CAPÍTULO I

Economía, mercados y geopolítica: el papel de los modelos de lenguaje natural en las ciencias sociales

Alvaro Ortiz
Tomasa Rodrigo

La digitalización de la información y el desarrollo de la inteligencia artificial están propiciando un cambio sin precedentes en la disponibilidad de nuevos datos. Gran parte de la información a la que hoy en día podemos acceder se produce de manera no estructurada. Gracias al desarrollo de los modelos de lenguaje natural, el texto se ha convertido en una de las principales fuentes de información y la posibilidad de trasladar "texto a números" se está convirtiendo en una poderosa herramienta de análisis en las ciencias sociales. El último exponente son los modelos generativos del lenguaje (*LLM,* por sus siglas en inglés). En este capítulo exploramos a través de varios ejemplos el papel que estos desarrollos pueden jugar dentro del análisis de las ciencias sociales con un especial foco en el ámbito económico.

Palabras clave: procesamiento de lenguaje natural, análisis de sentimento, *big data,* economía, geopolítica, mercados.

1. INTRODUCCIÓN

En los últimos años hemos sido testigos del rápido crecimiento y desarrollo de la inteligencia artificial (IA), que ha pasado a formar parte de nuestro día a día como, por ejemplo, a través de asistentes de voz, recomendadores de compras virtuales o plataformas digitales de entretenimiento. Este crecimiento no es un fenómeno aislado, sino que es el resultado del desarrollo y avances de distintos factores coincidentes en el tiempo.

En primer lugar, el principal factor es el aumento exponencial de la información con la digitalización y la capacidad de almacenar todos estos datos en la nube. Otro factor relevante es el crecimiento sustancial en el poder computacional y las capacidades de procesamiento de la información, que durante las últimas décadas ha avanzado de forma exponencial, permitiendo procesar cada vez una mayor cantidad de información en menos tiempo. Esto, además, ha venido acompañado de una reducción de los costes computacionales y de procesamiento de los datos, que hacen posible, y escalable, el uso de los mismos. Estos avances han impulsado a su vez el rápido desarrollo de algoritmos, metodologías y nuevos modelos adaptados a la naturaleza de los datos en la nube, que permiten sacar el máximo provecho a la alta frecuencia y granularidad de la información.

En la actualidad, la mayor parte de nuestra actividad diaria está digitalizada a partir de nuestra interacción con nuestro móvil, el ordenador, la tarjeta bancaria o nuestras redes sociales, entre otros muchos ejemplos. Mucha de esta información está no estructurada en forma de texto, vídeos, imágenes o voz (como, por ejemplo, el contenido de redes sociales, blogs, documentos personales, correos electrónicos, mensajes de texto, búsquedas en internet, fotografías, audio, vídeo...). Gracias al desarrollo de los algoritmos de procesamiento de lenguaje natural (PLN) y la inteligencia artificial, como los algoritmos de redes neuronales para el procesamiento de imágenes y vídeos, estos datos pueden ahora convertirse en datos estructurados para ser procesados y analizados.

Además de los individuos, las entidades público/privadas también generan ingentes cantidades de datos. El sector público, por ejemplo, genera datos sustanciales en forma de registros públicos, mientras que el sector privado produce información muy detallada fruto de su actividad, de alto valor para el análisis, como las señales que emite un terminal móvil, la huella que dejan las transacciones financieras o las interacciones en aplicaciones y plataformas digitales.

En los últimos años, el uso de las técnicas de procesamiento de lenguaje natural se ha visto revolucionado por la nueva generación de los modelos de redes neuronales profundas, conocidos como transformadores, que pueden detectar patrones sutiles y significados semánticos en el lenguaje. Esta reciente explosión de los grandes modelos de lenguaje (*Large Language Models* en inglés, *LLM*) ha generalizado el análisis de estos datos no estructurados en todas las disciplinas.

En este capítulo exploramos el uso de estas técnicas de PLN, que convierten el texto en números, detallando numerosas aplicaciones desarrolladas en BBVA Research para enriquecer y complementar el análisis económico, financiero, social y geopolítico.

En la primera sección, hacemos un repaso a la trayectoria del desarrollo de las técnicas de análisis de texto y procesamiento de lenguaje natural, desde la minería de texto a los grandes modelos de lenguaje, mostrando ejemplos en la literatura de cómo se han utilizado cada una de estas técnicas en economía. En la segunda sección, enumeramos las distintas aplicaciones desarrolladas por BBVA Research en el ámbito económico, empresarial, geopolítico y político. Finalmente, en la tercera sección, concluimos, enumerando los retos a futuro.

2. EL USO DEL LENGUAJE EN LAS CIENCIAS SOCIALES: SU EVOLUCIÓN DE LA MINERÍA DE TEXTO A LOS GRANDES MODELOS DEL LENGUAJE

El análisis de texto no es un área nueva de estudio, de hecho su uso se remonta a hace más de un siglo. Sin embargo, ha sido en la última década cuando ha experimentado una evolución transformadora, convirtiéndose en una poderosa fuente de valor para investigadores y analistas de datos lingüísticos.

Hasta las últimas dos décadas, el análisis de texto se centraba básicamente en la interpretación de una lectura humana profunda y detallada, la cual no se podía escalar a los grandes volúmenes de texto que tenemos disponibles hoy día. La creciente disponibilidad de información digitalizada en forma de texto ha propiciado el desarrollo y sofisticación de las técnicas de procesamiento del lenguaje para el análisis del mismo. Ash y Hansen (2023) ofrecen una visión detallada de los métodos utilizados en economía para el análisis de texto a partir de distintas metodologías, así como sus limitaciones, especialmente en el campo de la validación de resultados generados por los algoritmos de PLN.

En esta sección, hacemos un repaso a la evolución de las herramientas de PLN, que han evolucionado desde aplicaciones frecuentistas de conteo de palabras que se analizan de forma independiente a tener en cuenta el contexto y contenido semántico de las mismas, mejorando la capacidad de los modelos para representar la estructura temática subyacente en los datos y ayudando a una comprensión más profunda del lenguaje.

2.1. Bolsa de palabras (*Bag of Words*)

El modelo bolsa de palabras (*Bag of Words, BoW,* por sus siglas en inglés) es la forma más sencilla de representar documentos, que consiste en convertir el texto en un formato numérico donde cada documento se representa como un vector y cada dimensión del vector es una palabra que toma como valor el conteo de la ocurrencia de la misma en el documento.

Esta metodología cuantifica la presencia y frecuencia de palabras dentro de los textos, pero sin identificar ninguna relación entre ellas, es decir, no reconoce palabras derivadas, sinónimos, etcétera.

Para aplicar esta metodología, el primer paso se basa en la limpieza y preparación del texto, que se organiza en unidades básicas llamadas tokens[1] en la literatura de PLN, que normalmente son palabras, pero también pueden ser caracteres o subpalabras. De este conjunto de tokens se eliminan los caracteres no alfabéticos como los signos de puntuación y palabras de uso común (como "el/la", "en", "y"), con poco valor analítico y que no agregan significado al texto, conocidas en inglés como *stopwords*. Este paso ayuda a reducir el tamaño del conjunto de datos y a mejorar el tiempo de procesamiento de los mismos.

A continuación, se convierten todas las letras en minúsculas y se aplica una técnica de normalización del texto conocida como *stemming* que consiste en reducir las palabras a su forma raíz, cortando prefijos y sufijos según reglas del lenguaje (por ejemplo, de "niña" y "niñez" a "niñ").

Finalmente, se calcula la conocida matriz TF-IDF (Frecuencia de Término – Frecuencia Inversa de Documento, comúnmente conocida como matriz documento-término), que es una medida estadística para clasificar lo relevante que es una palabra para un documento en una colección de documentos. En dicha matriz, cada fila representa un documento y cada columna representa una palabra de toda la colección de documentos. Así, las celdas de la matriz contienen la frecuencia de cada palabra en cada documento. Esta métrica aumenta proporcionalmente al número de veces que una palabra aparece en el documento, pero se compensa con la frecuencia de la palabra en el corpus, lo que ayuda a ajustar el hecho de que algunas palabras aparecen más frecuentemente en general.

Tras las modificaciones enumeradas previamente, se crea la bolsa de palabras, que contiene todos los tokens que aparecen en el texto, ignorando el orden y el contexto en el que aparecen. Cada token representa una característica en el modelo.

El siguiente paso es la vectorización. Cada documento se transforma en un vector. Esto implica contar la frecuencia con la que cada token de la bolsa de palabras aparece en el documento. El resultado es un vector donde cada índice representa un token de la bolsa y cada valor representa la frecuencia de ese token en el documento analizado.

