RedNIE
Red Nacional de Investigadores en Economía

Medición de Incertidumbre Económica en Redes Sociales en Base a Modelos de Procesamiento de Lenguaje Natural

J. Daniel Aromí (IIEP UBA-Conicet/UCA)

DOCUMENTO DE TRABAJO N° 179

Septiembre de 2022

Citar como:
Aromí, J. Daniel (2022). Medición de Incertidumbre Económica en Redes Sociales en Base a Modelos de Procesamiento de Lenguaje Natural. *Documento de trabajo RedNIE N°179*.

Medición de incertidumbre económica en redes sociales en base a modelos de procesamiento de lenguaje natural

J. Daniel Aromí[1]

CAEE, FCE, UCA
IIEP-Baires, UBA-Conicet

Primera versión: 05/04/2022
Esta versión: 05/06/2022

Resumen:
Este trabajo propone un índice que describe las opiniones económicas transmitidas por usuarios argentinos en la red social Twitter. Luego de identificar mensajes económicos, éstos son clasificadas según la frecuencia con la que se utilizan palabras asociadas a incertidumbre. La evaluación cualitativa del índice sugiere un fuerte vínculo con eventos económicos y políticos de relevancia. Estimaciones de modelos estadísticos indican que el índice contiene información sobre el ciclo económico, la confianza del consumidor y la evolución del mercado cambiario. Análisis complementarios demuestran que el foco en el concepto de incertidumbre y el uso de técnicas de procesamiento de lenguaje natural constituyen elementos clave para el desempeño satisfactorio de este indicador de opiniones.

1. Introducción

La trayectoria de una economía está gobernada, en parte, por las expectativas y creencias de los diversos actores económicos. Los grandes volúmenes de información en redes sociales ofrecen la oportunidad de generar medidas novedosas y detalladas sobre la evolución de las opiniones económicas. Al mismo tiempo, es importante observar que no existe una receta clara sobre cuál es la forma más conveniente de resumir las grandes cantidades de mensajes intercambiados día a día. Por ejemplo, ¿cuáles son las manifestaciones que proveen más información sobre el estado de la economía? ¿En qué forma se pueden utilizar las herramientas de procesamiento de lenguaje natural?

Al enfrentar esta cuestión, vale observar que la teoría económica y la evidencia empírica indican que la percepción de incertidumbre es un aspecto clave que impacta en las decisiones de consumo de las familias, la inversión por parte de empresas y las condiciones en los mercados financieros (Bernanke 1983, Stiglitz y Weiss 1981, Bloom 2009, Orlik y Veldkamp 2014, Baker et al. 2016). La incertidumbre afecta especialmente las inversiones de largo plazo, como aquellas en infraestructura. El incremento de la incertidumbre aumenta el costo financiero o tasa de descuento y, por lo tanto, frena planes de expansión. De esta manera, la percepción de incertidumbre emerge como uno de los aspectos clave a ser medidos utilizando los grandes volúmenes de información en redes sociales.

[1] Se agradecen los valiosos comentarios de Lionel Barbagallo, Daniel Heymann, Ann Mitchell, Carlos Newland, Mariano Rabassa, Patricia Saporiti y las sugerencias recibidas en el Seminario de Economía de la UCA. Este trabajo fue en parte posible gracias a múltiples trabajos de investigación previos desarrollados junto a Sergio De Raco y Martín Llada.

Este trabajo presenta y evalúa un indicador de opiniones económicas en Argentina en base a los contenidos de la red social Twitter. El indicador se enfoca en la medición de manifestaciones asociadas a incertidumbre. Con este objetivo, se usan los contenidos de Twitter para entrenar modelos de aprendizaje no supervisado. Más específicamente, se entrena un modelo de representación vectorial de palabras para identificar palabras asociadas al concepto de "incertidumbre". Luego, se construye un índice que mide la frecuencia de este conjunto de palabras. Este ejercicio puede ser interpretado como la estimación de un factor latente que contiene información sobre el estado de las opiniones económicas y, más ampliamente, sobre el estado de la economía.

