Identificación y seguimiento de riesgos y vulnerabilidades en el sistema financiero mexicano a través de redes sociales y noticias

Extracto del Reporte de Estabilidad Financiera – Diciembre de 2024, Recuadro 1, pp. 19 – 23. Documento publicado el 11 de diciembre de 2024.

1. Introducción

Los bancos centrales, autoridades financieras, y organismos internacionales, entre otros, han sustentado el análisis de las condiciones financieras y económicas en modelos econométricos calibrados con datos estadísticos cuya recopilación y procesamiento requiere un gran esfuerzo operativo, por lo que su publicación, en ciertas ocasiones, se realiza con un retraso de semanas, meses, o incluso años.[1]

Los últimos años se han caracterizado por un incremento sin precedentes en la cantidad, diversidad y disponibilidad de datos, tanto por la popularización del internet como por el surgimiento de nuevos canales de comunicación; por ejemplo, los medios digitales y las redes sociales. Dicho incremento en la generación de datos coincide y, a su vez, ha incentivado el desarrollo de tecnología, *software* y herramientas analíticas para procesar datos (de fuentes y formatos más diversos y menos estandarizados), en mayor volumen y menor tiempo.[2]

En particular, los datos de texto publicados en internet, en artículos de noticias y redes sociales, son una fuente de información útil para analizar los canales y dinámicas de transmisión de la información y las expectativas de la población en relación con temas específicos, así como para identificar posibles fuentes de riesgo.

En este Recuadro se destacan algunos de los proyectos que el Banco de México ha desarrollado con base en la explotación de fuentes de datos no estructurados,[3] principalmente texto, y la implementación de metodologías de Aprendizaje Automático (*ML*, por sus siglas en inglés) e Inteligencia Artificial (IA), con el propósito de optimizar la identificación y el monitoreo de sucesos relevantes que podrían impactar en la estabilidad financiera del país. En particular, se abordan las metodologías de Procesamiento de Lenguaje Natural (*NLP*, por sus siglas en inglés) utilizadas para el aprovechamiento de datos extraídos de sitios *web* de noticias y *X* (antes *Twitter*), los cuales permiten obtener de manera oportuna indicadores de alta frecuencia.

En particular, se describen el índice de Incertidumbre de Política Económica (*EPU*, por sus siglas en inglés) y algunas de sus variantes, y el índice de riesgo bancario. Estos índices se elaboran con diversas fuentes de información y se actualizan constantemente, en respuesta a la velocidad con que se presentan cambios en la información. Por un lado, el índice *EPU* tiene por objetivo identificar y monitorear aquellos factores que pudieran impactar de forma negativa la estabilidad y las expectativas económicas en México, con origen y posible afectación en cualquiera de sus sectores. Por otro lado, el índice de riesgo bancario se enfoca en el monitoreo

[1] *BIS* (2022a).
[2] En particular, durante los últimos años, destaca la extensión en el uso de Inteligencia Artificial (IA), a través de metodologías variadas, con distintos objetivos y grados de complejidad.
[3] Los datos estructurados tienen un formato estandarizado que permite tanto a un software como a una persona acceder y explotar estos datos fácilmente. Generalmente se trata de formatos tabulares, con filas y columnas que definen claramente los atributos de los datos. En contraste, los datos no estructurados son aquellos que no poseen un modelo ni orden predefinidos, por lo cual se requieren herramientas de análisis más complejas; algunos ejemplos son archivos de texto, video, imágenes, etc.

del sector bancario específicamente, con el fin de identificar señales tempranas de posibles riesgos o vulnerabilidades que pudieran afectar la estabilidad de dicho sector o del sistema financiero en su conjunto.

2. Antecedentes

Las metodologías de análisis de datos no estructurados no son nuevas y abarcan un amplio espectro de herramientas y aplicaciones. Por ejemplo, en Baker et al. (2016) se propone la creación del índice *EPU*, para explotar la frecuencia de publicación de noticias relacionadas con incertidumbre, política y economía. Asimismo, en Cookson et al. (2023) se analiza el papel de las redes sociales durante el periodo de corridas bancarias en 2023, tomando como referencia información publicada en *X*. Por otro lado, a través del Proyecto *Ellipse*, el Banco de Pagos Internacionales (*BIS*, por sus siglas en inglés) desarrolla un sistema de alertas tempranas que incorpora el monitoreo y análisis de fuentes de datos no estructurados, incluidas noticias, para identificar riesgos y vulnerabilidades que pudieran afectar a las instituciones financieras.