Estos vectores se utilizan para entrenar modelos de aprendizaje automático como la clasificación de textos o análisis de sentimiento a partir de diccionarios asistidos. Son numerosas las aplicaciones de esta técnica en economía como primera aproximación para convertir texto en números y analizar estas nuevas fuentes de información. Permite, por ejemplo, cuantificar la frecuencia y presencia de palabras específicas relevantes en textos de diversa índole para analizar políticas económicas, identificando prioridades, cambios de política o impacto de

[1] Esta técnica también implica eliminar las características lingüísticas únicas del idioma, como los acentos y las eñes en español.

las mismas; monitorizar indicadores económicos donde la frecuencia y cambio en términos económicos específicos pueden señalar cambios en dichos indicadores antes de que se publiquen las cifras oficiales; o analizar comunicados de prensa o informes corporativos. Baker *et al.* (2016) se basan en esta metodología para construir un índice que mide la incertidumbre de la política económica a partir de la búsqueda de palabras clave en artículos de prensa de los principales periódicos estadounidenses y europeos. Este trabajo ha sido de las primeras iniciativas en el uso de las técnicas de análisis de texto en economía, convirtiéndose en un referente es este campo. Loughran y McDonald (2011 y 2016) realizan un estudio exhaustivo de cómo se utiliza el análisis de texto y los modelos de bolsa de palabras en finanzas, analizando su aplicación en el análisis de los comunicados financieros, informes de resultados y otros documentos corporativos.

De esta forma, el modelo de bolsa de palabras se puede usar como herramienta para transformar datos textuales no estructurados en datos estructurados y analizables, facilitando el análisis de texto. No obstante, aunque el modelo de bolsa de palabras es intuitivo y muy sencillo de implementar, es bastante limitado dado que ignora el orden de las palabras y el contexto, dificultando la comprensión más profunda del lenguaje.

2.2. Modelos de temas (*Topic Models*)

Los modelos de temas (*topic models,* en inglés) han supuesto un avance significativo en el análisis de texto y el procesamiento de lenguaje natural frente a los modelos de bolsa de palabras. Esta técnica se basa en la reducción de la dimensionalidad agrupando los tokens en temas subyacentes a partir de una colección de documentos, ayudando a identificar de qué se habla en cada documento.

Una de las primeras técnicas de reducción de la dimensionalidad es el análisis semántico latente (*LSA,* por sus siglas en inglés) (Deerwester *et al.,* 1990). Reconociendo las limitaciones de los modelos de bolsa de palabras para captar el significado y los temas de los documentos, esta metodología *LSA* fue el primer intento de incorporar información semántica. *LSA* utiliza técnicas de componentes principales en la matriz documento-término (que muestra la frecuencia relativa de cada palabra en un conjunto de documentos, como se define en la sección anterior) para reducir su dimensionalidad, identificando así las relaciones latentes entre palabras y documentos. Estas relaciones nos dan una pista de los temas subyacentes, pero no tienen un marco probabilístico asociado, lo que dificulta la interpretación de sus resultados (para más detalle consultar Hash y Hansen (2023)).

Para cubrir esta problemática, Hofmann (1999) desarrolló el análisis semántico latente probabilístico (*pLSA*), que introduce un modelo probabilístico que asume que un documento es una mezcla de temas y un tema es una mezcla de palabras. Esto permite una comprensión más intuitiva de que los documentos y las palabras se generan a partir de temas latentes. Cada documento se modela como una distribución probabilística sobre temas, y cada tema es una

distribución sobre palabras. No obstante, el modelo *pLSA* también tiene limitaciones, dado que no proporciona un modelo generativo completo y tiende a sobreajustarse con un gran número de parámetros como citan Ash y Hansen (2023).

Blei *et al.* (2003) resuelven esta problemática con la asignación de Dirichlet Latente (*LDA,* por sus siglas en inglés), que se trata de un modelo probabilístico basado en *pLSA,* pero incorpora la distribución de Dirichlet para las distribuciones de temas dentro de documentos y distribuciones de palabras dentro de temas, lo que ayuda a manejar el sobreajuste y proporciona un modelo más robusto. De esta forma, los componentes identificados representan el contenido temático latente que se encuentran en los patrones de co-ocurrencia de términos en cada documento que hacen referencia a un mismo tema.

Este modelo *LDA* asigna una distribución de temas a cada documento, lo que significa que un documento puede estar compuesto por varios temas en proporciones diferentes. Por ejemplo, un artículo de noticias podría contener un 70 % del tema "economía" y un 30 % del tema "finanzas".

El paso final es interpretar los temas identificados y utilizarlos para el análisis posterior. La interpretación de los conjuntos de tokens relacionados que conforman un tema es labor del analista y requieren una interpretación cuidadosa, dado que suelen depender del contexto y del conocimiento del dominio.

Una vez identificados estos componentes en temas, se puede estudiar la prevalencia de los diferentes temas a lo largo del tiempo, entre diferentes fuentes de información o como parte de un análisis más amplio como el análisis de sentimiento, identificando cómo cambia el sentimiento e importancia de cada uno de estos temas a lo largo del tiempo.

Esta técnica se ha convertido en la más común para la reducción de la dimensionalidad en procesamiento de lenguaje natural por su interpretabilidad y robustez. Son numerosos los trabajos que emplean *LDA* en economía, como Hansen *et al.* (2018), que usan un modelo *LDA* para entender las transcripciones del Comité de Mercado Abierto de la Reserva Federal, destacando cómo puede utilizarse esta metodología para entender las temáticas más relevantes en el lenguaje de bancos centrales y sus estrategias de comunicación o Bybee *et al.* (2023), que utilizan *LDA* para identificar de qué se habla en noticias financieras y empresariales del *Wall Street Journal* y su importancia para medir el ciclo económico.

Como variantes basadas en *LDA,* Blei y Lafferty (2006) desarrollaron los modelos de temas dinámicos (*DTM,* por sus siglas en inglés) para capturar la evolución de temas a lo largo del tiempo, introduciendo en el modelo *LDA* unos parámetros que evolucionan suavemente para estimar la prevalencia a lo largo del tiempo como parte del propio modelo de temas. Posteriormente, Roberts *et al.* (2013) presentaron el modelo temático estructural (*STM,* por sus siglas en inglés) que incorpora la estructura del corpus al modelo temático estándar del *LDA*. De esta forma, se tiene en cuenta no solo la prevalencia de los temas sino el contenido

temático. En la siguiente sección comentaremos algunas de las aplicaciones desarrolladas en BBVA Research basadas en el uso de *LDA* y *STM*.

Pese a su amplia aplicación y utilidad, estos modelos también cuentan con limitaciones dado que ignoran el orden de las palabras y no tienen en cuenta el contexto de forma que no captan frases con significado similar, pero expresadas de forma distinta, el sarcasmo o la ironía.

2.3. Representaciones vectoriales de palabras (*Word Embeddings*)

Una técnica más avanzada en el campo del procesamiento de lenguaje natural utilizada para representar palabras como vectores en un espacio multidimensional son los *word embeddings,* representaciones vectoriales de palabras o incrustaciones de palabras en español. Estos modelos permiten que las relaciones entre palabras informen del significado de las mismas y sus vectores capten la información semántica y sintáctica de las palabras.

Para ello, cada palabra se representa como un vector en un espacio de alta dimensionalidad. Estos vectores no son arbitrarios, sino que están diseñados de tal manera que palabras con significados o contextos similares tienen representaciones vectoriales similares. De esta forma, en un modelo de *word embeddings,* las palabras "gato" y "perro" estarán ubicadas cercanamente en el espacio vectorial, reflejando su relación, mientras que ambas tendrán una representación distinta a "sombrero", dado que no guardan ninguna relación semántica con esta última palabra. Estos modelos permiten, además, que el significado venga determinado por las palabras vecinas y funcionan mejor cuando se entrenan con grandes volúmenes de texto. Su buen funcionamiento depende en gran medida de la calidad y diversidad del corpus de entrenamiento utilizado, dado que es durante el entrenamiento cuando el modelo aprende a asociar palabras con su contexto.

Entre los modelos más populares para generar *word embeddings* está el *Word2Vec*, desarrollado por Mikolov *et al.* (2013a y 2013b), que se basa en la idea de que el significado de una palabra se puede inferir del contexto en el que aparece. Para ello, se usan redes neuronales poco profundas para incrustar palabras en un espacio vectorial continuo. Las redes neuronales se aplican para aprender de las representaciones vectoriales de palabras basadas en su contexto. La arquitectura de la red neuronal utilizada en los *embeddings* de palabras generalmente implica entrenar la red para predecir tanto las palabras de contexto circundantes dada una palabra objetivo, como al contrario. De esta forma, las incrustaciones de palabras generadas por *Word2Vec* capturan relaciones sintácticas y semánticas. Este modelo se basa en la arquitectura Skip-Gram que predice palabras de contexto a partir de las palabras objetivo. Así, para cada palabra objetivo, mira una ventana de palabras de contexto circundantes e intenta predecirlas. Como comentamos anteriormente, esta metodología funciona mejor cuanto mayor es el volumen de información.

