United States Code", ***Journal of Business & Technology Law***, vol. 10, núm. 2, pp. 297-374.

2. INTRODUCCIÓN A LAS TÉCNICAS DE MINERÍA DE TEXTOS

En la primera sección de este texto se planteó que la minería de textos es prometedora para los bancos centrales. En esta sección enumeramos los pasos principales de cualquier proyecto de minería de textos y proporcionamos una visión general de algunas técnicas específicas de la minería de textos. Todas las técnicas descritas en esta sección son una especie de *análisis del contenido* que resume de qué tratan los textos. Lo hacen contando el número de palabras de un corpus. La lógica subyacente es que la frecuencia de las palabras y su co-ocurrencia son buenos indicadores del tema o del sentimiento expresado en los textos.

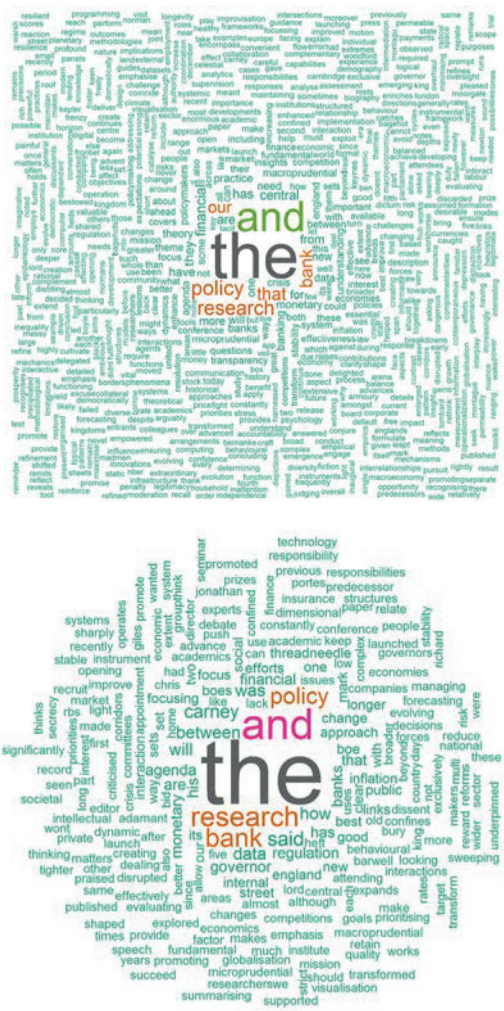
Para hacer más concreta esta intuición, considérense las *nubes de palabras* en la gráfica 5. La nube de palabras en la parte superior proviene del discurso inaugural del gobernador Carney en la conferencia sobre el OBRA organizada por el Banco (Carney, 2015); la de la parte inferior proviene del artículo sobre tal evento publicado en el *Financial Times* (Giles, 2015).

Como puede verse en la gráfica 5, los problemas ocurren con las palabras muy frecuentes y las poco frecuentes. Por ejemplo, las palabras *the* y *and* aparecen muy a menudo y, por ello, no nos ayudan a distinguir un documento de otro. Y al contrario, hay muchas palabras que sólo aparecen una vez.

La gráfica 6 simplifica las nubes de palabras de la gráfica 5 al mostrar únicamente aquellas que aparecen por lo menos dos veces en los textos y eliminar las palabras *the* y *and*. El contenido de ambas nubes es similar salvo en la mención que hizo Mark Carney de sí mismo en tercera persona, por lo que *Mr. Mark Carney* no aparece en la nube de la parte superior.

Gráfica 5

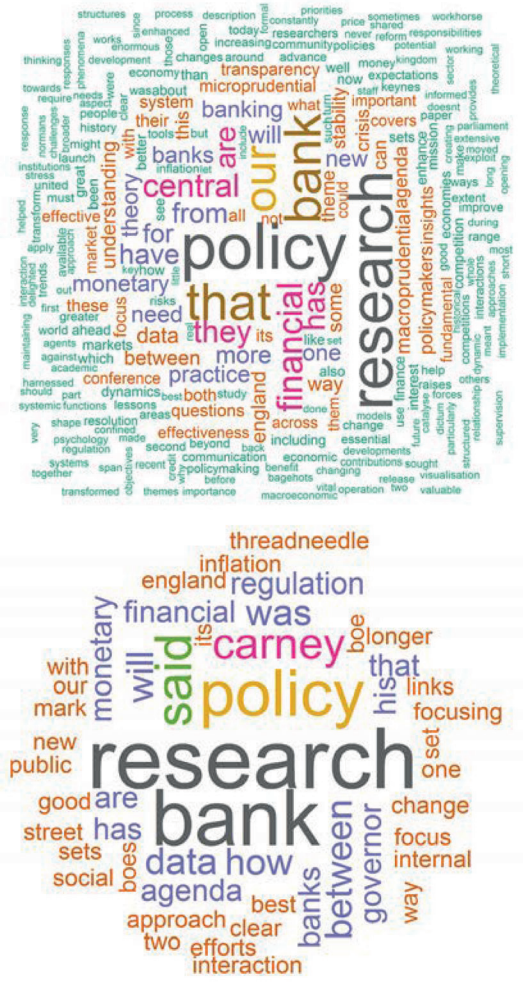
CONFERENCIA SOBRE EL DOCUMENTO OBRA DEL BANCO



Nota: Arriba, observaciones de Mark Carney; abajo, artículo del *Financial Times*.

Gráfica 6

NUBES DE PALABRAS SIMPLIFICADAS



Nota: Arriba, observaciones de Mark Carney; abajo, artículo del *Financial Times*.

Este ejemplo nos da dos lecciones sobre la minería de textos que estaremos retomando constantemente en este documento. La primera, que algunas palabras (visual y estadísticamente) relevantes pudieran no revelar nada significativo salvo lo obvio, por ejemplo, que el Sr. Carney no usó su propio nombre en su discurso. La segunda, que pudiera ser conveniente omitir ciertas palabras comunes, como *the* y *and* de nuestro marco de muestreo porque no agregan valor analítico. Dicho de otro modo, podemos eliminar palabras en la distribución de frecuencia sin perder comprensión sobre qué trata el corpus.

El cuadro 2 muestra estas nubes de palabras como una *matriz de término-documento (MTD)*, con filas que corresponden a términos únicos y

columnas que corresponden a cada documento. El elemento $X_{i,j}$ representa la cuenta del *i*-ésimo término en el documento *j*.

Cuadro 2

REPRESENTACIÓN MATRICIAL TÉRMINO-DOCUMENTO DE LAS NUBES DE PALABRAS

Término	Comentarios de Mark Carney sobre la agenda de investigación del Banco	Artículo en el ***Financial Times*** sobre la agenda de investigación del Banco de Inglaterra
policy	37	8
research	35	11
bank	28	10
that	28	3
our	25	2
financial	18	3
central	16	1
thinks	0	1
uses	0	1
wanted	0	1
won't	0	1

La tendencia a ser altamente dimensional y dispersa suele caracterizar a las matrices de término-documento como la del cuadro 2. Cualquier documento dado contendrá sólo un subconjunto de todos los términos únicos y las filas correspondientes a los términos no utilizados serán todas cero. Por ejemplo, la matriz completa de término-documento para los textos relacionados con el OBRA tiene 906 dimensiones (o términos únicos) y 42% del recuento de palabras es cero. Por lo anterior, la tarea fundamental consiste en extraer la información escasamente dimensional de los documentos que, por naturaleza, son altamente dimensionales. Esto equivale a una situación en la que un investigador tiene una base de datos con miles de *covariantes* y está intentando elegir cuál subconjunto de estas, o cuál resumen estadístico, debería incluirse en el análisis de regresión.

Por lo tanto, aunque la lógica de utilizar el recuento de palabras como indicador del contenido de los textos es simple, su puesta en práctica es más complicada. En las siguientes subsecciones nos centraremos en las técnicas para reducir la *dimensionalidad* (el número de palabras o variables) y la dispersión.

2.1 Preprocesamiento analítico

El primer paso en la minería de textos es definir el corpus en cuestión. Si el objetivo es generalizar, a partir del corpus, una población más grande de documentos, entonces se aplican las reglas estándar de muestreo. Los documentos deben ser representativos y seleccionarse aleatoriamente o mediante algún otro método de muestreo probabilístico. Un problema que a veces se presenta en esta etapa es la *duplicación,* que un corpus incluya el mismo documento varias veces. Por ejemplo, en las grandes bases de datos de periódicos, el mismo artículo aparece más de una vez, tal vez porque existen ediciones del mismo periódico en países diferentes. Tal duplicación puede distorsionar la inferencia al dotar de mayor representatividad a ciertos documentos. La eliminación de duplicaciones puede hacerse manualmente mediante la revisión del corpus para asegurarse de que cada documento sea una observación única, o mediante un algoritmo, como en Eckley (2015).