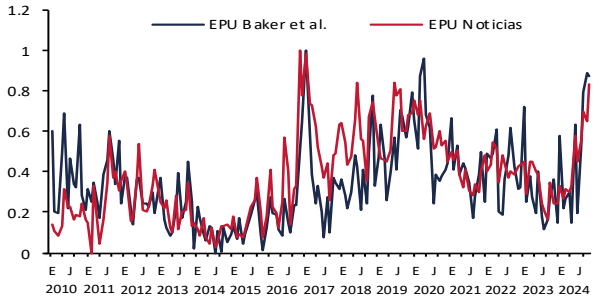
En términos generales, las técnicas de *NLP* permiten el análisis de la relación semántica del texto, mediante la detección de patrones de léxico, para extraer información útil.

3. Índice de incertidumbre de política económica y sus refinamientos

Como punto de partida, el Banco de México ha desarrollado metodologías basadas en la construcción de diccionarios con términos clave. Por ejemplo, se construyeron diccionarios para el análisis de sentimiento de documentos públicos, como el Reporte de Estabilidad Financiera, o bien, para el cálculo de nuevas versiones de índices *EPU* propios.[4]

A partir del primer semestre de 2024, en este *Reporte* se incluye la actualización de un índice *EPU* calculado con base en noticias obtenidas de sitios *web* de periódicos,[5] el cual representa la proporción de noticias relacionadas con incertidumbre en política económica respecto al total de noticias publicadas, cuyos valores son normalizados entre 0 y 1, para facilitar su interpretación (Gráfica 1).

Gráfica 1
Índice de Incertidumbre de Política Económica
Nivel de incertidumbre



Cifras a noviembre de 2024.
Fuente: Banco de México.

El índice *EPU* puede, a su vez, descomponerse para analizar cómo es que la incertidumbre local responde a los niveles de incertidumbre en otros países desarrollados y emergentes, o a factores idiosincráticos.

[4] A partir del *Reporte* publicado en junio de 2020, el Banco de México da seguimiento al nivel de incertidumbre en política económica a través de la actualización de índices *EPU* propios, construidos con base en información de *X*, así como de los sitios *web* de noticias.
[5] Mediante técnicas de *web scraping*, se extrajo el contenidoB de las noticias en los sitios *web* de distintos periódicos, en secciones relacionadas con economía, finanzas, mercados y negocios. Para la lista de fuentes, ver Sección III. Actualización de retos del entorno económico, evolución de la incertidumbre, en este *Reporte*.

Para capturar mayor información, el contenido de los diccionarios se ha ampliado mediante la identificación de términos similares con base en el modelo *Word2Vec*,[6] que representa las palabras en un espacio vectorial, mediante el entrenamiento de una red neuronal, tomando en cuenta el contexto en el que aparecen; lo anterior, con el objetivo de extender la cobertura de los diccionarios e incrementar la precisión de los resultados.[7]

Asimismo, es posible identificar los temas principales asociados con la incertidumbre. Para ello, se implementan técnicas no supervisadas para la segmentación por grupos de conjuntos de texto, en los que se identifican los temas principales a los que pertenece cada conjunto de texto. Por ejemplo, el algoritmo *Latent Dirchlet Allocation* (*LDA*),[8] se puede emplear en el análisis de noticias relacionadas con temas específicos (p. ej. elecciones en otras jurisdicciones, comercio internacional, etc.) con el fin de identificar, asociado a esos temas específicos, temas relevantes para México. Otro ejemplo es el uso del modelo *BERTopic*,[9] el cual se puede emplear para la desagregación e identificación de los factores de incertidumbre que influyen en el nivel del índice *EPU*, con la ventaja de que este incorpora un Modelo de Lenguaje a Gran Escala, permitiendo la representación del texto en su conjunto para contextualizar los datos.[10] Haciendo uso del modelo *BERTopic*, Banco de México ha desarrollado metodologías que permiten obtener una clasificación automatizada de las noticias que componen el índice *EPU*, optimizando la detección de los determinantes de incertidumbre en México.