Otro conocido modelo basado en *word embeddings* son los vectores globales para la representación de palabras (*Global Vectors for Word Representation, GloVe,* por sus siglas en inglés) (Pennington *et al.,* 2014). Este modelo *GloVe* se basa en otra aproximación para la incrustación de palabras, centrándose en las matrices de co-ocurrencia palabra-palabra, diseñado así para construir vectores de palabras que codifican la co-ocurrencia local. De esta forma, se construye una matriz de co-ocurrencia global en la que se recoge la frecuencia de palabras que aparecen juntas en todo el corpus analizado. No obstante, *GloVe* también incorpora información del contexto local de cada palabra, considerando la probabilidad de co-ocurrencia de palabras. Esto le ayuda a capturar tanto las relaciones semánticas entre palabras en todo el texto, matizando estas relaciones a partir del contexto local. Este modelo proporcionó un enfoque complementario a *Word2Vec* y mostró cómo se pueden combinar diferentes tipos de información (como la factoriación matricial global con métodos de contexto local) para obtener representaciones de palabras robustas.

Los modelos basados en *word embeddings* son utilizados en una variedad de aplicaciones de PLN, incluyendo la traducción automática, la extracción de información, el análisis de sentimiento y los chatbot, dado que entienden mejor el significado de las palabras en diferentes contextos. En economía, Bloom *et al.* (2021) utilizan los *word embeddings* para identificar frases asociadas a las nuevas tecnologías en patentes, ofertas de empleo e informes de resultados para estudiar la difusión de empleos relacionados con las nuevas tecnologías. Por otro lado, Ash y Gennaro (2023) utilizan esta metodología de *word embeddings* para construir una escala de emocionalidad en los discursos políticos pronunciados en el Congreso de los Estados Unidos durante 1858-2014.

No obstante, los *word embeddings* también tienen ámbitos de mejora como la dificultad para capturar el significado de palabras polisémicas (palabras con múltiples significados) o manejar palabras nuevas (fuera del vocabulario incluido en el contexto analizado).

2.4. Grandes modelos de lenguaje (*Large Language Models*)

Finalmente, los grandes modelos de lenguaje (*Large Language Models, LLM,* por sus siglas en inglés) representan la evolución más disruptiva y reciente en el procesamiento del lenguaje natural e inteligencia artificial. Estos modelos son sistemas avanzados de aprendizaje automático diseñados para entender, interpretar y generar texto, ofreciendo una comprensión profunda del mismo y una capacidad de respuesta lingüística sin precedentes.

Sus miles de millones de parámetros les permiten procesar y generar lenguaje con gran precisión y están entrenados con ingentes cantidades de texto para identificar patrones lingüísticos, gramaticales y semánticos del idioma, adaptándose además a diferentes estilos y dialectos y resolviendo el problema de la polisemia o palabras nuevas que presentan los *embeddings*. Los *LLM* no solo son capaces de comprender el texto de entrada, sino también de generar texto de salida coherente y contextualmente relevante, resolviendo preguntas de distinta índole, resumiendo documentos o traduciendo los mismos, generando ideas, etcétera.

Muchos de estos modelos están basados en arquitecturas de redes neuronales avanzadas, como los modelos *Transformers*[2], que han revolucionado el procesamiento del lenguaje entrenando a los algoritmos para que también "presten atención" a las características relevantes del contexto específico. Estos modelos *Transformers* fueron desarrollados por Vaswani *et al.* (2017) en el documento de trabajo seminal en la literatura de PLN *Attention is All You Need.* La innovación clave de los modelos *Transformers* es el uso de mecanismos de autoatención, que permiten al modelo ponderar la importancia de diferentes palabras en una frase o secuencia, enfocándose dinámicamente en diferentes partes de los datos de entrada al hacer predicciones o generar datos de salida en forma de texto.

En esta tecnología *Transformers* se basan los conocidos modelos de lenguaje como es el caso de BERT (Devlin *et al.,* 2019), desarrollado por Google. Su lanzamiento supuso un cambio significativo en el análisis con PLN frente a modelos previos al incorporar una profunda comprensión del contexto y sutileza del lenguaje. Este modelo aprende a partir de grandes corpus a predecir palabras ocultas en función de su contexto, así como predecir si una frase es continuación lógica de otra para entender las relaciones entre ellas. Posteriormente, Facebook AI desarrolló RoBERTa (Liu *et al.,* 2019), una versión optimizada de BERT que mejora significativamente el rendimiento del modelo y se entrena con volúmenes de corpus significativamente mayores. Sin embargo, han sido en los últimos dos años cuando estos modelos *Transformers* han evolucionado a velocidad de vértigo, cambiando la forma en la que podemos usar la inteligencia artificial (IA) generativa en nuestro día a día. En 2022, Google presentó PALM (Chowdhery *et al.,* 2022) como modelo de IA generativa que mejora los modelos anteriores en términos de escalabilidad, comprensión y flexibilidad. Recientemente, Google junto con DeepMind, han presentado PALM 2, que soporta más de cien idiomas y también es capaz de escribir código, y finalmente han lanzado a final de 2023 Gemini (Gemini y Google, 2024), su modelo más avanzado de inteligencia artificial generativa a partir de tareas multimodales que introduce notables capacidades de comprensión de imágenes, audio, vídeo y texto. No ha sido menos significativa la evolución de la familia de modelos *GPT* (*Generative Pre-trained Transformer,* por sus siglas en inglés) (Radford *et al.,* 2018; Brown *et al.,* 2020; OpenAI, 2023), culminado con su última versión ChatGPT 4, diseñado para entender el lenguaje humano y generar respuestas como si de una persona se tratara. ChatGPT 4 se ha convertido en los últimos meses en la solución de IA generativa más utilizada del mercado basado, al igual que Gemini, en modelos multimodales para trabajar también con imágenes y código. No obstante, la evolución exponencial de estos modelos y la competencia de las grandes tecnológicas como Open AI, Google y Facebook AI por conseguir la mejor solución en AI evidencian el enorme potencial actual y futuro de estas herramientas para el análisis.

Estos modelos ya han sido aplicados a distintos campos dentro de la economía. Korinek (2023) describe distintos casos de uso en los que los *LLM* son de utilidad para la investigación económica en la generación de ideas, escritura, investigación, análisis de datos y codificación. Shapiro *et al.* (2022) muestra el uso de Bert entre otras metodologías de PLN para el análisis

[2] Véase Phuong y Hutter (2022) para una explicación más detallada y precisa de los modelos *Transformers*.

de sentimiento y desarrolla una nueva medida del sentimiento económico a partir de artículos de prensa económica y financiera de enero de 1980 a abril de 2015. Hansen *et al.* (2023a) usan modelos *LLM,* ajustando un modelo estándar basado en la metodología *transformers* para que tenga en cuenta la estructura lingüística específica de las ofertas de empleo con el objetivo de medir y caracterizar el cambio al trabajo a distancia producido tras la pandemia. Por otro lado, Hansen *et al.* (2023b) evalúan la capacidad de los modelos *GPT* para clasificar la orientación política de los anuncios del Comité Federal de Mercado Abierto en Estados Unidos en relación con la valoración humana.

Mas allá del ámbito académico, los grandes modelos del lenguaje también han sido aplicados en el ámbito corporativo para aprovechar el potencial de estas tecnologías. El trabajo de Dabravolski *et al.* (2023) describe el modelo *GPT* personalizado por Bloomberg con 50.000 millones de parámetros, entrenado con datos financieros recopilados a lo largo de cuarenta años con el objetivo de ayudar a la compañía en sus aplicaciones con análisis de texto.

A pesar de su avanzada capacidad, los modelos *LLM* también tienen algunos desafíos aún pendientes de resolver. Estos modelos pueden replicar o amplificar sesgos presentes en los datos de entrenamiento. Además, la interpretación de matices lingüísticos y culturales, así como la gestión de ambigüedades y polisemias del idioma, sigue siendo un área en desarrollo.

3. APLICACIONES DE LAS TÉCNICAS DE LENGUAJE NATURAL EN LAS CIENCIAS SOCIALES EN BBVA RESEARCH

El procesamiento del lenguaje natural se ha convertido en una herramienta muy valiosa en las ciencias sociales, ofreciendo nuevas perspectivas que complementan y enriquecen el conocimiento de la sociedad. Estas técnicas, descritas en la sección anterior, ayudan al investigador en la comprensión del comportamiento humano y los patrones sociales.

Desde BBVA Research son numerosas las aplicaciones desarrolladas a partir del uso de distintas metodologías de procesamiento de lenguaje natural descritas anteriormente. A continuación, enumeramos algunas de ellas en el campo de la economía, mercados y ámbito empresarial, así como en el área geopolítica y política.