Una vez que el corpus se ha especificado, el siguiente paso es pasarlo a un formato analizable. Esto suele ser un proceso engorroso y tardado. Por ejemplo, los documentos en formato PDF pudieran convertirse y guardarse como archivos de texto (txt). Por otro lado, si los textos están sólo impresos, como sucede en muchos archivos, pudiera ser necesario digitalizarlos y convertirlos a archivos de texto mediante *programas para reconocimiento óptico de caracteres.*

Una vez que los textos se han convertido al formato apropiado, el siguiente paso consiste en segmentar el documento en *tokens*. Este paso implica representar el texto como una lista de palabras, números, signos de puntuación y posiblemente otros símbolos, como los de monedas o el de marca registrada.[12] Reiteramos que, en teoría, parece sencillo, pero no lo es tanto en la práctica.

Considérese este extracto, de una sola oración, del enunciado de la misión del Banco: *Promoting the good of the people of the United Kingdom* [Promover el bienestar del pueblo del Reino Unido]. El número de palabras en esta oración podría contarse de dos maneras por lo menos. Una es contar cada palabra discreta o token. Haciéndolo así, el extracto tiene diez tokens. Pero otra manera de contar el número de palabras en la oración es contando las palabras distintas. Así que, si bien el extracto incluye diez tokens, hay sólo siete palabras distintas, porque el token *the* aparece tres veces y *of* aparece dos veces. Podemos representar el número de tipos de palabras como un vector [1,3,1,2,1,1,1], como en el cuadro 3. En la minería de textos, las representaciones vectoriales de texto se denominan *representaciones de bolsa de palabras.*

Obsérvese que la representación vectorial en el cuadro 3 trata a *United* y *Kingdom* como palabras independientes. Sin embargo, otra representación, que supuestamente refleja mejor el significado o *semántica* de la oración, pudiera tratar ambos tokens como un único concepto, es decir, United Kingdom.

Escribir un algoritmo para dividir en tokens el texto de manera que siempre transmita el significado correcto es difícil. Por ejemplo, si un algoritmo representara cada caso en el que *United* está seguido de la palabra *Kingdom* como una aparición del término *United Kingdom*, entonces el algoritmo trataría incorrectamente "united" y "kingdom" en la siguiente oración como un token en vez de dos: *The marriage of Isabella of Castile to Ferdinand of Aragon created a united kingdom in Spain* [El matrimonio de Isabel de Castilla y Fernando de Aragón creó un reino unido en España].

Además de ejemplificar lo difícil que es segmentar en tokens el texto para que el recuento de palabras transmita significado con precisión, la oración anterior también muestra que un recuento

Cuadro 3

UNA REPRESENTACIÓN VECTORIAL DE PALABRAS

Término	*Documento: Enunciado de la misión del Banco (extracto)*
promoting	1
the	3
good	1
of	2
people	1
United	1
Kingdom	1

Cuadro 4

UNA REPRESENTACIÓN VECTORIAL DE TOKENS

Término	*Documento: Enunciado de la misión del Banco (extracto)*
Promoting	1
the	3
good	1
of	2
people	1
United Kingdom	1

[12] Una manera de segmentar en *tokens* es mediante una función de expresión regular que permita a los investigadores buscar patrones en el texto y dividir el texto con base en tales patrones. Por ejemplo, una búsqueda de las palabras *Bank of England* pudiera utilizar una expresión estándar que busque una palabra que comience con la letra *B* seguida de tres caracteres alfabéticos, un espacio, la palabra *of*, un espacio y una palabra que empieza con la letra *E*.

de palabras desajustado pudiera ser un mal indicador del contenido distintivo de un documento.

En el ejemplo, la palabra *the* es la moda. Así que, como dijimos al principio de esta sección, un aspecto fundamental de la minería consiste en reducir la dimensionalidad de las representaciones de bolsa de palabras para eliminar el *ruido* y centrarse en el contenido distintivo de los documentos. Hay varias técnicas para lidiar con las palabras que son irrelevantes para el contenido del corpus. Mencionamos algunas a continuación.

1) *Eliminar los signos de puntuación y las palabras raras.* Palabras como *el*, *de*, *que* y similares son muy comunes, pero ayudan poco a distinguir el contenido de un documento del contenido de otro. Así que tales palabras vacías, como se les conoce, con frecuencia se descartan de la muestra.[13] Por ejemplo, si descartáramos las palabas *of* y *the*, entonces el extracto del enunciado de la misión del Banco de Inglaterra quedaría representado de la siguiente manera:

Tabla 5

REPRESENTACIÓN VECTORIAL UTILIZANDO PALABRAS VACÍAS

Término	*Documento: Enunciado de la misión del Banco (extracto)*
Promoting	1
good	1
people	1
United Kingdom	1

2) *Representar las palabras con su raíz lingüística común.* Otro procedimiento para reducir la dimensionalidad es la *lematización*. La lematización utiliza el etiquetado de partes del discurso para determinar la categoría gramatical (parte del discurso) a la que pertenece cada palabra (sustantivo, pronombre, verbo, etc.) y convertirla a su forma base. Por ejemplo, *pienso*, etiquetado como verbo, se convertiría en *pensar*, pero no si fuera etiquetado como sustantivo. El etiquetado de partes del discurso es difícil porque las palabras suelen pertenecer a más de una categoría. Por ejemplo, la palabra *libro* es tanto un sustantivo como un verbo conjugado. Cuando una palabra recibe un etiquetado múltiple, pueden utilizarse algoritmos sintácticos para determinar el etiquetado correcto con base en las palabras vecinas. Considérese la oración: *Fed increases interest rate* [La Fed sube la tasa de interés] (Jurafsky y Manning, 2012). En inglés cada una de estas palabras puede ser un sustantivo o un verbo. Por ejemplo, la palabra *Fed* puede referirse a la Reserva Federal o al pasado del verbo *feed*. Un algoritmo sintáctico podría desambiguar cada uno de estos tokens mediante referencia a otro, hasta que a cada token se le haya asignado una categoría gramatical.

3) *Extracción de raíces.* En la práctica, muchos mineros de textos simplemente extraen las raíces de las palabras porque este procedimiento tiende a ser más rápido y más sencillo que la lematización.[14] La extracción de raíces implica cortar los afijos y contar sólo las raíces. Por ejemplo, la palabra "bancario" contiene la raíz "banca" y el afijo "rio". Por lo tanto, "bancario" y "banca", una vez extraída su raíz, serían tratadas como dos apariciones del mismo token.[15]

[13] Una lista de palabras vacías se puede conseguir en: <http://snowball.tartarus.org/algorithms/english/stop.txt>.

[14] El algoritmo de Porter es popular para los textos en inglés.

[15] El resultado después de la extracción de raíces tal vez no sea una palabra que ocurra naturalmente y pudiera carecer de significado interpretativo. Por ejemplo, la raíz de *inflation* es *inflat*. Como ya dijimos, la lematización proporciona un resultado más refinado, es decir, por lo general arroja la forma de diccionario de una palabra.

4) Conversión a minúsculas. La conversión a minúsculas implica pasar todos los tokens alfabéticos a minúscula. Aunque en algunos casos esto podría oscurecer el significado de algunos nombres propios (por ejemplo, un acrónimo como *US* para *United States* podría convertirse erróneamente en el pronombre *us,* nosotros en inglés), este procedimiento casi siempre considera irrelevante que la palabra se encuentre al principio de una oración. No obstante, hay casos en los que la conversión a minúsculas puede resultar engañosa. Considérese la oración: *The Bank of England is the main regulator of XYZ bank* [El Banco de Inglaterra es el principal regulador del banco XYZ]. La conversión a minúsculas trataría *Bank* y *bank* como apariciones del mismo token cuando, de hecho, se trata de entidades jurídicas diferentes.

2.2 Técnicas booleanas

Hasta ahora hemos considerado los pasos de preprocesamiento comunes en cualquier investigación de minería de textos. Ahora toca el turno a los distintos tipos de técnicas de minería de textos.

Tal vez el método de minería de textos más sencillo es realizar una búsqueda booleana de uno o más términos. *Las técnicas de búsqueda booleana* combinan términos individuales o frases con operadores lógicos como AND (y), OR (o) y NOT (no) para formar expresiones de búsqueda. Después, a cada documento se le asigna un 1 o un 0 dependiendo de si la expresión es verdadera (= 1) o falsa (= 0). Dado que la búsqueda booleana ya forma parte de los principales buscadores de internet, tiene la gran ventaja de que los investigadores no tienen que ingresar en los textos sin procesar si estos ya han sido indizados. Por ejemplo, Google y Yahoo arrojan un número aproximado de sitios web que cumplen con los criterios de la búsqueda booleana.

Un ejemplo de investigación relevante para los bancos centrales que utilizan técnicas booleanas es la realizada por Baker, Bloom y Davis (2013), quienes calcularon la incertidumbre asociada a la política económica analizando los principales periódicos de Estados Unidos y Europa. Contaron diariamente cuántos de los artículos publicados a partir de 1985 contenían palabras relacionadas con la incertidumbre y la economía. Específicamente, los autores computaron el número de artículos diarios que cumplían con los siguientes criterios de búsqueda:

1) el artículo contiene *incierto* OR *incertidumbre,* AND

2) el artículo contiene *económica* OR *economía,* AND

3) el artículo contiene *congreso* OR *déficit* OR *reserva federal* OR *legislación* OR *regulación* OR *casa blanca*.

Las series de tiempo de este *recuento normalizado* –conforme al número total de artículos de periódico cada mes– actúa entonces como variable sustituta de la incertidumbre respecto a la política.[16] Como se observa en la gráfica 7, su índice pudo consignar importantes sucesos políticos y de los mercados financieros.