4. Índice de riesgo bancario

El índice de riesgo bancario se enfoca en el monitoreo del sector bancario dada su relevancia para la salud del sistema financiero en su conjunto.[11]

El objetivo de este índice es dar seguimiento a la percepción de riesgo entre especialistas, en relación con el sector bancario, mediante el análisis de publicaciones en *X* (tuits), dada la popularidad de esta red social entre medios de comunicación, personajes relevantes en el sector y el público en general. Esto se logra mediante el análisis de sentimiento del contenido de los tuits, según tres metodologías distintas.

Datos

Para calcular el índice, se elabora una lista de cuentas verificadas pertenecientes a periódicos de circulación nacional y extranjeros, agencias de prensa y de televisión, y agencias calificadoras,[12] y se realiza la extracción de tuits publicados en dichas cuentas. Se conservan únicamente aquellos tuits que contienen el nombre de algún

[6] Mikolov et al., (2013).

[7] Para la construcción del índice *EPU*, tanto los diccionarios de referencia originales como sus ampliaciones, se enfocan en términos en español, dado que las fuentes de datos corresponden a periódicos de circulación nacional.

[8] LDA es un modelo generativo para la modelación de tópicos en un conjunto de textos. Partiendo del cálculo de la frecuencia de las palabras en cada documento, se infiere la probabilidad de que cada palabra del vocabulario y cada artículo, en el conjunto de datos, pertenezcan a un grupo específico, asignando cada elemento de acuerdo con la mayor probabilidad de pertenencia (Blei et al., 2003).

[9] La arquitectura *BERT* (*Bidirectional Encoder Representations from Transformers*) permite obtener directamente la representación vectorial no solo de palabras, sino también de documentos. *BERTopic* incorpora el modelo *Sentence-BERT* (Reimers y Gurevych, 2020), un reentrenamiento del modelo *BERT*, el cual mejora la calidad de la vectorización de textos y reduce significativamente el tiempo de procesamiento para las tareas de cálculo de similitudes y la clasificación de noticias. Dicho modelo está calibrado para el procesamiento de texto en distintos lenguajes, incluido el español.

[10] Grootendorst (2022).

[11] Elaborado con base en el trabajo de Rho et al. (2021).

[12] En línea con Rho et al. (2021), se seleccionaron cuentas específicas para basar el análisis en fuentes confiables, las cuales podrían tener influencia en la percepción del público sobre el sector bancario. En particular:

- Periódicos de circulación nacional: @eleconomista, @Reforma, @reformanegocios, @Milenio, @lajornadaonline, @Excelsior, @elsolde_mexico, @El_Universal_Mx, @diario24horas, @LaRazon_mx, @CapitalMexico, @Reporte_Indigo, @heraldodemexico, @lacronicadehoy, @sdpnoticias.
- Periódicos extranjeros: @el_pais, @elpais_america, @Forbes_es, @Forbes_Mexico).
- Agencias de prensa y televisión: @AP_Noticias, @AFPespanol, @XHespanol, @EFEMexico, de televisión @bbcmundo.
- Agencias calificadoras: @MoodysLatAm, @FitchLatAm.

banco, o bien, mencionan los términos “banco(a)” o “bancario(a)”. Posteriormente, en el texto se aplica una limpieza y lematización,[13] y se anonimizan los bancos.

Metodología

El cálculo del índice de riesgo bancario se basa en la clasificación de los tuits, de acuerdo con la polaridad de su contenido, a partir de un algoritmo basado en la votación de tres clasificadores distintos:

- Diccionario. Construido por el Banco de México, con base en el trabajo de Loughran y McDonald (2016).
- *B4MSA* (*Baseline for Multilingual Sentiment Analysis*). Es un clasificador de sentimiento multilenguaje basado en un modelo de *Máquina de Soporte Vectorial* (*SVM*, por sus siglas en inglés), para clasificación.
- *ULMFIT* (*Universal Language Model Fine-Tuning for Text Classification*). Está basado en una red neuronal entrenada para clasificación de textos financieros.