Una de las ventajas de la metodología aquí propuesta es que, si bien utiliza herramientas de procesamiento de lenguaje natural, de todas maneras, el indicador resultante puede ser interpretado con naturalidad. En una primera etapa del procesamiento de textos, se utiliza una lista de palabras clave para identificar mensajes relacionados con economía. Esta lista es generada con la asistencia de un modelo de detección de temas (Blei et al. 2003) entrenado con mensajes compartidos en Twitter por usuarios argentinos. En una segunda etapa, se computa la frecuencia de palabras asociadas a incertidumbre. Esta lista corresponde a las 100 palabras más cercanas a la palabra "incertidumbre" según un modelo de representación vectorial de palabras (Pennington et al. 2014) que fue entrenado con mensajes económicos emitidos por usuarios argentinos.

La evaluación cualitativa del índice resultante sugiere que éste provee información valiosa sobre la trayectoria de las opiniones económicas y sobre su vínculo con las cambiantes condiciones de la economía. Complementariamente, análisis formales indican que el índice anticipa información sobre los niveles de actividad económica, la confianza del consumidor y los mercados de activos financieros. Por ejemplo, un modelo simple de vectores autorregresivos indica que un shock de una desviación estándar en el indicador es seguido, en promedio, por una caída acumulada de 0,6% en el nivel de actividad tres meses después.

Adicionalmente, análisis formales muestran que tanto el foco en el concepto "incertidumbre" como la utilización de herramientas de procesamiento de lenguaje natural son elementos importantes para el desempeño satisfactorio del indicador. Es decir, cuando se utilizan diccionarios de palabras predefinidos comúnmente utilizados para medir el nivel de positividad de las opiniones o cuando se utiliza un diccionario compuesto exclusivamente por las pocas palabras cuya raíz coincide con la de "incertidumbre" se verifican asociaciones que son menos intensas, menos robustas o, directamente, inexistentes.

Este trabajo contribuye a la literatura de procesamiento de texto aplicados a contextos económicos (Gentzkow et al. 2019). Más específicamente, este ejercicio está vinculado con un conjunto de trabajos que proponen indicadores de opiniones económicas útiles para el análisis macroeconómico (Tetlock 2007, Baker et al. 2016, Shapiro et al. 2020, Aromí 2020, Bybee et al. 2021). Dentro de este grupo de contribuciones, existen antecedentes que consideraron el caso de la red social Twitter y implementaron medidas de incertidumbre económica (Baker et al. 2021, Becerra et al. 2021, Istat 2022). A diferencia de estos trabajos, buscando obtener un índice más informativo, en este trabajo se utilizan técnicas de procesamiento de lenguaje natural. En parte, este ejercicio toma elementos que fueron propuestos en Aromí (2017&2020) para la generación de un indicador en base a los contenidos del diario económico The Wall Street Journal.

En la próxima sección se detallan los datos utilizados y la metodología implementada. En la sección 3 se describe en forma cualitativa la información capturada por el índice. La sección 4 reporta el contenido informativo del índice a partir de la estimación de modelos estadísticos y

reporta los resultados observados ante cambios en la metodología. En la última sección se presentan algunas discusiones a modo de conclusión.

2. Datos y metodología

El indicador es computado a partir de mensajes compartidos en la red social Twitter por usuarios cuyo vínculo con Argentina es determinado a partir de la ubicación reportada por el usuario. Aproximadamente, los mensajes procesados constituyen una muestra que representa el 1% de los mensajes compartidos.

El diseño del indicador responde a dos consideraciones. Por un lado, es evidente que los modelos de procesamiento de lenguaje natural constituyen una herramienta muy valiosa para la extracción de información. Entre otros beneficios, estas herramientas permiten representar datos de muy alta dimensionalidad en una forma estructurada con una dimensionalidad órdenes de magnitud menor. Por otro lado, resulta de interés implementar metodologías que faciliten la interpretación de la información transmitida por el índice. En base a estas consideraciones, se propone un índice que utiliza modelos de procesamiento de lenguaje natural de una forma transparente que puede ser comunicada y comprendida sin dificultades.