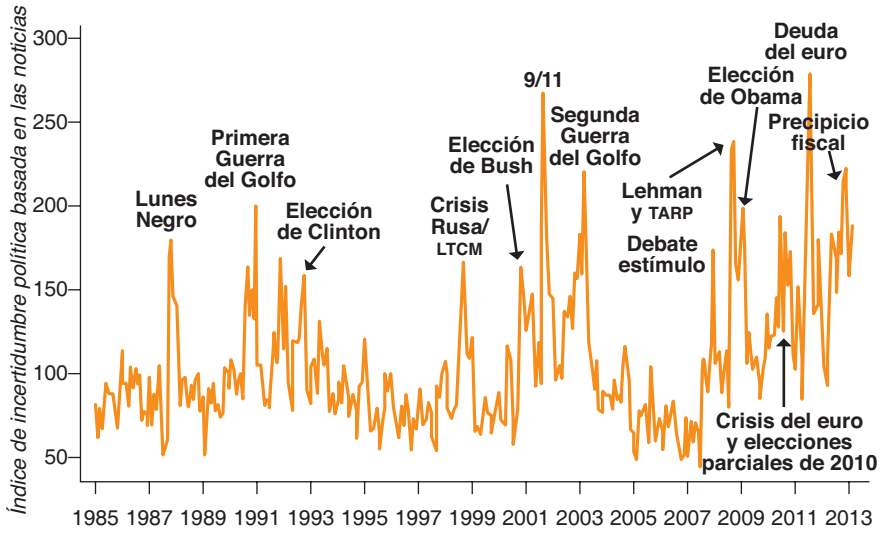
Otro ejemplo de minería booleana de textos es el trabajo reciente del Banco en la antesala del referendo por la independencia de Escocia. En ese entonces existía preocupación por las secuelas de una votación a favor, como por ejemplo, corridas contra los bancos escoceses (Banco de Inglaterra, 2014). Dadas las fluctuaciones en los resultados de las encuestas de opinión y el sentimiento público, el Banco monitoreó estos riesgos mediante un análisis del tráfico en Twitter. Se dio seguimiento a los tuits que contenían términos y combinaciones de términos relativos a corridas bancarias durante una semana completa, pero particularmente en la noche de la votación. Al final, el tráfico fue poco y resultó evidente el voto opuesto a la independencia.[17]

[16] Ver <http://www.policyuncertainty.com> para más detalles.

[17] En general, las redes sociales tienen propiedades que las hacen atractivas como fuente de datos. (O'Connor *et al.* 2010; Bollen *et al.*, 2011). En particular, los datos de las redes sociales se generan durante la noche, gracias a lo cual las autoridades pueden vigilar los riesgos después del cierre

Gráfica 7

ÍNDICE DE INCERTIDUMBRE POLÍTICA BASADA EN LAS NOTICIAS



Fuente: Baker, Bloom, and Davis (2013).

En conclusión, podría decirse que las mediciones booleanas del riesgo y la incertidumbre basadas en textos tienen cuatro puntos fuertes en comparación con las mediciones más usuales, como el índice VIX. Primero, al modificar los términos de búsqueda utilizados para los artículos, se puede consignar la incertidumbre más allá de los mercados financieros. Segundo, la fuente de incertidumbre suele enunciarse explícitamente en texto: "la incertidumbre causada por el referendo", por ejemplo. En contraste, con las medidas de incertidumbre basadas en el precio de los activos financieros, la fuente de la incertidumbre no siempre está clara. Tercero, si buscamos texto en los medios con fechas más antiguas, podemos obtener series de tiempo mucho más amplias que las basadas en el precio de las opciones. Por ejemplo, el VIX sólo está disponible a partir de finales de los años ochenta, mientras que Baker, Bloom y Davis remontaron su índice hasta 1900. Cuarto y último, podemos obtener mediciones multinacionales de la incertidumbre centrándonos en el texto publicado en los medios de distintos países, mientras que el VIX simplemente es una variable sustituta de la incertidumbre en los precios de las acciones en Estados Unidos, que pueden ser un mal indicador de la incertidumbre en otros países.

2.3 Técnicas de diccionario

Otra serie de estrategias para la minería de textos son las técnicas de diccionario, como las utilizadas por Nyman y coautores, tal como señalamos en la sección 1. Las técnicas de diccionario por lo general consisten de dos pasos:

1) Primero, definir una lista de palabras clave para consignar el contenido de interés.

2) Luego, representar cada documento en términos de la frecuencia (normalizada) de las palabras en el diccionario.

A diferencia de las técnicas booleanas, las técnicas de diccionario miden la intensidad del uso

de los mercados y antes de su apertura.

de las palabras, que pudiera ser una medida más adecuada del contenido del corpus. Sin embargo, la aplicación de las técnicas de diccionario implica que el investigador tenga acceso a los textos sin procesar.

Por ejemplo, permitamos que el diccionario D = {labour, wage, employ} [trabajo, salario, empleo] sea las raíces de diversas palabras relacionadas con los mercados laborales. Podemos entonces representar cada documento d como la participación s_d de palabras que están en el diccionario (ecuación 2).

Frecuencia normalizada de palabras en un documento

$$s_d = \frac{\begin{pmatrix}\text{Número de ocurrencias de } \textit{trabajo} + \\ \text{número de ocurrencias de } \textit{salario} + \\ \text{número de ocurrencias de } \textit{empleo}\end{pmatrix}}{\text{Total de palabras en el documento } d} \quad (2)$$

Los estudios publicados proporcionan varios ejemplos en los que se han empleado técnicas de diccionario para analizar textos sobre finanzas y economía. Por ejemplo, Tetlock (2007) escribió un artículo muy citado que utiliza diccionarios (es decir, léxicos) para medir el tono de la columna "Abreast of the Market" que se publica en el *Wall Street Journal*.[18] En particular, emplea los diccionarios Harvard IV – 4, cuyas listas de palabras reflejan muchas categorías, incluido el sentimiento positivo y negativo, dolor y placer, y rituales y procesos naturales, entre otros.[19] Tetlock luego cuenta el número de palabras en la columna de cada día entre 1984 y 1999. Su primer hallazgo es que la mayor parte de la variación en el recuento de palabras entre columnas refleja virajes entre el optimismo y el pesimismo. Su segundo hallazgo es que un aumento en las noticias negativas influye de manera estadísticamente significativa en los resultados del día siguiente. Después de Tetlock (2007), otros estudios han empleado la misma estrategia básica de contar palabras de interés en el contexto de los mercados financieros y correlacionarlas con el precio de los activos (por ejemplo, Aase, 2011).

Sin embargo, reiteramos que hay muchos casos en los que un mero recuento de palabras puede ser engañoso. Considérese la clase de minería de textos conocida como *análisis del sentimiento*, llamada también *minería de opinión* o *análisis de subjetividad.* La finalidad del análisis del sentimiento consiste en detectar el sentir expresado respecto a cierto objeto, o como se dice en el medio, un objetivo. El objetivo del sentimiento pudiera ser una persona, un suceso, una institución, un objeto, por nombrar unos cuantos. La manera más sencilla de analizar el sentimiento es la detección de polaridad, es decir, una clasificación binaria de sentimiento positivo y negativo.

Sin embargo, la medición precisa del sentimiento es complicada por varias sutilezas lingüísticas comunes como la negación, la ironía, la ambigüedad, las expresiones idiomáticas y los neologismos (Jurafsky y Manning, 2012). Considérese, por ejemplo, el siguiente párrafo:

> One would have expected building society X to have been an excellent institution. It had a top-notch CEO and world-renowned analysts. Its services were highly rated by retail clients and its operations were efficient. Yet it was a total failure. [*Uno pensaría que la sociedad de constructores X habría sido una institución excelente. Tenía un director general de primera y analistas de renombre mundial. Sus servicios recibían buena calificación de los clientes minoristas y sus operaciones eran eficientes. Aun así, fue un fracaso total.*]

Un mero recuento del total de las palabras positivas y negativas no transmitiría el sentimiento del párrafo anterior.

2.4 Ponderación de palabras

Un simple recuento pudiera no ser apropiado porque puede exagerar la importancia de un número

[18] Ver <http://www.wsj.com/news/types/abreast-of-the-market>.

[19] Ver <http://www.wjh.harvard.edu/~inquirer>.

pequeño de palabras muy frecuentes. Esto puede ser causa de problemas por dos razones:

1) La ley de Zipf, una observación científica de que la frecuencia de una palabra es inversamente proporcional a su jerarquía relativa en un corpus. Esto significa que una diferencia mínima en la jerarquía relativa de una palabra puede significar una gran diferencia en términos del recuento real de palabras. Así que, depender del recuento simple de palabras pudiera exagerar su importancia comparativa.

2) Si una palabra se utiliza en muchos documentos de un corpus, entonces su poder para discriminar entre dos documentos es menor que si sólo aparece en unos pocos documentos. No obstante, pudiera ser recomendable dar mayor peso a las palabras que aparecen en pocos documentos porque pudieran indicar diferencias reales en el contenido.