3.1. El uso del análisis de texto en economía

En BBVA Research hemos desarrollado una amplia gama de indicadores en tiempo real con técnicas de procesamiento de lenguaje natural y análisis de sentimiento para monitorizar la evolución de temas candentes con impacto en la economía y construir indicadores difícilmente medibles con datos numéricos.

3.1.1. De noticias globales a indicadores en tiempo real

La información disponible en medios de comunicación en forma de noticias *online*, radio y televisión se han convertido en los últimos años en una rica base de datos, altavoz de temas de índole muy diversa de interés social. Gracias a las técnicas de PLN podemos analizar de qué se habla, cómo, dónde y cuándo se habla de cualquier tema, persona u organización.

En BBVA Research llevamos más de una década trabajando con una base de datos en la nube llamada GDELT (Base de Datos Global de Eventos, Lenguaje y Tono)[3], que extrae, procesa y analiza noticias en medios de comunicación a nivel mundial en más de cien idiomas diariamente, desde fuentes de medios globales, nacionales, regionales hasta locales, todos ellos traducidos al inglés automáticamente.

Esta fuente de datos *big data* utiliza diferentes diccionarios para identificar miles de temas, entre los que se incluye todo el glosario de temas del Banco Mundial[4], para clasificar y categorizar la información. Los algoritmos utilizados por GDELT también identifican emociones, organizaciones, ubicaciones, fuentes de noticias y eventos en todo el mundo. Además, generan un sentimiento promedio de cada pieza de información.

Centrándonos en el sentimiento, GDELT aplica más de 40 diccionarios diferentes que clasifican palabras asociadas con tonos positivos y negativos para calcular el tono o sentimiento promedio de todos los documentos que contienen una o más menciones de los eventos o temas que queremos monitorizar de acuerdo a la siguiente fórmula:

$$\overline{S_h} = \frac{\Sigma w_h^+ - \Sigma w_h^-}{\Sigma w_h} * 100 \qquad [1]$$

Donde S es el sentimiento por pieza de información h, y W representa las palabras en cada artículo. La puntuación varía de -100 (extremadamente negativo) a +100 (extremadamente positivo), aunque los valores comunes se sitúan entre -10 y +10, siendo 0 indicativo de neutral.

Usando esta base de datos de noticias globales, estudiamos la evolución de temas económicos difícilmente medibles como la incertidumbre de política económica (*EPU*). Para ello, monitorizamos el sentimiento medio y la cobertura mediática de este tema incluido en la taxonomía de GDELT[5]. La construcción del índice es un producto ponderado formado por la

[3] Véase www.gdelt.org para más información.

[4] Véase el siguiente enlace para acceder a la lista detallada de temas definidos por el Banco Mundial https://vocabulary.worldbank.org/taxonomy/1737.html

[5] Dentro de la taxonomía de GDELT monitorizamos el tema "epu_economy" que incluye noticias relacionadas principalmente con política fiscal y gasto público, política monetaria y tipos de interés, política comercial y aranceles, política regulatoria y asuntos legales, pero también indicadores económicos y tendencias del mercado, acontecimientos internacionales y tensiones geopolíticas, crisis de salud pública y pandemias, desastres naturales y fenómenos meteorológicos, política energética y medioambiental, política laboral y de empleo, crecimiento económico y recesión, inflación y deflación, mercados bursátiles y financieros, banca e instituciones financieras, beneficios empresariales e informes financieros, confianza y gasto de los consumidores, fabricación y producción, vivienda y sector inmobiliario, política de infraestructuras y transporte.

cobertura relativa del tema, medida como el número de noticias relacionadas con el término de búsqueda de *EPU* ese día sobre el número total de noticias publicadas ese mismo día, multiplicado por el sentimiento medio de estos artículos. Multiplicamos el indicador resultante por -1 para facilitar la interpretación, de forma que valores positivos indican mayor incertidumbre por un aumento de la cobertura y/o un empeoramiento del sentimiento medio. La construcción del índice se puede resumir con la siguiente fórmula:

$$EPU_{t,i} = \frac{\Sigma_{s \in G} COV_{t,i}}{\Sigma_G COV_{t,i}} * avg\left(-\overline{S_{t,i}}\right) \quad [2]$$

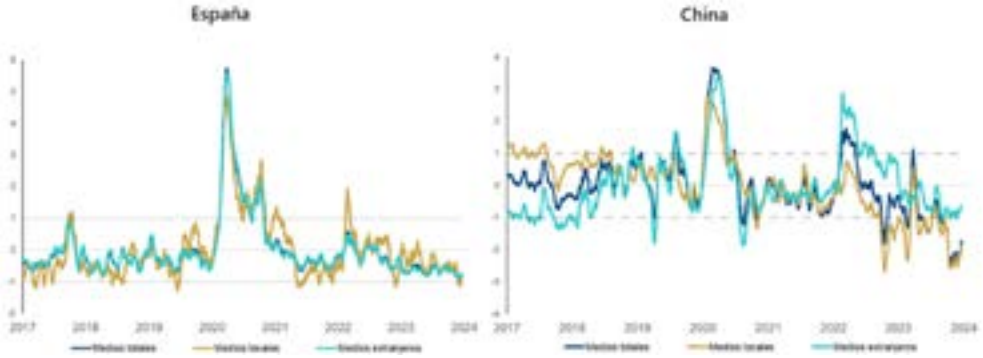
Donde *t* es el tiempo (día) e *i* es el país. *s* se refiere a los artículos de noticias sobre *EPU* y *G* a toda la base de datos. *COV* es a la cobertura mediática y *S* es el sentimiento como vimos en la fórmula [1].

La información se recoge a nivel país. Para presentar indicadores homogéneos, todos los índices se normalizan y se transforman aplicando un promedio móvil ponderado de 28 días[6] para reducir el ruido de la información diaria e identificar señales más claramente.

Figura 1.

Índices *big data* BBVA Research de incertidumbre de política económica

Media móvil 28 días. Indicador normalizado desde 2017



Nota: Valores positivos (negativos) indican mayor (menor) incertidumbre relativo a la media del periodo de 2017 hasta la actualidad.

Fuentes: BBVA Research y www.gdelt.org

La figura 1 muestra el indicador *Big Data* BBVA Research de Incertidumbre de Política Económica para el caso de España y China. Estos indicadores por país nos muestran que a mayor valor (bien por una mayor cobertura al tema, por un empeoramiento del sentimiento

[6] Se aplica media móvil 28 días para incluir cuatro semanas completas con el mismo número de días de la semana y evitar el ruido que pueda generar que no todos los días de la semana hay el mismo volumen de información.

o por la combinación de ambas), mayor es la preocupación por temas de incertidumbre de política económica. Además, podemos distinguir la evolución del indicador por la fuente de información, clasificándolos en medios locales del país, fuentes extranjeras y medios totales, que nos ayudan a identificar las diferentes percepciones o sesgos de los medios de comunicación en determinadas regiones. En el caso de España existe una mayor similitud en las percepciones de incertidumbre de política económica en prensa local y extranjera que en el caso de China. Sobre estas diferencias por origen de los medios de comunicación, en el caso de España, destaca que en medios locales españoles el índice empeora frente a la percepción en prensa extranjera en enero de 2021 debido a las implicaciones económicas de la última oleada monitorizada del COVID-19 con la variante ómicron, acompañada del primer mes de datos de inflación positivos tras el estadillo del COVID-19 y los cuellos de botella en las cadenas globales de valor. El indicador local también sobresale a principios de 2022 con el estadillo de la guerra de Ucrania y más recientemente en 2023 con el periodo electoral, primero por las elecciones municipales en mayo, seguidas de las elecciones generales en julio. En el caso de China, destaca la brecha significativa entre las señales de mayor incertidumbre que arroja el indicador en prensa extranjera frente al de prensa local desde inicios de 2022, al contrario que sucedía en 2017 y 2018.

Utilizando la información de los medios de comunicación también analizamos un tema candente relacionado con las cadenas globales de valor como es la crisis de los semiconductores, impulsada por la guerra comercial entre EE. UU. y China, los problemas relacionados con la capacidad de la industria después del COVID-19, el mal tiempo durante 2020-2022 y eventos geopolíticos clave como las tensiones entre China y Taiwán y la guerra entre Rusia y Ucrania.

Figura 2.

Indicador Global *Big Data* BBVA Research de Semiconductores (2018-2023)
Media móvil 28 días. Indicador normalizado desde 2018

Nota: Valores positivos (negativos) indican mayores (menores) tensiones relativo a la media del periodo de 2018 hasta la actualidad.

Fuentes: BBVA Research y www.gdelt.org

Los semiconductores son esenciales para las tecnologías modernas y desempeñan un papel clave en algunas industrias estratégicas. Con el uso del *big data* de noticias globales y técnicas de PLN e inteligencia artificial, desarrollamos indicadores de sentimiento mediático en tiempo real para analizar su evolución en el tiempo (figura 2) siguiendo la metodología descrita en el caso anterior de los indicadores *EPU* (Fórmula [2]). El indicador es, por tanto, un índice ponderado de la cobertura relativa del tema de los semiconductores por el sentimiento medio y multiplicado por -1 de tal forma que valores mayores indican mayor riesgo, tensiones o peor sentimiento.