Cada modelo asigna una de tres posibles puntuaciones: -1 (polaridad positiva), 1 (negativa) y cero (neutral). La clasificación final para cada tuit se obtiene como la moda de los resultados individuales, asignando un sentimiento neutral si no se determina.

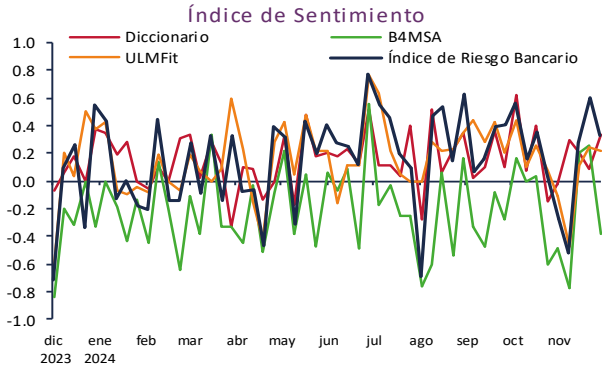
Finalmente, el índice de riesgo bancario se calcula como el promedio semanal de la clasificación final de los tuits, obtenida con base en los tres clasificadores, con valores que oscilan en el intervalo $[-1,1]$, manteniendo una relación directa entre el nivel del índice y la percepción de riesgo.

Resultados

De acuerdo con el índice obtenido, durante el periodo que abarca este *Reporte*, el nivel de estrés en la banca múltiple se mantuvo mayoritariamente en niveles ligeramente por encima del promedio histórico (Gráfica 2).

Gráfica 2

Índice de Riesgo Bancario Semanal



Cifras al 01 de diciembre de 2024.
Fuente: Banco de México.

Asimismo, se identifican los principales factores que contribuyen al nivel del índice. Al alza: fallas en los servicios digitales de algunas instituciones, interrupciones en los servicios debido a la temporada de huracanes, incremento en la incertidumbre política y el riesgo de una reducción en la calificación crediticia de México. A la baja: resultados positivos en las utilidades, particularmente en la banca digital, y la inclusión de nuevos participantes en el sector.[14]

[13] La lematización es la técnica que permite asignar a cada palabra de una frase su lema, es decir, la forma canónica bajo la cual nos la encontraríamos en un diccionario.

[14] Durante la semana del 25 de noviembre al 01 de diciembre se observa una divergencia en el indicador asociado con la metodología de Diccionario, en comparación con el índice de riesgo bancario. Esto se debe a que, en ciertos casos, la existencia de términos clasificados con polaridad negativa conduce

5. Consideraciones finales

La relevancia de las redes sociales y los medios de comunicación digitales como canales de difusión de información masiva, va en aumento. Estos medios actúan como amplificadores de información, poniéndola a disposición de millones de personas en cuestión de segundos. Por una parte, esto representa beneficios para la identificación oportuna de fuentes de riesgo potenciales. Por otra parte, dicha capacidad de difusión representa una fuente de riesgo, dado que podría amplificar o inducir reacciones de los mercados y de la población en general. Por ello, es importante que las autoridades lleven a cabo un seguimiento constante sobre los temas en tendencia relacionados con el sistema financiero y la forma en que cambios en la información pueden anticipar o inducir eventos que tengan un impacto en la estabilidad financiera.