Para abordar estos problemas, una manera común de ponderación que se utiliza en la minería de textos es la *frecuencia de término-frecuencia inversa de documento (tf.idf)*, conforme a la ecuación 3.

$$3 \qquad tf.idf_{t,d} = \left(1+\log f_{t,d}\right)\cdot \log\left(\frac{D}{d_{ft}}\right), \tag{3}$$

donde D es número total de documentos en el corpus, df_t es el número de documentos en los que aparece el término *t* y $f_{t,d}$ es la frecuencia del término *t* en el documento *d.*

El primer factor en la ecuación 3 es la frecuencia del término *t* en el documento *d,* que otorga menor peso a las palabras que aparecen con más frecuencia. El segundo factor es la frecuencia inversa en el documento del término *t,* que asigna mayor peso a las palabras que aparecen con menos frecuencia.

Un ejemplo de la ponderación *tf.idf* aparece en Loughran y Mcdonald (2011). Su punto de partida es una crítica a los diccionarios Harvard IV-4 utilizados por Tetlock (2007). Los diccionarios de Tetlock contienen palabras como *impuesto*, *costo* y *pasivo* que, si bien transmiten un sentimiento negativo en un contexto general, tienen un tono más neutral en el contexto de los mercados financieros, pues describen prácticas contables cotidianas. Por lo anterior, Loughran y Mcdonald proponen un diccionario específico de las finanzas y demuestran que puede predecir mejor el rendimiento de los activos que los diccionarios genéricos.[20] Sin embargo, tras una ponderación *tg.idf*, el desempeño de los diccionarios genéricos mejora de manera notable.[21] Esto se debe a que palabras como *impuesto* aparecen en muchos documentos y, por lo tanto, tienen menos peso que otras palabras *realmente* negativas.

2.5 Modelos de espacio vectorial

Hasta ahora hemos considerado técnicas que identifican los principales temas *dentro de* los textos mediante una serie predefinida de palabras clave. Ahora consideraremos las técnicas para medir la similitud de los temas *entre* textos.

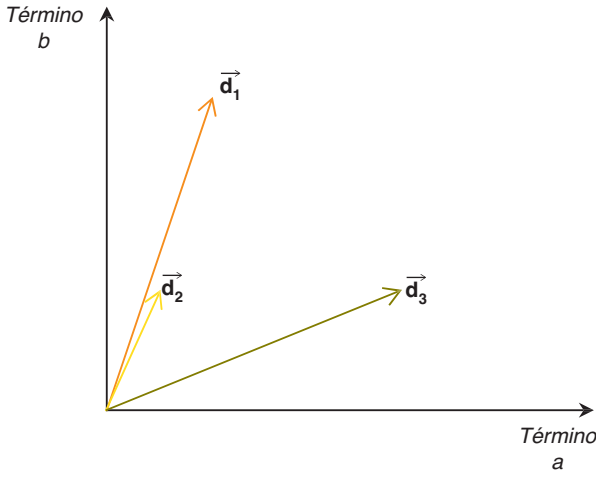
Una manera de medir la similitud de los documentos es utilizando la distancia euclidiana simple. Por ejemplo, Kloptchenko *et al*. (2004) utilizan la distancia euclidiana para encontrar agrupaciones de informes financieros. Sin embargo, como muestra la gráfica 8, esta medición de la distancia tiene limitaciones. La gráfica representa tres documentos hipotéticos, cada uno de los cuales contiene dos términos: *a* y *b*. Supóngase que los documentos 1 y 2 utilizan los términos *a* y *b* casi en la misma proporción. Sin embargo, debido a que el documento 1 pudiera ser mucho más largo que el documento 2, su distancia es bastante significativa. De hecho, el documento 3, que utiliza el término *a* en relación con el término *b* mucho más que el documento 2, se mediría como más parecido al documento 1, simplemente por su longitud similar.

[20] Disponible en <http://www3.nd.edu/~mcdonald/Word_Lists.html>.

[21] Aunque todavía encuentran que su diccionario específico de finanzas tiene un mayor poder explicativo.

Gráfica 8

ESTA GRÁFICA REPRESENTA TRES DOCUMENTOS, CADA UNO CONTIENE DOS TÉRMINOS *A* Y *B.*



Nota: documentos 1 y 2 tienen contenido muy similar aunque todavía se encuentran muy separados por la diferencia de longitud.

Este ejemplo muestra las distorsiones que pueden darse cuando se utiliza la distancia euclidiana para medir la distancia entre documentos. Una medida que evita estos problemas es *la similitud coseno,* que refleja el ángulo formado por dos vectores. Volviendo a la gráfica 8, podemos ver que el ángulo formado entre los documentos 1 y 2 es muy pequeño; apuntan a la misma dirección dado que utilizan los dos términos en proporciones casi idénticas. Sin embargo, debido a que las frecuencias de los términos difieren en los documentos 1 y 3, el ángulo es mayor. Si un vector del documento contuviera sólo el término *a* y el otro sólo el término *b,* los vectores serían *ortogonales.* Así que, midiendo el coseno del ángulo θ formado por dos documentos en el espacio vectorial se obtiene una medida de similitud independiente de la longitud de los documentos. La fórmula para computar la similitud coseno se proporciona en la ecuación 4.

4 $$\cos\theta = \frac{\vec{d}_1 \cdot \vec{d}_2}{\|\vec{d}_1\|\|\vec{d}_2\|},$$

donde $\cdot$ es el operador producto punto y $\|\vec{d}_t\|$ es la longitud del vector que representa al documento *i.*

Hoberg y Phillips (2010) proporcionan un ejemplo de modelación con espacio vectorial en economía. Tomaron declaraciones de productos de empresas de los formularios 10K presentados ante la Comisión de Valores y Bolsas de Estados Unidos y computaron su similitud de coseno. Luego utilizaron esta puntuación de similitud como variable sustituta del sector al que pertenecen las empresas. Los autores argumentaron que su análisis proporciona una medida mucho más rica y continua de la intercambiabilidad de productos que los códigos tradicionales de clasificación sectorial.[22]

Sin embargo, la similitud de coseno no es una panacea. El cuadro 6 muestra dos documentos hipotéticos. Ambos tratan sobre la educación, pero emplean palabras diferentes para abordar el tema. Esta característica del lenguaje se conoce como sinonimia, es decir, el mismo tema subyacente puede describirse haciendo uso de muchas palabras diferentes. Aunque el tema es el mismo en ambos documentos, por emplear un vocabulario diferente su similitud de coseno será baja.

Cuadro 6

SINONIMIA EN UNA MUESTRA DE DOCUMENTOS

Documento 1

escuela	universidad	colegio	maestro	profesor
0	5	5	1	2

Documento 2

escuela	universidad	colegio	maestro	profesor
10	0	0	4	0

Nota: este cuadro muestra dos documentos con contenido diferente pero una similitud coseno baja debido a polisemia.

[22] Ver, por ejemplo, el sistema británico de clasificación estándar de actividades económicas industriales (Office for National Statistics, 2007).

La polisemia también está relacionada con el problema de la sinonimia: una misma palabra puede tener significados diferentes en contextos diferentes. Considérense los dos documentos en el cuadro 7.

Cuadro 7

POLISEMIA EN UNA MUESTRA DE DOCUMENTOS

Documento 1

tanque	marino	rana	animal	naval	guerra
5	5	3	2	0	0

Documento 2

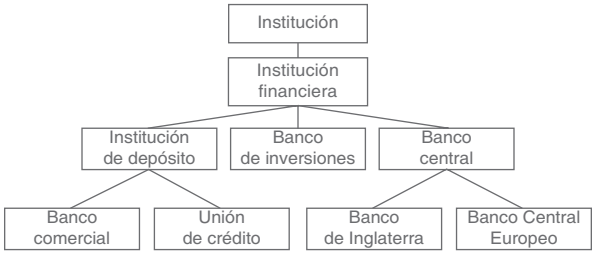
tanque	marino	rana	animal	naval	guerra
5	5	0	0	4	3

Nota: este cuadro muestra dos documentos con contenido diferente pero una similitud de coseno relativamente alta debido a polisemia.