El Indicador Global Big Data BBVA Research de Semiconductores muestra que la crisis de los semiconductores y su posterior normalización se ha visto impulsada por varios factores, como la guerra comercial entre Estados Unidos y China, la escasez relacionada con el COVID-19 y los problemas de capacidad, el mal tiempo, las tensiones geopolíticas y los avances en la diplomacia internacional. La guerra comercial entre EE. UU. y China en 2018-2019 generó un aumento del indicador a la zona de riesgo a mediados de 2019. Posteriormente, el COVID-19 provocó alteraciones sin precedentes en las cadenas de suministro mundiales que, junto con un aumento de la demanda de productos electrónicos debido al trabajo a distancia, generaron un aumento del indicador debido a la escasez de oferta. Pero fue la guerra entre Rusia y Ucrania en 2022 el mayor amplificador de los problemas en las cadenas de suministro mundiales, lo que llevó al mayor aumento del indicador en el horizonte temporal estudiado. Desde mediados de 2023, el indicador muestra una relajación en el mercado gracias al alivio de las tensiones entre EE. UU. y China y una mejora de los cuellos de botella en la producción de los mismos.

El uso del PLN y el análisis de texto en medios de comunicación también nos permite entender relaciones e interconexiones entre temas, geografías, personas, organizaciones… teniendo en cuenta la co-ocurrencia de los mismos en prensa a través del análisis de redes. Para entender las posibles implicaciones económicas de un hipotético conflicto entre China y Taiwán tiene especial relevancia el peso de Taiwán en la Industria Mundial de Semiconductores a través de la empresa Taiwan Semiconductor Manufacturing Company (TSMC), que es particularmente relevante para la producción de chips para los dispositivos más avanzados, utilizados especialmente por los países desarrollados. Para analizar el papel sistémico de TSMC en la industria de semiconductores, construimos una red de noticias globales con GDELT basada en la co-ocurrencia de empresas y sectores con la empresa TSMC en la misma noticia (figura 3).

Los nodos o vértices de la red en la figura 3 representan empresas y las aristas indican la relación entre ellas medida como el conteo de noticias donde aparecen ambas empresas mencionadas. Existen varias medidas de centralidad, como el grado, la proximidad y la interrelación o intermediación. Con el objeto de analizar el papel sistémico de TSMC, implementamos un algoritmo de centralidad de autovectores en la red para ponderar el tamaño de cada nodo en función del número de aristas o relaciones, así como la relevancia de estas conexiones de aristas, dadas sus propias relaciones con otros nodos en la red. Un mayor tamaño del nodo significa que un nodo está conectado a muchos otros nodos que, a su vez, tienen también

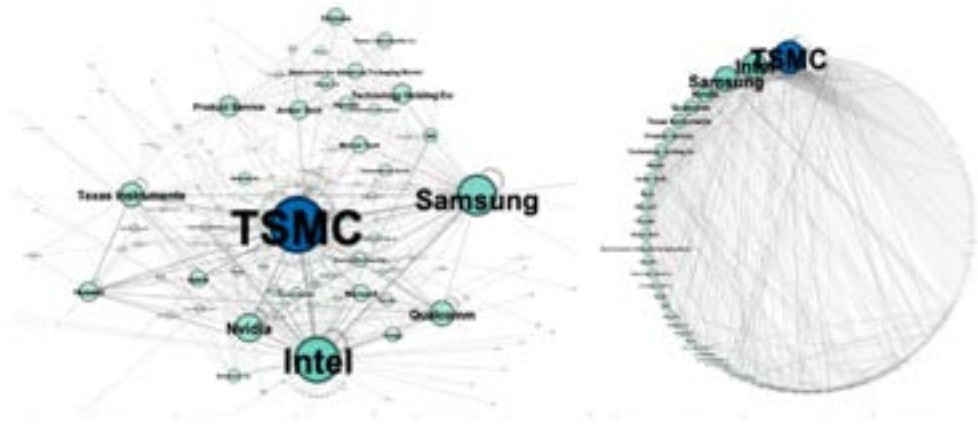
muchas relaciones con otros nodos, lo que refleja mejor la naturaleza sistémica de un nodo en la red.

Esta red de interconexiones entre compañías nos permite describir las posibles relaciones en las noticias de la crisis de semiconductores de Taiwán con el resto de la industria. En ella vemos la importancia y centralidad de TSMC en la industria. La red también muestra la relación de TSMC con el resto de las empresas de la industria donde grandes fabricantes de dispositivos integrados (conocidos como *IDMs,* por sus siglas en inglés, *Integrated device manufacturers*) como Intel, Samsung, Texas Instruments y Micron Tech tienen un peso relevante por su relación con TSMC, dado que subcontratan parte de su producción para tecnologías avanzadas. Se identifican también en la red clientes relevantes de TSMC como Nvidia y Qualcomm, seguidas por Broadcom y AMD. Adicionalmente, encontramos grandes empresas tecnológicas como Apple, Huawei y Microsoft, que dependen de TSMC para la producción de chips personalizados.

Figura 3.

Análisis de redes de empresas de semiconductores y TSMC (2022)

Algoritmos Multi-gravedad Force Atlas (izquierda) y esquema circular (derecha)



Nota: El tamaño de los nodos está ponderado por la centralidad de los autovectores.

Fuentes: BBVA Research y www.gdelt.org

En definitiva, el papel de TSMC como proveedor clave de chips avanzados y su posición en la cadena de valor de semiconductores hacen que TSMC se convierta en un elemento de riesgo sistémico importante en la red para la industria de semiconductores y las economías avanzadas. Su interconexión con importantes actores globales refleja la complejidad y la interdependencia del sector dada la importancia de TSMC en la cadena de suministros globales y su papel crítico en la producción de tecnologías avanzadas. Estos vínculos son fundamentales para entender la dinámica del mercado y las posibles repercusiones de cualquier perturbación en la industria.

3.1.2. Análisis del lenguaje de los bancos centrales a partir de los comunicados de política monetaria

Además de las noticias globales, las técnicas de procesamiento de lenguaje natural permiten analizar cualquier texto, sea cual sea su naturaleza. Los comunicados de política monetaria de los bancos centrales se han convertido en una herramienta clave para controlar las expectativas de inflación y analizar dichos comunicados nos ayuda a entender mejor la estrategia de los mismos en sus decisiones de política monetaria y, por tanto, su impacto en la economía real.

Siguiendo la metodología aplicada por Hansen y McMahon (2016), analizamos el lenguaje de los bancos centrales sobre política monetaria a partir de sus comunicados de prensa o declaraciones, actas y discursos publicados en la web. Tras limpiar, transformar y procesar el texto convertido en vectores y resumido en la matriz documento-término, utilizamos los modelos de temas dinámicos basados en la asignación de Dirichlet Latente (*LDA*) que, como explicamos en la sección anterior, tienden a distribuir palabras en un conjunto reducido de temas para maximizar las probabilidades de las palabras de aparecer juntas para cada tema dado. Es un algoritmo no supervisado donde el investigador tiene que interpretar cada tema examinando la colección de palabras clave en cada componente. Una vez que tenemos identificados los temas, realizamos análisis de sentimiento basado en la aproximación del léxico, que consiste en analizar dentro de estos temas cuantas palabras positivas y negativas le acompañan de acuerdo a diccionarios asistidos para obtener un sentimiento medio del tema a lo largo del tiempo. Para ello, usamos el diccionario Loughran-McDonald (2011), que fue creado específicamente para analizar textos de índole financiera y el diccionario de la FED para la estabilidad financiera (Correa *et al.,* 2017). Estos diccionarios identifican palabras negativas y positivas, incluyendo palabras específicas de la jerga financiera y nos dan una mejor aproximación en el sentimiento que diccionarios creados a partir de lenguajes más generales.