Más específicamente, el cómputo del indicador es implementado en dos etapas: selección de tweets y clasificación de contenidos. En la primera etapa, se seleccionan los tweets económicos a partir de un conjunto de palabras clave que son identificadas con la asistencia de un modelo de detección de temas (LDA, Blei et. al. 2003). En una segunda etapa, se mide la frecuencia de palabras relacionadas con incertidumbre en el subconjunto de tweets económicos. Esta lista de palabras es identificada a través de un modelo de representación vectorial de palabras (GloVe, Pennington et al. 2014).

2.1 Datos

El índice reportado en este documento es construido a partir de una colección de 210 millones de mensajes compartidos en la red social Twitter para el período 2011-2022. Estos mensajes corresponden a usuarios que indican como localización “Argentina” en su perfil de usuario. La fuente principal de estos mensajes es la muestra de 1% de tweets que es distribuida en tiempo real a través de la API de Twitter. Complementariamente, esta colección fue extendida usando nuevamente la API de Twitter para recuperar mensajes enviados por una selección aleatoria de usuarios argentinos que fueron identificados usando la muestra inicial. El conjunto de mensajes en la base de datos representa, aproximadamente, el 1% de los mensajes compartidos por los usuarios argentinos de Twitter.

2.2 Selección de mensajes económicos

La selección de mensajes económicos es implementada a través de una lista de palabras económicas. Todo mensaje que incluye al menos una palabra de esta lista es seleccionado. La lista es generada combinando la información provista por un modelo de temas (LDA, Blei et al. 2003) y el juicio de experto.

La primera etapa en este proceso involucra entrenar el modelo de temas LDA (Blei et al. 2003). Según este modelo generativo, cada tema puede ser visto como una distribución de

probabilidad de palabras. Dado un conjunto de documentos, el modelo es entrenado para estimar las distintas distribuciones de probabilidad que, con mayor probabilidad, pudieron haber generado los documentos. De esta manera, el modelo identifica una estructura que genera información que puede ser utilizada para clasificar textos. El modelo no genera etiquetas para cada tema estimado, es por eso que, luego de estimar el modelo, el analista necesita inspeccionar las palabras más frecuentes en cada tema para asignar etiquetas (por ej. "economía").

Para el entrenamiento del modelo se construyeron documentos a partir de los mensajes compartidos por 20000 usuarios argentinos. Para cada usuario, se construyó un documento concatenando una secuencia de tweets recientes. Se usó una cota superior de 12000 caracteres para cada documento. La estimación del modelo exige especificar un parámetro que determina el número de temas y un vocabulario. El número de temas elegido fue 50 y el vocabulario correspondió al conjunto de 4000 palabras más frecuentes en la colección de documentos.

Siguiendo la práctica usual con este tipo de modelos, para cada tema estimado, se inspeccionó las palabras cuya frecuencia es particularmente alta. De esta manera, se logró identificar un tema asociado a discusiones económicas. En el caso de este tema, se observa que las 100 palabras con frecuencia particularmente alta[2] incluyen palabras cuya raíz es "econ" (económica, economía …) y palabras de uso generalizado en discusiones económicas (mercado, precios, inflación, cobrar, pagar, producir, servicios, producción). Complementariamente, se hallan palabras asociadas a discusiones más específicas como: finanzas (deuda, crédito, bono, tarjeta), moneda (usd, dólar/es, pesos), trabajo (empleo, empleador, sueldo, laboral), finanzas públicas (impuesto/s), empresas (empresas, comercio, negocio/s) y distribución del ingreso (pobres, pobreza, ricos). Este conjunto de 100 palabras representativos del tema economía fue seleccionado para definir una primera lista de palabras clave.

Complementariamente, buscando lograr construir una lista más exhaustiva de términos económicos, se generó una segunda lista de palabras. Para ello, en primer lugar se computó la frecuencia de palabras en el conjunto de tweets que incluían alguna de las 100 palabras clave identificadas en el paso anterior. Luego, se seleccionaron las palabras cuya frecuencia es alta en este subconjunto de tweets versus la frecuencia observada en el total de tweets de la base de datos. De esta manera se obtuvo una segunda lista de 100 palabras.