Supóngase que el documento 1 es sobre animales y el documento 2, sobre cuestiones bélicas. El problema es que la misma palabra tiene diferentes significados dependiendo del contexto. Por ejemplo, en una conversación sobre animales, *tanque* se refiere al lugar donde viven peces o anfibios; en una conversación sobre guerra, se refiere a un arma mecanizada. Esto significa que dos documentos esencialmente sin relación alguna pueden mostrar una similitud de coseno elevada.[23]

[23] Además de la sinonimia y la polisemia, también podemos considerar la relación jerárquica entre palabras utilizando algoritmos basados en tesauros (Jurafsky y Manning, 2012). Decimos que una palabra es el *hipónimo* de otra si la primera palabra es una subcategoría de la otra. Por ejemplo, en la gráfica que sigue, obsérvese que *receptoras de depósitos* es un hipónimo de *institución financiera*. Y, a su vez, *institución financiera* es un *hiperónimo* de *receptora de depósitos*. Una de las maneras más sencillas de saber si dos palabras son similares es viendo qué tan próximas están en la jerarquía. Definimos la longitud de ruta, pathlen(w_1, w_2), como uno más el número de orillas en la ruta más corta en el gráfico de hiperónimos entre w_1 y w_2. Luego, la *similitud de longitud de rutas* puede definirse mediante

$$pathsim(w_1, w_2) = \frac{1}{pathlen(w_1, w_2)}$$



En el diagrama superior, la longitud de ruta entre *banco comercial* e *institución* es igual a la que existe entre *banco comercial* y *banco de inversión*. Sin embargo, semánticamente, *banco comercial* está relacionado más específicamente con *banco de inversión* que con la idea genérica de *institución*. Esto sugiere que necesitamos una medida más adecuada. Una posible mejoría es aumentar palabras en el tesauro con valores de contenido de información (CI). Los valores de CI se computan con base en el conteo de frecuencia de las palabras que se encuentran en el texto. Para cada palabra *w* en el tesauro, el CI se define como el logaritmo negativo de la probabilidad de esa palabra. Otras técnicas más complejas, como la medida de Resnik (res), toman como dato de entrada dos palabras w_1 y w_2 y producen una medida de similitud. La técnica se basa en la idea del mínimo ancestro común (least common subsumer, LCS), es decir, la palabra más específica que es un ancestro común de las dos palabras. Por ejemplo, el LCS de *banco comercial* y *unión de crédito* es *receptora de depósitos*. El LCS se evalúa directamente desde el CI: $res(w_1, w_2) = CI(LCS(w_1, w_2))$.

2.6 Análisis semántico latente

Un supuesto de los modelos de espacio vectorial es que las palabras en un texto son conceptualmente independientes de las demás. Pero esto no siempre es así. Por ejemplo, *naval* y *marino* pudieran considerarse como expresiones superficiales de un tema latente más fundamental, como es *guerra*.

Los *modelos de variable latente* que ahora vamos a comentar adoptan este enfoque. Suponen que las palabras no son independientes y, más bien, están interrelacionadas por temas subyacentes no observados. Estos modelos tienen cuatro virtudes. Abordan la sinonimia al asociar cada palabra en el vocabulario con un tema latente dado. Consignan la polisemia al permitir que cada palabra tenga asociaciones con múltiples temas. También asocian cada documento con temas más que con

palabras. Y permiten a los algoritmos encontrar la mejor asociación entre palabras y variables latentes, sin utilizar listas o categorías de palabras predefinidas, en contraste con la técnicas booleanas y basadas en diccionarios.

Análisis semántico latente (latent semantic analysis, *LSA*; Deerwester *et al*. 1990) constituye uno de los primeros ejemplos del método de variable latente. El LSA comienza por representar la matriz término-documento mediante la descomposición de valores singulares, que consiste en la búsqueda de los componentes principales de las filas y las columnas de una matriz término-documento. Es decir, el LSA calcula las combinaciones lineales de términos que explican gran parte de la variación de términos entre documentos, así como las combinaciones lineales de documentos que explican gran parte de la variación de documentos entre términos. Entonces, la idea consiste en aproximar la matriz de término-documento utilizando únicamente los componentes principales y medir la similitud de los documentos con la aproximación, más que con la matriz de término-documento verdadera.[24] La hipótesis es que los componentes principales representan los temas en común y que los componentes descartados representan las elecciones de palabras idiosincrásicas.

Una aplicación que hizo un banco central recientemente del análisis semántico latente se muestra en un artículo de Acosta (2014), quien estudia el efecto de una mayor transparencia en las reuniones del Comité Federal de Mercado Abierto (Federal Open Market Committee, FOMC) de la Reserva Federal de Estados Unidos.[25] La Reserva ha publicado varias versiones estenográficas de las reuniones del FOMC desde octubre de 1993. Antes de esa fecha, los miembros del FOMC no estaban conscientes de que sus deliberaciones estaban siendo grabadas. Pero después de que el Congreso ejerciera presión para que la Reserva se volviera más transparente, el expresidente de esta, Alan Greenspan, descubrió que el personal había estado transcribiendo palabra por palabra las reuniones desde mediados de los años setenta. Así, convino en publicar las transcripciones previas y darlas a conocer en lo sucesivo con un rezago de cinco años.

Acosta analizó si esa mayor conciencia de los miembros del FOMC respecto a que los comentarios en las reuniones estaban siendo grabados y serían divulgados cambió su comportamiento. El autor aplica la descomposición de valores singulares y utiliza los 200 componentes principales para medir la similitud de documentos. Lo que descubrió fue un mayor apego a las convenciones después de que empezaron a publicarse las transcripciones.

2.7 La asignación de Dirichlet latente

Un problema de la descomposición de valores singulares es que los temas que produce no son probabilísticos. La *asignación de Dirichlet latente* (latent Dirichlet allocation, LDA) corrige esto.

La LDA es un *modelo de elementos mixtos* en el que palabras y documentos son probabilidades asignadas y se asocian a múltiples temas. Esto contrasta con los *modelos deterministas de membrecía* única en el que palabras y documentos se asignan únicamente a un tema. Este aspecto probabilístico de la LDA es importante. Supóngase un tema sobre inflación y otro sobre desempleo. Ahora considérese la palabra *tasa*. De entrada, no se sabe bien a bien a qué tema asociar *tasa*, pues un tema sobre inflación o desempleo podría incluir la *tasa de inflación* o de la *tasa de participación* en la fuerza laboral, respectivamente. Permitir la asignación probabilística de palabras a temas permite esta flexibilidad semántica.

En términos formales cada documento tiene su propia distribución de probabilidad en los temas.

[24] Sin embargo, hay otras maneras de realizar la descomposición de la matriz. Por ejemplo, Hendry (2012) aplica la *factorización de matrices no negativas* para estudiar si la comunicación de los bancos centrales afecta a los mercados y, de ser así, cómo lo hace.

[25] Hendry y Madeley (2010) y Masawi *et al*. (2014) aplican el análisis semántico latente a documentos de bancos centrales.

Luego, para cada palabra en cada documento se realiza una asignación de tema y, dependiendo de la asignación, una palabra del tema correspondiente.[26] Un ejemplo ilustra lo anterior. Supóngase que el hablante A y el hablante B hablan acerca de los temas 1 y 2. El hablante A dedica dos terceras partes de su tiempo al tema 1 y el hablante B dedica dos terceras partes de su tiempo al tema 2. En este caso, las palabras observadas por A pueden pensarse que se eligieron de la siguiente manera. Para cada palabra, A elabora un tema y, siendo la probabilidad del 0.67, este es el tema 1; y con una probabilidad del 0.33, este es el tema 2. Una vez que se elabora un tema para una palabra, la propia palabra se elabora a partir de la distribución de probabilidad asociada a cada tema. Para el hablante B, la probabilidad de que una palabra trate sobre el tema 1 es 0.33 y sobre el tema 2 es 0.67. Pero una vez que una palabra se asigna a un tema dado, se obtiene de la distribución de ese tema que es común a ambos hablantes.

La LDA ha encontrado amplia aplicación en las ciencias computacionales y la estadística, y está empezando a aparecer en la economía. Por ejemplo, Hansen y sus coautores (2014) estudiaron los efectos de una mayor transparencia del banco central. Al igual que Acosta, encontraron evidencia de un mayor apego a las convenciones después de difundirse las transcripciones del FOMC.

Los dos temas que calcularon aparecen en la gráfica 9, que representa temas como nubes de palabras en las que el tamaño de la palabra es proporcional a su probabilidad en el tema. No se imponen categorías durante el cálculo. Estas son sencillamente las agrupaciones 25ª y 40ª de palabras estimadas. El tema 25 (arriba) claramente trata acerca de la inflación, mientras que el tema 40 (abajo) claramente trata acerca del riesgo. Así que podemos utilizar el resultado de la LDA para medir el contenido subyacente de una manera directamente significativa.

2.8 Clasificación jerárquica descendente

Además de la LSA y la LDA, hay otras técnicas de minería de textos que parten de la base de que las palabras no son independientes unas de otras, sino reflejan temas subyacentes. Por ejemplo, Schonhardt-Bailey (2013) y Vallès y Schonhardt-Bailey (2015) estudiaron los bancos centrales que utilizan el software Alceste.[27] En pocas palabras, Alceste intenta descubrir categorías estables de términos que estén asociadas máximamente en el interior, mientras que están mínimamente asociadas a otras

[26] Una descripción más formal del modelo es la siguiente. Supóngase un corpus conformado por K temas, donde un tema individual β_k es un vector de probabilidad sobre los elementos únicos V en el vocabulario. Por lo tanto, todas las palabras pueden aparecer en todos los temas, aunque con un peso diferente. Al mismo tiempo, cada documento individual en el corpus es una distribución de probabilidad θ_d sobre los temas K. Cada palabra individual $w_{d,n}$ en el documento d se genera mediante un proceso de dos pasos. Primero, una asignación temática $z_{d,n}$ se obtiene de θ_d. Segundo, una palabra se obtiene de $\beta_{z_{d,n}}$. Esto contrasta con el modelo de membrecía única en el que a un documento se le asignaría un único tema $z_{d,n}$, y cada palabra en el documento d se obtendría de β_{z_d}.