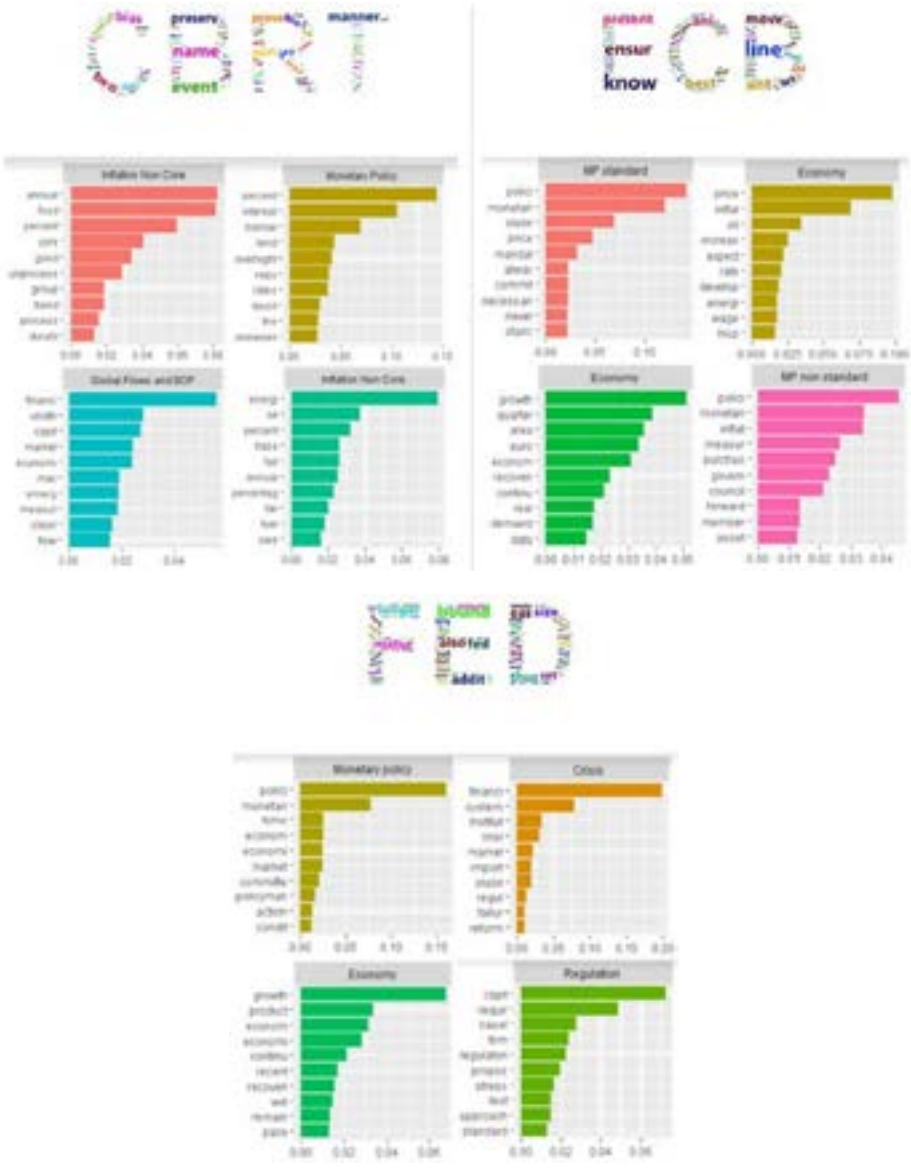
Con esta aproximación, analizamos los comunicados de varios bancos centrales, entre ellos el Banco Central Europeo (ECB, por sus siglas en inglés) o la Reserva Federal (FED) dentro de los países desarrollados y el Banco Central de Turquía (CBRT, por sus siglas en inglés) como país emergente durante los años 2016-2018. Para todos ellos obtenemos los principales temas identificados en sus comunicados como una composición de los términos relacionados más importantes y frecuentemente utilizados.

En la figura 4 podemos ver para el corpus analizado de cada banco central, qué temas emergen para cada uno de ellos. Términos de política monetaria como tasas de interés son un denominador común en todos los comunicados de bancos centrales, aunque existen matices relevantes dado que usan distintas herramientas de política monetaria. En el caso del BCE, destaca el uso de la política monetaria no estándar durante el periodo estudiado. En el caso del Banco Central de Turquía, destaca el tema de los flujos de capital por su importancia para las economías emergentes, especialmente tras la crisis financiera global.

Figura 4.

Principales temas identificados para el Banco Central de Turquía, Banco Central Europeo y la Reserva Federal (2016-2018)

Modelo estructural de temas basado en *LDA* para identificar los componentes



Fuentes: BBVA Research a partir de los comunicados e informes de política monetaria del CBRT, ECB y la FED.

Figura 5.

Evolución de los temas de política monetaria y sentimiento de los comunicados del Banco Central de Turquía a partir del análisis de texto

Modelo estructural de temas basado en *LDA* para identificar los componentes, la evolución de su peso en el tiempo y análisis de sentimiento

Fuentes: BBVA Research a partir de los comunicados e informes de política monetaria del CBRT.

Una vez identificados los temas, estudiamos su evolución en el tiempo, analizando cómo varía el peso o importancia de cada tema en cada comunicado, así como el sentimiento medio asociado. La figura 5 resume este análisis para el Banco Central de Turquía. El primer gráfico muestra cómo identificamos a partir de las técnicas de *LDA* y *STM* de qué se habla en las minutas del CBRT, mostrando tres ejemplos de temas identificados de política monetaria, inflación y actividad. De todos estos temas, gracias al modelo *STM,* podemos estudiar su evolución a lo largo del tiempo en cuanto al peso relativo de cada tema en los comunicados. Así por ejemplo, el segundo gráfico muestra la evolución de los temas de política monetaria en detalle, donde vemos cómo a lo largo del tiempo con la crisis financiera global, la política monetaria tradicional fue disminuyendo su importancia frente a las herramientas de política monetaria macroprudencial, que habían estado prácticamente inexistentes antes de la crisis. De todas estas series, tenemos, además, la evolución del sentimiento medio asociado. El tercer gráfico muestra el tono medio en todos los comunicados del CBRT e identificamos una estrecha relación del sentimiento con la postura del banco central, donde valores negativos van asociados a una posición de endurecimiento de la política monetaria y positivos con una flexibilización de la misma.

3.2. El uso de los informes corporativos como indicador de la evolución económica por sector de actividad

En el ámbito empresarial, la aplicación de las técnicas de PLN es de gran utilidad para analizar la información publicada por las empresas en sus informes trimestrales y anuales,

donde tienen que reportar su actividad y desempeño financiero, incluyendo datos financieros y resultados, información sobre segmentos de mercado, nuevos planes de productos e investigación y actividades de desarrollo en programas futuros. Es, por ello, que son una fuente de información de gran utilidad para entender el desempeño, perspectivas de futuro e interpretaciones del pasado de las empresas, así como los principales riesgos y oportunidades que se perciben en cada sector de actividad y, por ello, en la economía.

Con este fin, en BBVA Research analizamos los informes trimestrales y anuales de las empresas estadounidenses, las cuales están obligadas a presentar información veraz y detallada sobre su salud financiera a la SEC (Comisión de Bolsa y Valores de EE. UU.). La SEC pone a disposición pública todos los archivos electrónicos de estos informes corporativos a través del sistema de Recopilación, Análisis y Recuperación de Datos Electrónicos (EDGAR). Nos centramos en la información de las empresas del SP500, dado que son las 500 empresas de gran capitalización y cotizadas en la bolsa de valores de EE. UU. que cubren aproximadamente el 80 % del mercado de acciones estadounidense por capitalización.

Analizamos de ellas más de 21 millones de documentos desde 1994 a 2019 a los que les aplicamos dos enfoques diferentes para obtener información sobre los principales problemas abordados en el texto. En primer lugar, aplicamos análisis de temas utilizando la asignación de Dirichlet Latente como algoritmo de aprendizaje automático no supervisado para identificar temáticas en función de la relación de palabras en el texto. En segundo lugar, usamos la metodología *Word2Vec*, que es una red neuronal que obtiene las relaciones no lineales entre palabras para agruparlas de acuerdo a su contexto, por lo que nos dirá, dada una palabra o conjunto de palabras, el resto de las palabras similares que aparecen en el mismo contexto, generando un diccionario adaptado a las particularidades del texto analizado.

Finalmente, después de obtener los temas identificados y los diccionarios a partir de palabras clave, aplicamos el sentimiento utilizando el diccionario de Loughran y McDonald (2011) para obtener el tono o sentimiento promedio de los documentos.

Así, la combinación de ambas técnicas complementa la comprensión más profunda y detallada del texto. Por tanto, dejamos que los datos hablen para identificar de qué están hablando las empresas por sector de actividad y cómo están hablando a partir del sentimiento medio, y además identificamos, para términos estratégicos como el riesgo o la incertidumbre, cómo se relacionan en estos informes y su contexto.

Los principales temas identificados en los informes hacen referencia a sectores de actividad, dado que empresas del mismo sector tienden a comentar contextos similares. La figura 6 resume los resultados de los temas identificados con *STM* donde se destacan algunos de ellos con nubes de palabras en las que el tamaño y el color hacen referencia a la mayor probabilidad de la palabra de pertenecer a ese grupo. De esta forma, en la primera nube de palabras, los tokens de energía, electricidad o transmisor nos ayudan a identificar el sector energético. Del mismo modo, a partir de estos tokens, identificamos temas financieros, comercio al por menor, sector inmobiliario o automotriz entre otros.

Figura 6.

Principales temas identificados en los informes corporativos de las empresas SP500 (1994-2019)

Modelo estructural de temas basado en *LDA* para identificar temas

Fuentes: BBVA Research a partir de los informes corporativos de las empresas del SP500.