Finalmente, esta lista de palabras fue combinada con la lista original proveniente del modelo LDA y el nuevo conjunto expandido fue inspeccionada para definir la lista definitiva en función de juicio de experto. La lista final contiene 153 palabras clave y es reportada en el apéndice A.

De esta manera, detectando la presencia de palabras clave de esta lista definitiva, se identificaron 7 millones de mensajes sobre economía. Es decir, aproximadamente 3,5% de los mensajes fueron clasificados como económicos. La siguiente nube de palabras ilustra el contenido de este conjunto de mensajes:

[2] Dado un tema, para identificar las palabras características, se computó la razón entre la probabilidad de cada palabra condicional al tema en cuestión y la probabilidad no condicional de cada palabra. Palabras características de un tema son aquellas para las que, en términos relativos, el ratio toma valores altos.

Gráfico 1: Nube de palabras en base a 7 millones de mensajes económicos

La visualización sugiere que un diverso conjunto de temas económicos son cubiertos por este conjunto de mensajes. Una inspección de estas palabras sugieren que entre los aspectos discutidos se encuentra el consumo (comprar, quiero, compra, ropa), el trabajo (trabajar, trabajando, sueldo, trabajo, trabajadores), las finanzas (deuda, dólares, dólar, tarjeta), las políticas económicas (gobierno, ministro, economía, impuestos) y cuestiones sociales (pobreza, pobres).

2.3 Estimación de la incertidumbre

Siguiendo la metodología propuesta en Aromí (2017, 2020), la incertidumbre manifestada en los mensajes económicos es estimada computando la frecuencia de palabras relacionadas con incertidumbre como proporción del total de palabras. Esta lista es identificada midiendo distancia entre palabras por medio de un modelo de representación vectorial de palabras.

Más específicamente, en este trabajo se estima o entrena el modelo GloVe (Pennington et al. 2014) usando el conjunto de 7 millones de tweets económicos. Este modelo genera una representación numérica de palabras en base a las estadísticas de co-ocurrencia, es decir, a partir de la cantidad de veces que una palabra aparece en la vecindad de otra palabra. Estas estadísticas de co-ocurrencia tiene una muy alta dimensionalidad. Por ejemplo, comúnmente se trabaja con vocabularios con decenas de miles de palabras y, por lo tanto, cientos de millones de pares de palabras cuya co-ocurrencia es representada a través de una matriz dispersa. El modelo GloVe implementa una forma conveniente de reducción de dimensionalidad y, de esta forma, facilita tareas como la identificación de palabras

relacionadas. La distancia entre las palabras es comúnmente computada utilizando la similitud del coseno.[3]

Para la estimación del modelo se definió un vocabulario con las palabras que, en el conjunto de 7 millones de mensajes económicos, tienen una frecuencia mayor o igual a 100. Este vocabulario contiene 38957 palabras.[4] Luego de estimar el modelo se computó la distancia entre las representaciones vectoriales de palabras y se recuperaron las 100 palabras más cercanas a "incertidumbre". La lista de palabras es reportada en el apéndice 2 y el gráfico 2 provee una visualización en la que el tamaño de las palabras está asociado a la cercanía con "incertidumbre".

Gráfico 2: Nube de palabras asociadas a incertidumbre

La nube de palabras muestra los diversos aspectos que, según las estadísticas de co-ocurrencia, están vinculados al concepto de incertidumbre. No sorprendentemente, el primer rasgo que se destaca al inspeccionar la lista es la presencia de palabras que refieren a problemas o percepciones negativas. Por ejemplo, se encuentran palabras que indican dificultades económicas (recesión, crisis, volatilidad, inflación, inestabilidad) y palabras que manifiestan emociones negativas (ira, miedo, temor, angustia, tristeza, desesperación, impotencia). Entre otras palabras con contenido negativo encontramos también a: destrucción, injusticia, agrava, destrucción, desconfianza, conflictividad y caos. Sin embargo, también encontramos palabras que no tienen un contenido negativo pero que, de todas maneras, están vinculadas a escenarios inciertos. Dentro de este conjunto de palabras, se destaca un grupo que demuestra cómo la incertidumbre está asociada a discusiones de coyuntura económica (panorama, coyuntura, actualidad, contexto, situación, escenario, optimismo, política, cambiaria, solvencia). Según esta inspección informal, es posible

[3] Para una descripción más detallada de la forma en que se implementa y se interpreta el modelo GloVe ver Aromí (2020).