[27] Alceste es el acrónimo de Analyse des Lexèmes Co-occurents dans les Énnoncés Simples d'un Texte (Análisis de los Lexemas Coocurrentes dentro de los Enunciados Simples de Texto). Image-Zafar distribuye el programa. Ver <http://www.image-zafar.com/>. El programa fue desarrollado originalmente por Max Reinert. Sus diversas publicaciones en más de 20 años, principalmente en francés, documentan los primeros pasos de Alceste (Reinert, 1983; Reinert, 1987; Reinert, 1990; Reinert, 1993; Reinert, 1998; y Reinert, 2003). Una reproducción de código abierto, basada en R, de Alceste se encuentra en el programa informático de Iramuteq <http://www.iramuteq.org/>. Desde 1983, una comunidad de investigadores y analistas de textos cada vez más interdisciplinaria ha empleado Alceste como técnica de minería de textos (Noel-Jorand *et al.*, 1995; Lahlou, 1996; Jenny, 1997; Noel-Jorand *et al.*, 1997; Brugidou, 1998; Guerin-Pace, 1998; Bauer, 2000; Brugidou, 2000; Brugidou, 2003; Noel-Jorand *et al.*, 2004; Schonhardt-Bailey, 2005; Schonhardt-Bailey, 2006; Bara *et al.*, 2007; Schonhardt-Bailey, 2008; Schonhardt-Bailey *et al.*, 2012; Weale *et al.*, 2012; Schonhardt-Bailey, 2013; Vallès y Schonhardt-Bailey, 2015). Peart (2013) utiliza Alceste para estudiar si las preferencias de los miembros del CPM permanecen estables con el tiempo.

Gráfica 9

ESTIMACIÓN DE DOS TEMAS DE LAS TRASCRIPCIONES DEL CORPUS DEL FOMC DURANTE EL PERIODO DE ALAN GREENSPAN

Fuente: Hansen, McMahon y Prat (2014).

categorías.[28] Para hacer esto, Alceste construye una matriz que cruza todos los enunciados –denominados unidades elementales de contexto (elementary context units, ECU)[29]– y todas las palabras, y cuyas celdas indican la presencia o ausencia de esa palabra en el enunciado, de manera parecida a una matriz de término-documento, pero con una unidad más pequeña de análisis textual. Las celdas indican la ausencia o presencia de esa palabra en el enunciado, representada por un 0 o un 1, respectivamente. El cuadro 8 muestra un ejemplo de este tipo de matriz.

Alceste luego divide el contenido de esta tabla en dos categorías, con el fin de maximizar la similitud de las ECU en la misma categoría y, al mismo tiempo, maximizar la diferencia entre categorías. El conjunto total de ECU en la matriz inicial constituye la primera categoría. El algoritmo luego busca una división que minimice el número de palabras traslapadas. El traslape se mide mediante el *valor ji cuadrado* (χ^2) de un cuadro con dos filas, comparando las distribuciones observadas con las esperadas. El algoritmo luego intenta maximizar los valores χ^2 repitiendo el proceso de división (*clasificación jerárquica descendente*), es decir, al probar si mediante dividir las clases en subclases más pequeñas mejora los valores x. El proceso iterativo de clasificación jerárquica descendente

[28] Una manera de proceder más sencilla sería simplemente contar el número de ocurrencias y coocurrencias. Por ejemplo, Ronnqvist y Sarlin (2012) investigan la co-ocurrencia de los nombres de bancos finlandeses en un importante foro financiero en línea. Meten nombres de bancos y sus co-ocurrencias en el mismo registro en una red, donde el tamaño del nodo y el peso de la arista lo proporcionan el número de ocurrencias y de coocurrencias, respectivamente. Aunque es posible percibir algunos cambios en la concentración y fuerza de las conexiones, un proceso más objetivo consiste en consignar esos cambios temporales utilizando las medidas de la centralidad de la red. Los autores detectan un incremento en el número de ocasiones en las que se mencionan a los bancos finlandeses juntos, durante y después de la crisis financiera.

[29] Las unidades elementales de contexto o enunciados son *oraciones medidas* que el programa construye automáticamente con base en la puntuación del texto. En el análisis de textos, un problema persistente y difícil consiste en la longitud óptima de un *enunciado*. Y es que el lenguaje podría analizarse en términos de oraciones, párrafos, frases y demás. Alceste resuelve este problema al no tratar de identificar directamente la longitud del enunciado. Más bien, produce clasificaciones que son independientes de la longitud de los enunciados. Se crean dos clasificaciones, cada una de las cuales utiliza unidades de contexto de distinta longitud; sólo las categorías que aparecen en ambas clasificaciones se conservan para análisis y tales categorías son independientes de la longitud de los enunciados. Esto deja sin clasificar cierto número de ECU, con lo cual se aproxima una medición de la bondad del ajuste. La calidad de la división se mide construyendo un cuadro que cruza todas las categorías obtenidas en la primera clasificación y todas las categorías obtenidas en la segunda clasificación. El resultado es un *cuadro de* χ^2 *señaladas*, es decir, un cuadro de datos con vínculos positivos y negativos entre la categorías. Este cuadro ayuda a elegir las categorías que comparten el mayor número de ECU.

Cuadro 8

PALABRA MEDIANTE UNA MATRIZ DE ECU

	ECU						
Palabra	*1*	*2*	*3*	*4*	*5*	*…*	*j*
1	0	1	1	1	1	…	0
2	1	1	0	0	0	…	1
…	…	…	…	…	…	…	…
i	0	1	1	0	1	…	1
Totales	2	31	5	10	67	…	2

finaliza cuando un número predeterminado de iteraciones ya no produce divisiones estadísticamente significativas.

Por otro lado, para cada categoría se produce una lista de palabras y la fuerza de asociación entre cada palabra y la categoría se expresa mediante un valor χ^2 y un coeficiente *fi* (ϕ), donde la distribución observada de palabras se compara con una esperada. Por ejemplo, si el vocabulario es diferente en ambas categorías, la distribución observada se desviará sistemáticamente de una distribución esperada consistente de independencia de palabras. Las relaciones entre categorías también pueden descomponerse y examinarse de forma espacial utilizando el factor de *análisis de correspondencia.* Un ejemplo de producción de análisis de correspondencia ya se mostró en la gráfica 3. Las posiciones de los puntos dependen de las correlaciones, donde la distancia refleja el grado de coocurrencia.[30] Con respecto a los ejes, el análisis de correspondencia pretende identificar una cantidad máxima de asociación en el primer eje (horizontal). El segundo eje (vertical) busca dar cuenta de un máximo de la asociación restante.[31]

2.9 Aprendizaje automático supervisado

El análisis semántico latente, la asignación Dirichlet latente y la clasificación jerárquica descendente son ejemplos de algoritmos de aprendizaje automático sin supervisión. En contraste, los algoritmos de aprendizaje automático supervisados comienzan cuando el investigador clasifica manualmente los datos de entrenamiento en categorías predefinidas, como en los métodos basados en diccionarios. Con el fin de evitar el problema de sobreajuste, el algoritmo es posteriormente validado en otra serie de documentos, denominados datos de comprobación, antes de que sea aplicado al resto del corpus (Grimmer y Stewart, 2013).

Tal vez la aplicación más fructífera de las técnicas de aprendizaje supervisado en economía es cuando el investigador tiene categorías de textos bien justificadas.[32] Una posible aplicación para los bancos centrales consiste en asociar texto con una orientación rigurosa o condescendiente, partiendo de Apel y Grimaldi (2012).[33] Una aplicación bien conocida en economía utilizando datos de textos se encuentra en Gentzkow y Shapiro (2010). Sus datos de entrenamiento consisten en una muestra grande de discursos ante el Congreso de Estados Unidos. Cada discurso lleva una categoría que corresponde al partido del orador e identifica

[30] Para hacer esto, el análisis de correspondencia utiliza la *distancia ji cuadrado*, que se asemeja a la distancia euclidiana entre puntos en el espacio físico. Cada diferencia al cuadrado entre coordenadas se divide entre el elemento correspondiente del perfil promedio (donde el perfil es una serie de frecuencias divididas entre su total). La razón para utilizar el concepto de ji cuadrado es que permite transformar las frecuencias dividiendo las raíces cuadradas de las frecuencias esperadas, con lo cual se normalizan las varianzas. Esto puede compararse con el análisis factorial, donde los datos en diferentes escalas se estandarizan.

[31] Sin embargo, muchos casos requieren más de dos dimensiones para consignar la dimensionalidad de los datos. Así, Alceste proporciona el porcentaje que se acumula a cada dimensión, pero limita la representación gráfica a dos y tres dimensiones.

[32] Alternativamente, pudieran haber sido ya asignadas categorías como parte de los metadatos.

[33] Sin embargo, pudiera ser difícil clasificar los documentos de antemano (Grimmer y Stewart, 2013). Por ejemplo, la naturaleza multifacética de las reuniones, los discursos y las conversaciones pudiera dificultar reducir un documento a un único tema.

las frases partidistas. Luego, asignan una calificación a otro corpus conformado por artículos de prensa, como izquierdista o derechista, con base en la presencia o ausencia de frases partidistas.