Para complementar este análisis, aplicamos la metodología *Word2Vec* para generar un conjunto de diccionarios con el fin de entender el contexto de términos económicos y financieros clave como la inflación, el crecimiento económico, la depreciación, el riesgo o la incertidumbre. La figura 7 muestra los resultados de este análisis. La distribución de palabras asociadas con cada término de interés se ilustra en el gráfico de dispersión (segundo gráfico). Los resultados muestran que hay una red significativa de interrelacio-

Figura 7.

Distribución de palabras en la generación de diccionarios y su evolución en el tiempo

Modelo *Word2Vec* para la generación de diccionarios de acuerdo a la proximidad entre palabras

Fuentes: BBVA Research a partir de los informes corporativos de las empresas del SP500.

nes entre palabras que hacen referencia al mismo contexto y, en general, las palabras clave escogidas están bien diferenciadas entre ellas, puesto que no hay muchas palabras mezcladas entre contextos. Inflación y depreciación guardan una mayor similitud, al igual que riesgo e incertidumbre. Sobre estos dos últimos términos, mostramos en la figura 7 la nube de palabras del diccionario creado para riesgo (primer gráfico) e incertidumbre (tercer gráfico), donde el tamaño (mayor) y el color (más oscuro) de las palabras hacen referencia a la proximidad (mayor) entre ellas. Vemos así qué riesgo está más asociado con términos financieros y de mercado, mientras que incertidumbre se relaciona más con preocupaciones y fragilidades económicas. Con estos diccionarios, podemos monitorizar con mayor robustez (dado que se han creado específicamente para el corpus analizado), la evolución de estas temáticas a partir de palabras clave a lo largo del tiempo, sectorialmente y por tipo de empresa, entendiendo mejor la heterogeneidad entre temas y sectores de actividad económica.

3.3. Monitorización de indicadores geopolíticos y políticos a partir de los medios de comunicación

En el campo geopolítico, el uso de estas técnicas de PLN se hace especialmente relevante dado que la información cuantitativa existente es escasa. El análisis de texto en las noticias nos permite rastrear múltiples eventos, proporcionando respuestas rápidas a preguntas cada vez más complejas en un mundo cada vez más fragmentado.

En BBVA Research llevamos casi una década utilizando estos métodos en el campo geopolítico para obtener respuestas en tiempo real y de forma muy granular y detallada. Así, hemos desarrollado desde indicadores de intensidad de conflictos, protestas, riesgo geopolítico y tensiones políticas, hasta mapas dinámicos de flujos migratorios con mucho detalle para ver el origen y destino de los migrantes durante la crisis humanitaria con la guerra de Siria.

Empezando por los indicadores de intensidad de protesta y conflicto, utilizamos la información de GDELT, donde cada evento se codifica según el sistema de codificación de eventos CAMEO (*Conflict and Mediation Event Observations*) desarrollado por Gerner *et al.* (2002). CAMEO es un esquema de codificación ampliamente utilizado para sistematizar el análisis de eventos políticos y sociales y dividirlos en una escala que va desde la cooperación material y verbal hasta el conflicto verbal y material. A partir de este sistema de CAMEO, identificamos todos los eventos relacionados con protesta y conflicto y monitorizamos la cobertura mediática de los mismos a través de GDELT.

De esta forma, construimos unos indicadores de intensidad de protesta y de conflicto, que capturan el volumen total de artículos de noticias por día que incluyen cualquier men-

ción de estos eventos identificados en CAMEO sobre protestas y conflictos. El número total de eventos de protestas y conflictos cada día y en cada país se divide por el número total de todos los eventos registrados por GDELT para ese día en ese país con objeto de elaborar un índice de intensidad por país. De esta manera, el indicador rastrea el nivel de prevalencia de la actividad de protesta y conflicto a lo largo del tiempo, corrigiendo por el aumento exponencial en la cobertura mediática en los últimos años y las diferencias de la cobertura mediática entre geografías.

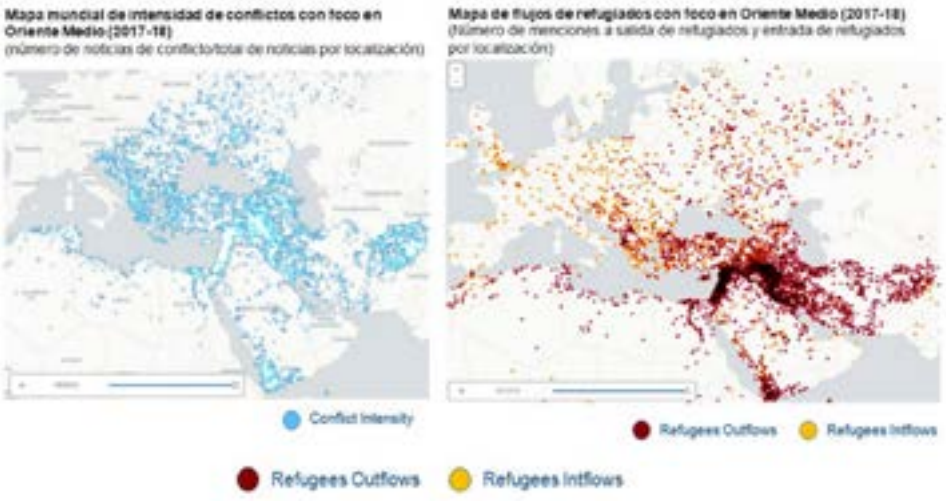
En los últimos años no hemos estado exentos de conflictos locales con efectos globales como la guerra de Siria, la guerra de Ucrania o la más reciente invasión de Hamas. Medir su evolución nos permite cuantificar su impacto social y económico.

En las figuras 8 y 9 podemos ver cómo hemos medido estos conflictos con distintos focos. En el caso de la guerra de Siria, construimos un mapa dinámico del indicador de intensidad de conflicto durante 2017 y 2018 con foco en Europa y Oriente Medio para identificar donde se localizan estos conflictos y cómo evolucionan en el tiempo (figura 8). El principal foco de conflicto se situaba en Siria e Irak, pero también destacan conflictos congelados como en Yemen y en Afganistán (primer mapa). Como consecuencia de la guerra de Siria, se produjo un éxodo masivo de refugiados con importantes implicaciones sociales y económicas para la región y para Europa. En el segundo mapa de la figura 8 mostramos

Figura 8.

Evolución de los conflictos en Oriente Medio y flujos migratorios a raíz de la Guerra de Siria (2017-2018)

Indicadores de intensidad de conflictos y de salida/entrada de refugiados



Fuentes: BBVA Research y www.gdelt.org

estos flujos migratorios donde en color rojo recogemos el volumen de noticias que hablan de salidas de refugiados en esa geografía y en amarillo la llegada de refugiados. Turquía fue el principal receptor de estos flujos migratorios con más de 3,5 millones de migrantes, como muestra el mapa, que refleja muy bien la trazabilidad de estos movimientos. La dinámica en el tiempo nos muestra el paso de los migrantes por Turquía, reflejando cómo una gran parte de ellos se quedaban en el país y otros continuaban su camino por los Balcanes hasta llegar a Europa Central.

La evolución temporal de los indicadores de conflicto por país y en frecuencia diaria permite identificar rápidamente cambios de tendencia como, por ejemplo, con la invasión rusa en Ucrania en febrero de 2022 o la reciente invasión de Hamas a inicios de octubre de 2023 (figura 9). En el caso de la guerra de Ucrania, el indicador tanto para Rusia como para Ucrania, muestra que el conflicto continúa vivo con valores de riesgo alto (por encima de una desviación típica con respecto a su media histórica desde 2019), sin signos de desescalada (primer gráfico). El aumento de las tensiones en los indicadores de conflicto ha sido especialmente significativo en el caso de Hamas (segundo gráfico), que además tiene implicaciones importantes para la región con riesgo de escalada e importantes efectos globales y económicos.

Aparte de conflictos y protestas, utilizando la taxonomía de GDELT, podemos rastrear otros temas geopolíticos, sociales y económicos como el riesgo geopolítico o las tensiones políticas. De esta forma, en BBVA Research desarrollamos un indicador de riesgo geopolítico basado en la metodología de Caldara e Iacoviello (2022) para iden-

Figura 9.

Indicador Big Data BBVA Research de intensidad de conflictos: Rusia, Ucrania y Oriente Medio (2017-2024)

Número de noticias de conflicto/total de noticias por geografía

Fuentes: BBVA Research y www.gdelt.org

tificar palabras clave relacionadas con el riesgo geopolítico como violencia, disturbios, relaciones exteriores[7],...