Adicionalmente, si bien se percibe que este tipo de fuentes de información ofrece beneficios, tales como mayor disponibilidad y frecuencia de actualización en los datos, es importante destacar también la posibilidad de que se presenten intermitencias o dificultades en el acceso a las plataformas en cuestión, por lo cual sería recomendable incorporar medidas de contingencia ante cualquier eventualidad.[15] El Banco de México evalúa constantemente cómo hacer uso de nuevas tecnologías y metodologías que permitan explotar de forma óptima el acceso a datos de alta frecuencia, con el fin de identificar oportunamente vulnerabilidades y amenazas a la estabilidad financiera del país. Ejemplos de ello son el índice *EPU* y el índice de riesgo bancario, los cuales se complementan entre sí y permiten dar seguimiento a los riesgos en el sistema financiero. En particular, de acuerdo con la literatura, el incremento en la incertidumbre económica podría influir directamente en la percepción de riesgo y comportamiento del sector bancario (Wu et al., 2020), lo cual destaca la importancia de realizar el análisis de ambos indicadores en conjunto. Esto, a su vez, permitiría profundizar en otros estudios, como el análisis de la incidencia que tiene la incertidumbre en el apetito por riesgo de los bancos y en el otorgamiento de crédito, así como en las expectativas sobre el deterioro de los activos debido al incremento en la probabilidad de incumplimiento de los deudores.

6. Referencias

Baker, S., Bloom, N. & J. Davis (2016): "*Measuring Economic Policy Uncertainty*". The Quarterly Journal of Economics, Volume 131, noviembre 2016.

Bank of International Settlements (*BIS*) (2019): "*The use of big data analytics and artificial intelligence in central banking*", IFC Bulletin, No. 50, mayo 2019.

BIS (2022a): "*Ellipse: regulatory reporting and data analytics platform*", marzo 2022. Disponible en: https://www.bis.org/about/bisih/topics/suptech_regtech/ellipse.htm

BIS (2022b): "*Machine learning in central banking*", IFC Bulletin, No. 57, noviembre 2022.

Blei, D., Ng, A. & M. Jordan (2003): "*Latent Dirichlet Allocation*", Journal of Machine Learning Research No. 3, enero 2003.

Cookson, J., Fox, C., Gil-Bazo, J., Imbet, J. & C. Schiller (2023): "*Social Media as a Bank Run Catalyst*", Université Paris-Dauphine, Research Paper No. 4422754, junio 2023.

Devlin, J., Chang, M., Lee, K. & K. Toutanova (2019): "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*". Google AI Language, mayo 2019.

a la asignación de tuits en esta categoría, en contraposición al sentimiento que trasmite el texto en conjunto. Lo anterior, permite resaltar una ventaja del algoritmo de votación, el cual tiene por objetivo robustecer el modelo e incrementar la precisión de los resultados.

[15] Por ejemplo, durante 2023 se experimentaron dificultades en la actualización del índice *EPU* elaborado con información de *X*, debido a modificaciones en los esquemas de suscripción a la plataforma y acceso a las herramientas de descarga de datos. Por ello, como medida de contingencia, se construyó un índice *EPU* alternativo, elaborado con información de los portales de noticias, el cual se publica en este *Reporte* a partir del primer semestre de 2024. Esta medida permite disponer de información actualizada, reduciendo la dependencia de un proveedor de datos específico. Adicionalmente, cuando no se puede acceder a una fuente de datos alterna, es posible calibrar modelos de *ML* e IA para obtener estimaciones robustas para los periodos en los cuales no se cuenta con toda la información, con base en información histórica y otras variables auxiliares.

Grootendorst, M. (2022): "*BERTopic: Neural topic modeling with a class-based TF-IDF procedure*", marzo 2022.

Loughran, T. & McDonald, B., (2016): "*Textual Analysis in Accounting and Finance: A Survey*", Journal of Accounting Research, Vol. 54, septiembre 2016.

Mikolov, T., Chen, K., Corrado, G. & J. Dean (2013): "*Efficient Estimation of Word Representations in Vector Space*", septiembre 2013.

Rho, C., Fernández, R. & B. Palma (2021): "*A Sentiment-based Risk Indicator for the Mexican Financial Sector*", Banco de México Working Paper series, abril 2021.

Yerushalmy, J. (2023): "*The First Twitter-Fuelled Bank Run': How Social Media Compounded SVB's Collapse*", The Guardian, marzo 2023.

Reimers, N. & I. Gurevych (2020): "*Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*". Technische Universität Darmstadt, octubre 2020.

Wu, J., Yao Y., Chen, M. & B. Jeon (2020): "*Economic uncertainty and bank risk: Evidence from emerging economies*", Journal of International Financial Markets, Institutions & Money, vol. 68, agosto 2020.