Un algoritmo popular de aprendizaje automático supervisado es Naïve Bayes.[34] Naïve Bayes aplica la regla de Bayes de que la categoría más probable de un documento es aquella que maximiza el producto de dos factores, *P(c)* y *P(dlc),* donde *d* es el documento y *c* es la categoría. El factor *P(c)* se conoce como la probabilidad previa de la categoría. Consigna con qué frecuencia ocurre una categoría en los datos de entrenamiento. El factor *P(dlc)* se conoce como la probabilidad. Consigna la probabilidad de que un documento *d* dada la categoría, cuando *d* puede representarse como un vector de palabras $d = x_1 + x_2 + x_3 + \cdots + x_n$, donde *n* es el número total de palabras. Para cada palabra, el factor de probabilidad puede estimarse viendo el número de veces que la palabra aparece en esa categoría particular, como coeficiente de todas las palabras asociadas con esa categoría en los datos de entrenamiento. En la práctica, estas probabilidades se calculan llevando todos los textos dentro de una categoría particular a un único documento combinado de entrenamiento para la categoría y, posteriormente, contando las frecuencias relativas de w_i como un coeficiente del número general de palabras *w* en los datos de entrenamiento.[35]

Un ejemplo de investigación utilizando Naïve Bayes es el estudio de Moniz y Jong (2011) sobre los efectos de las minutas del CPM del Banco respecto a las expectativas de la tasa de interés futura. Los autores emplean Naïve Bayes en combinación con otras técnicas de minería de textos comentadas en este documento. Primero, buscaron las palabras en las páginas de Wikipedia sobre *banca central* e *inflación* y consignaron las palabras que se asocian[36] a *crecimiento económico*, *precios*, *tasas de interés* y *crédito bancario*. Estas palabras luego se utilizan como categorías en un modelo de Naïve Bayes asignándolas a oraciones en las minutas del CPM. Las categorías asignadas se utilizan posteriormente para construir un índice de sentimiento utilizando un método sencillo basado en diccionario.[37] Por último, el algoritmo de LDA se utiliza para encontrar palabras que pudieran actuar como intensificadores y atenuadores del sentimiento, por ejemplo, "aumento" y "moderado", respectivamente. La combinación de métodos de Moniz y Jong destaca un aspecto importante de la minería de textos en la práctica: las técnicas supervisadas y no supervisadas suelen complementarse y emplearse en distintas etapas del proceso de minería de textos.

[34] Naïve Bayes es *ingenuo* en dos sentidos. Primero, parte del supuesto de una bolsa de palabras sencilla de que el orden de las palabras no importa, por lo que sólo considera la frecuencia de las palabras en un documento. Segundo, supone que la probabilidad de cada palabra que aparece en una categoría dada es independiente de la presencia de otras palabras, a pesar de que ya hemos señalado que es probable que se trate de un supuesto erróneo (Jurafsky y Manning, 2012).

[35] Por supuesto, es posible que una palabra asociada a una categoría no aparezca en los datos de entrenamiento. Supóngase que el banco X tiene sus oficinas principales en Londres. Pero que la palabra *Londres* no aparece en un documento de entrenamiento de 2,000 palabras clasificado como *banco X*. Como resultado, la probabilidad estimada de Londres dado el tema del banco X sería cero. Con el fin de evitar este resultado falso, podemos utilizar un procedimiento denominado alisamiento de Laplace agregando 1, que simplemente implica sumar 1 a la siguiente ecuación:

$$\hat{P}\left(w_i \middle| c_j\right) = \frac{cuenta\left(w_i, c_j\right) + 1}{\sum_{w \in V} cuenta\left(w_i, c_j\right) + V} \quad ,$$

donde w_i es una palabra determinada, c_j es una categoría determinada y V representa todas las palabras en el corpus.

[36] Los autores utilizan TextRank, un algoritmo de clasificación basado en gráficas para determinar grupos de palabras asociadas (Mihalcea y Tarau, 2004).

[37] Los autores utilizan el diccionario General Inquirer <http://www.wjh.harvard.edu/~inquirer/>, como ya se señaló.

3. CONCLUSIÓN

El propósito de este documento ha sido demostrar el valor adicional que pueden obtener los bancos centrales si aplican las distintas técnicas de minería de textos, así como ejemplificar el uso de estas por quienes formulan las políticas y abordar los principales temas de investigación que interesan a los bancos centrales. Para concluir, deseamos destacar que la promesa de la minería de textos para los bancos centrales no es sólo hipotética sino que se ha comprobado. Por ejemplo, Kevin Warsh (2014) citó publicaciones sobre minería de textos (Schonhardt-Bailey, 2013; Hansen, McMahon y Prat, 2014) como influencias importantes de las recomendaciones finales de política que formaron parte de su análisis, a saber, que el Banco divulgara más información sobre sus deliberaciones. Warsh destacó que si bien "los estudios que buscan dar sentido a millones de palabras habladas" son "intimidantes e imperfectos", la minería de textos "ha hecho grandes avances en nuestra comprensión" de los bancos centrales (Warsh, 2014). En términos más generales, este texto ha mostrado de qué manera la minería de textos puede ser una contribución útil al arsenal analítico de los bancos centrales y ayudarles a alcanzar sus objetivos de política.

BIBLIOGRAFÍA

Aase, K. G. (2011), *Text Mining of News Articles for Stock Price Predictions*, tesis de maestría, Norwegian University of Science and Technology.

Acosta, J. M. (2014), *FOMC Responses to Calls for Transparency: Evidence from the Minutes and Transcripts Using Latent Semantic Analysis*, tesis distinguida, Department of Economics, Stanford University, disponible en <http://economics.stanford.edu/content/honors-thesis-2014>.

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

Baker, S. R., N. Bloom y S. J. Davis (2013), *Measuring Economic Policy Uncertainty*, Chicago Booth Research Paper, núm. 13-02.

Banco de Inglaterra (2014), "Inflation Report Q&A, 13th August 2014", disponible en <http://www.bankofengland.co.uk/publications/Documents/inflationr eport/2014/conf130814.pdf>.

Banco de Inglaterra (2015), "One Bank Research Agenda Discussion Paper", disponible en <http://www.bankofengland.co.uk/research/Documents/onebank/dis cussion.pdf>

Bara, J., A. Weale y A. Biquelet (2007), "Analysing Parliamentary Debate with Computer Assistance", *Swiss Political Science Review*, vol. 13, núm. 4, pp. 577-605.

Bauer, M. (2000), "Qualitative Researching with Text, Image and Sound: A Practical Handbook", en M. W. Bauer y G. Gaskell, *Classical Content Analysis: A Review*, Sage Publications, Londres, pp. 131-151.

Bennani, H., y E. Farvaque (2014), "Speaking in Tongues? Diagnosing the Consistency of Central Banks' Official Communication", disponible en <http://www.econ.cam.ac.uk/epcs2014/openconf/modules/request.php?module=oc_program&action=view.php&id=198>.

Bholat, D. (2015), "Big Data and Central Banks", *Bank of England Quarterly Bulletin,* vol. 55, núm. 1, pp. 86-93.

Blinder, A. S., M. Ehrmann, M. Fratzscher, J. de Haan y D. J. Jansen (2008), "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence", ECB Working Paper Series*,* núm. 898.

Bollen, J., H. Mao y X. Zeng (2011), "Twitter Mood Predicts the Stock Market", *Journal of Computational Science*, vol. 2, núm. 1, pp. 1-8.

Brugidou, M. (1998), "Epitaphes, l'image de Francois Mitterrand à travers l'analyse d'une question ouverte posée à sa mort", *Revue Française de Science Politique*, vol. 48, núm. 1, pp. 97-120.

Brugidou, M. (2000), "Les discours de la revendication et de l'action dans les éditoriaux de la presse syndicale (1996-1998)", *Revue Française de Science Politique*, vol. 50, núm. 6, pp. 962-992.

Brugidou, M. (2003), "Argumentation and Values: An Analysis of Ordinary Political Competence Via an Open-ended Question", *International Journal of Public Opinion Research*, vol. 15, núm. 4, pp. 413-430.

Bulir, A., M. Cihak y D. J. Jansen (2014), *Does the Clarity of Inflation Reports Affect Volatility in Financial Markets?*, IMF Working Paper, núm. 14/175*.*

Carney, M. (2013), "Crossing the Threshold to Recovery", discurso en el Banco de Inglaterra, 28 de agosto.

Carney, M. (2015), "One Bank Research Agenda: Launch Conference", discurso en el Banco de Inglaterra, 25 de febrero.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer y R. Harshman(1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society for Information Science,* vol. 41, núm. 6, pp. 391-407.

Eckley, P. (2015), *Measuring Economic Uncertainty Using News-media Textual Data*, MPRA Paper, núm. 64874*,* disponible en <http://mpra.ub.uni-muenchen.de/64874/>.

Gai, P., A. Haldane y S. Kapadia (2011), "Complexity, Concentration and Contagion", *Journal of Monetary Economics,* vol. 58, núm. 5, pp. 453-470.

Gentzkow, M., y J. M. Shapiro (2010), "What Drives Media Slant? Evidence from U.S. Daily Newspapers", *Econometrica,* vol. 78, núm. 1, pp. 35-71.

Giles, C. (2015), "Bank of England Mark Carney Expands Research Agenda", *Financial Times,* 25 de febrero de 2015.

Grimmer, J., y B. M. Stewart (2013), "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts", *Political Analysis*, vol. 21, pp. 267-297.