A partir de las noticias identificadas, generamos un índice basado en el tono y la cobertura relativa de estas palabras clave en la base de datos de GDELT de acuerdo a la fórmula [2] detallada en esta sección. De la misma manera, para las tensiones políticas, recopilamos todos los artículos de noticias que mencionan temas relacionados con política siguiendo la taxonomía de GDELT, como elecciones y campañas, partidos políticos y políticos, instituciones gubernamentales, políticas y escándalos políticos entre otros y construimos un indicador por país en frecuencia diaria de acuerdo a la misma metodología descrita en la fórmula [2].

Con todo ello, generamos un conjunto de indicadores que nos permiten monitorizar el panorama geopolítico en alta frecuencia y con gran precisión (figura 10), no solo identifi-

Figura 10.

Cuadro resumen de los principales indicadores *big data* geopolíticos de BBVA Research antes del conflicto Hamas – Israel y en la actualidad

Media móvil 28 días. Indicadores normalizados desde 2017

Fuentes: BBVA Research y www.gdelt.org

[7] El Índice de Riesgo Geopolítico de BBVA Research recoge la intersección de dos grupos de palabras clave. Las búsquedas deben contener al menos un tema incluido en la taxonomía GDELT de cada grupo (grupo 1 y grupo 2). El grupo 1 incluye los siguientes temas: guerra, conflicto, violencia revolucionaria, rebelión, disturbios violentos, mantenimiento de la paz, acuerdos de reconocimiento mutuo, alto el fuego, tratados, parlamento y legislaturas, militares, tropas, energía nuclear, energía hidroeléctrica, terror, rebeldes guerrilleros e insurgentes, secuestro, alianza, comité de resistencia popular de grupo, insurgencia, resistencia social de grupo, cooperación militar, armada y rebeldes entre los términos más relevantes. El grupo 2 incluye los siguientes temas: acto perjudicial, anunciante, riesgo, preocupación en todo el mundo, especie en peligro, crisis, problema, disputas, procedimientos de despido, boicot, perturbación, refuerzo militar, sanciones, bloqueo, vulnerabilidad y riesgos financieros, cuarentena, ultimátum, declaración, brote, anunciador, armero, persecución, choque, redada, conflicto armado, acto de fuerza, amenaza de bomba, asesinato, huelga. Véase Caldara e Iacoviello (2022) para mayor detalle.

cando puntos de tensión, sino posibles efectos contagio entre geografías o sectores. Su seguimiento en frecuencia diaria nos permite contar con una herramienta de alerta temprana para entender que está pasando en el mundo. En el primer gráfico de la figura 10 mostramos cómo estaban estos indicadores antes de la invasión de Hamas como referencia al conflicto armado más reciente, y en el gráfico de la derecha una imagen más actualizada. La figura muestra que se ha producido un aumento generalizado de los focos de conflicto y el riesgo geopolítico, acompañado en el caso de Oriente Medio y en algunos países latinoamericanos de un aumento de las tensiones políticas. El descontento social y el aumento de incertidumbre de política económica se mantiene contenido hasta el momento.

4. CONCLUSIONES

La digitalización y la interacción social en tiempo real con dispositivos móviles, ordenadores, redes sociales y plataformas digitales ha generado ingentes volúmenes de datos no estructura- dos (textos, videos, imágenes, voz...). Gracias al avance de los algoritmos de procesamiento de lenguaje natural y las técnicas de inteligencia artificial, estos datos pueden convertirse en información estructurada para el análisis. Además, la disminución significativa de los costes de cómputo y procesamiento de la información y el desarrollo de modelos y algoritmos para tratar estos datos, han revolucionado la forma de hacer analítica en las ciencias sociales.

En este capítulo hacemos un seguimiento a la evolución de los algoritmos de análisis de texto, desde la bolsa de palabras a los grandes modelos de lenguaje, destacando sus ventajas y desafíos y enumerando ejemplos de cómo se han utilizado en economía. Finalmente, mostramos cómo en BBVA Research utilizamos estas técnicas de procesamiento de lenguaje natural para el análisis económico, financiero, social y geopolítico.

En economía, analizamos el lenguaje de bancos centrales, temas clave difícilmente cuantificables con datos tradicionales como la incertidumbre de política económica a partir de los medios de comunicación o la crisis de los semiconductores. En el ámbito financiero, mostramos el análisis de las empresas estadounidenses del SP500 para estudiar la evolución sectorial y términos económicos y financieros clave. Finalmente en el campo geopolítico, mostramos indicadores en tiempo real para monitorizar conflictos, protestas, riesgo geopolítico y tensiones políticas.

Todo ello muestra el potencial del análisis de texto en las ciencias sociales, que alcanzan su máximo exponente en la actualidad con los grandes modelos de lenguaje (*LLM*), que comprenden y analizan en profundidad el lenguaje humano. No obstante, existen desafíos relevantes que condicionarán su desarrollo y viabilidad futura como la importancia de asegurar un uso justo de los mismos, sin sesgos y éticamente correctos, que además tienen que ir acompañados del desarrollo de marcos regulatorios que garanticen la privacidad de los usuarios, la seguridad de los datos y la penalización de un mal uso del contenido generado

por los mismos. Si conseguimos hacer frente a estos desafíos, las oportunidades de futuro y el potencial de la inteligencia artificial es incalculable para el análisis, donde la imaginación es el límite.

Referencias

Ash, E. y Gennaro, G. (2023). Emotion and Reason in Political Language. *The Economic Journal, Volume 133,* Issue 650.

Ash, E. y Hansen, S. (2023). Text Algorithms in Economics. *Working Paper.*

Baker, S. R., Bloom, N. y Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics, 131*(4), pp.1593–1636. Oxford University Press.

Blei, D. y *Lafferty, J.* (2006). *Dynamic Topic Models. ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning, 2006.* 113-120. 10.1145/1143844.1143859.

Blei, D. M., Ng, A. Y. y Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*(null), pp. 993–1022.

Bloom, N., Hassan, T., Kalyani A., Lerner, J. y Tahoun, A. (2021). The diffusion of disruptive technologies. *CEP Discussion Papers,* dp1798. Centre for Economic Performance, LSE.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan,A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G.,Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen,M., Sigler, E., Lit- win, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. y Amodei, D. (2020). Language Models are Few-Shot Learners. *In Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc.

Bybee, L., Kelly, B., Manela, A. y Xiu, D. (2023). Business News and Business Cycles. *Journal of Finance.* Forthcoming.

Caldara, D. e Iacoviello, M. (2022). Measuring Geopolitical Risk. *American Economic Review, 112*(4), pp. 1194-1225.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P. *et al.* (2022). PaLM: Scaling language modeling with Pathways. arXiv preprint arXiv:2204.02311.

Correa, R., Keshav, G., Londono-Yarce, J. M. y Nathan M. (2017). Constructing a Dictionary for Financial Stability. IFDP Notes. Washington: Board of Governors of the Federal Reserve System, June 2017. https://doi.org/10.17016/2573-2129.33

Dabravolski, V., Dredze, M., Gehrmann, S., Irsoy, O., Kambadur, P., Lu, S., Mann, G., Rosenberg, D. y Wu, S. (2023). *BloombergGPT: A Large Language Model for Finance.*

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. y Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), pp. 391–407.

Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota. Association for Computational Linguistics.

Gemini Team Google (2023). Gemini: A Family of Highly Capable Multimodal Models Gemini Team. *Working Paper.*

Gerner, D., Jabr, R. y Schrodt, P. (2002). *Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions.*

Hansen, S., Lambert, P. J., Bloom, N., Davis, S. J., Sadun, R. y Taska, B. (2023a). Remote Work across Jobs, Companies, and Space. *Working Paper.*

Hansen, S., Lundgaard, A. y Kazinnik, S. (2023b). Can ChatGPT Decipher Fedspeak? *Working Paper.*

Hansen, S. y McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, 99, S114–S133.

Hansen, S., McMahon, M. y Prat, A. (2018). Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics,133*(2), pp. 801–870.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99* (pp. 289–296). San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Korinek, A. (2023). Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature, 61*(4), pp. 1281-1317.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. y Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.arXiv preprint arXiv:1907.11692.

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

Loughran, T. y McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research, 54,* issue 4, p. 1187-1230.

Mikolov, T., Chen, K., Corrado, G. y Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. y Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546.

OpenAI. (2023). GPT-4 Technical Report.

Pennington, J., Socher, R. y Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in naturallanguage processing* (*EMNLP*) (pp.532–1543).

Phuong, M. y Hutter, M. (2022). Formal Algorithms for Transformers. *Working paper.*

Radford, A., Narasimhan, K., Salimans, T. y Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. *Working Paper, OpenAI Blog.*

Roberts, M. E., Stewart, B. M. y Airoldi, E. M. (2013). Structural Topic Models. *Working paper.*

Shapiro, A. H., Sudhof, M. y Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics, 228*(2), pp. 221–243.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. y Polosukhin, I. (2017). Attention is All you Need. In *Advances in NeuralInformation Processing Systems,* volume 30. Curran Associates, Inc.