Guerin-Pace, F. (1998), "Textual Analysis. An Exploratory Tool for the Social Sciences", Population: An English Selection*,* número especial de *New Methodological Approaches in the Social Sciences*, vol. 10, núm. 1, pp. 73-95.

Haldane, A. (2012), "The Dog and the Frisbee", discurso en el Banco de Inglaterra, 31 de agosto.

Haldane, A. (2015), "The Promise of New Data and Advanced Analytics", discurso en el Banco de Inglaterra, 25 de febrero.

Hansen, S., M. McMahon y A. Prat (2014), "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach", CEP Discussion Paper, Centre for Economic Performance, LSE, núm. DP1276.

Hendry, S. (2012), *Central Bank Communication or the Media"s Interpretation: What Moves Markets?*, Bank of Canada Working Paper, núm. 2012-9.

Hendry, S., y A. Madeley (2010), *Text Mining and the Information Content of Bank of Canada Communications*, Bank of Canada Working Paper, núm. 2010-31.

Hoberg, G., y G. M. Phillips (2010), "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis", *The Review of Financial Studies*, vol. 23, núm. 10, pp. 3773-3811.

Humpherys, S., K. C. Moffitt, M. B. Burns, J. K. Burgoon y W. F. Felix (2011), "Identification of Fraudulent Financial Statements Using Linguistic Credibility Analysis", *Decision Support Systems*, vol. 50, pp. 585-594.

Jansen, D. J., y J. de Haan (2010), *An Assessment of the Consistency of ECB Communication Using Wordscores*, Nederlandsche Bank Working Paper, núm. 259.

Jenny, J. (1997), "Techniques and Formalized Practices for Content and Discourse Analysis in Contemporary French Sociological Research", *Bulletin de Méthodologie Sociologique*, vol. 54, pp. 64-112.

Jurafsky, D., y C. Manning (2012), "Natural Language Processing", discurso en línea, disponible en <https://www.coursera.org/course/nlp>.

Kloptchenko, A., C. Magnusson, B. Back, A. Visa y H. Vanharanta (2004), "Mining Textual Contents of Financial Reports", *The International Journal of Digital Accounting Research,* vol. 4, núm. 7, pp. 1-29.

Lahlou, S. (1996), "A Method to Extract Social Representations from Linguistic Corpora", *Japanese Journal of Experimental Social Psychology,* vol. 35, núm. 3, pp. 278-291.

Li, W. P., P. Azar, D. Larochelle, P. Hill y A. W. Lo (2015), "Law is Code: A Software Engineering Approach to Analyzing the United States Code", *Journal of Business & Technology* Law, vol. 10, núm. 2, pp. 297-374.

Loughran, T., y B. McDonald (2011), "When Is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks", *Journal of Finance*, vol. 66, núm. 1, pp. 35-65.

Masawi, B., S. Bhattacharya y T. Boulter (2014), "The Power of Words: A Content Analytical Approach Examining Whether Central Bank Speeches Become Financial News", *Journal of Information Science*, vol. 40, núm. 2, pp. 198-210.

McLaren, N., y R. Shanbhogue (2011), "Using Internet Search Data as Economic Indicators", *Bank of England Quarterly Bulletin,* vol. 51, núm. 2.

Mihalcea, R., y P. Tarau (2004), "TextRank: Bringing Order into Texts", en Association for Computational Linguistics*, Proceedings of EMNLP 2004*, pp. 404-411.

Moniz, A., y F. de Jong (2011), "Predicting the Impact of Central Bank Communications on Financial Market Investors' Interest Rate Expectations", *Lecture Notes in Computer Science*, vol. 8798, pp. 144-155.

Nergues, A., J. Lee, P. Groenewegen y I. Hellsten (2014), "The Shifting Discourse of the European Central Bank: Exploring Structural Space in Semantic Networks", en *Signal-image Technology and Internet- Based Systems (SITIS), 2014 Tenth International Conference*, pp. 447, 455, 23-27.

Nivre, J., J. Hall y J. Nilsson (2006), "MaltParser: A Data-driven Parser-generator for Dependency Parsing", en *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (LREC2006), pp. 2216-2219.

Noel-Jorand, M. C., M. Reinert, M. Bonnon y P. Therme (1995), "Discourse Analysis and Psychological Adaptation to High Altitude Hypoxia", *Stress Medicine*, vol. 11, pp. 27-39.

Noel-Jorand, M. C., M. Reinert, S. Giudicelli y D. Dassa (1997), "A New Approach to Discourse Analysis in Psychiatry, Applied to Schizophrenic Patient Speech", *Schizophrenia Research*, vol. 25, pp. 183-198.

Noel-Jorand, M. C., M. Reinert, S. Giudicelli y D. Dassa (2004), "Schizophrenia: The Quest for a Minimum Sense of Identity to Ward Off Delusional Psychosis", *The Canadian Journal of Psychiatry*, vol. 49, núm. 6, pp. 394-398.

Nyman, R., D. Gregory, S. Kapadia, P. Ormerod, D. Tuckett y R. Smith (2015), *News and Narratives in Financial Systems: Exploiting Big Data for Systemic Risk Assessment*, material mimeografiado.

O'Connor, B., R. Balasubramanyan, B. R. Routledge y N. A. Smith (2010), "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", *Proceedings of the 4th International Conference on Weblogs and Social Media*, pp. 122-129.

Office for National Statistics (2007), *UK Standard Industrial Classification of Economic Activities 2007*.

Peart, J. (2013), "How Do Appointment Processes Affect the Policy Outputs of Monetary Policy Committees?", disponible en <http://johnpeart.org/dissertation/>.

Reinert, M. (1983), "Une methode de classification descendante hierarchique: application a l'analyse lexicale par contexte", *Les Cahiers de l'Analyse des Donnees*, vol. 8, núm. 2, pp. 187-198.

Reinert, M. (1987), "Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud", *Bulletin de Méthodologie Sociologique*, vol. 13, pp. 53-90.

Reinert, M. (1990), "Alceste. Une methodologie d'analyse des donnees textuelles et une application: *Aurelia* de Gerard de Nerval", *Bulletin de Methodologie Sociologique*, vol. 26, pp. 24-54.

Reinert, M. (1993), "Les 'mondes lexicaux' et leur 'logique' à travers l'analyse statistique d'un corpus de récits de cauchemars", *Langage et Société*, vol. 66, pp. 5-39.

Reinert, M. (1998a), "Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste", *Proceedengs of the 4th JADT (Journées d'Analyse des Données Textuelles)*, Universidad de Niza, JADT.

Reinert, M. (1998b), *Alceste Users' Manual (English version)*, Image, Toulouse.

Reinert, M. (2003), "Le rôle de la répétition dans la représentation du sens et son approche statistique dans la méthode Alceste", *Semiotica*, vol. 147, núm. 1-4, pp. 389-420.

Ronnqvist, S., y P. Sarlin (2012), "From Text to Bank Interrelation Maps", *Computational Intelligence for Financial Engineering & Economics*, 2104 IEEE Conference, pp. 48-54.

Rosa, C., y G. Verga (2006), "On the Consistency and Effectiveness of Central Bank Communication: Evidence from the ECB", *European Journal of Political Economy*, vol. 23, núm. 1, pp. 146-175.

Schonhardt-Bailey, C (2005), "Measuring Ideas More Effectively: An Analysis of Bush and Kerry's National Security Speeches", *Political Science and Politics*, vol. 38, núm. 3, pp. 701-711.

Schonhardt-Bailey, C. (2006), *From the Corn Laws to Free Trade: Interests, Ideas and Institutions in Historical Perspective*, MIT Press, Cambridge, MA.

Schonhardt-Bailey, C. (2008), "The Congressional Debate on Partial-birth Abortion: Constitutional Gravitas and Moral Passion", *British Journal of Political Science*, vol. 38, pp. 383-410.

Schonhardt-Bailey, C., E. Yager y S. Lahlou (2012), "Yes, Ronald Reagan's Rhetoric Was Unique –But Statistically, How Unique?", *Presidential Studies Quarterly*, vol. 42, núm. 3, pp. 482-513.

Schonhardt-Bailey, C. (2013), *Deliberating American Policy: A Textual Analysis*, MIT Press, Cambridge, MA.

Siklos, P. L. (2013), *The Global Financial Crisis and the Language of Central Banking: Central*

Bank Guidance in Good Times and in Bad, CAMA Working Paper, núm. 58/2013.

Tetlock, P. C. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market", *The Journal of Finance*, Vol. LXII, núm. 3.

Upshall, M. (2014), "Text Mining: Using Search to Provide Solutions", *Business Information Review,* vol. 31, pp. 91-99.

Vallès, D. W., y C. Schonhardt-Bailey (2015), "Forward Guidance as Central Bank Discourse: MPC Minutes and Speeches under King and Carney", presentado en el Political Leadership and Economic Crisis Symposium, Yale University (febrero).

Warsh, K. (2014), "Transparency and the Bank of England's Monetary Policy Committee", reseña de Kevin Warsh.

Weale, A., A. Bicquelet y B. Judith (2012), "Debating Abortion, Deliberative Reciprocity and Parliamentary Advocacy", *Political Studies*, vol. 60, pp. 643-667.