[4] La estimación del modelo GloVe requiere especificar el valor de un conjunto de parámetros. En esta implementación, los vectores de palabras tienen dimensión 100, la ventana para el cómputo de co-ocurrencias es de tamaño 10 y el parámetro x_max, que impone una cota al peso de palabras muy frecuentes, fue fijado igual a 100.

conjeturar que la lista de palabras obtenida permite estimar un factor latente con información valiosa sobre el estado de la economía.

Dado un período de tiempo y el conjunto de mensajes correspondientes a ese mes, el índice de incertidumbre es definido como la razón entre la frecuencia de las 100 palabras más cercanas a "incertidumbre" y el número total de palabras en ese conjunto de mensajes. Formalmente el índice satisface la siguiente expresión:

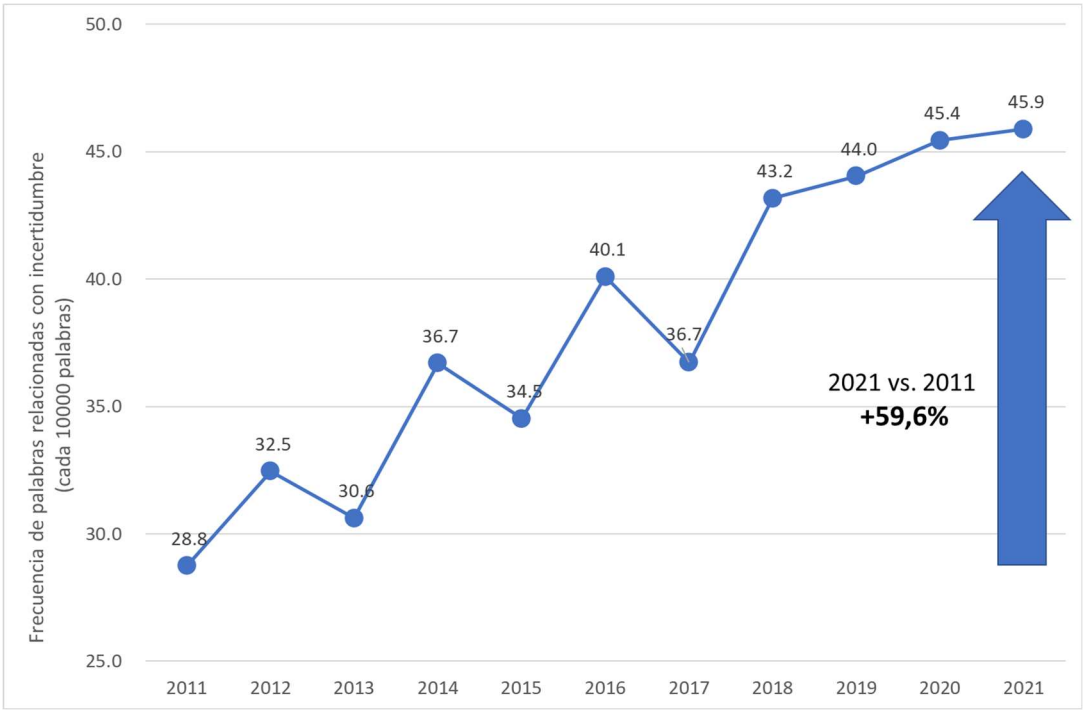
$$I_t = \frac{\#\ palabras\ relacionadas\ con\ incertidumbre}{\#\ total\ de\ palabras} * 10000$$

Esta especificación constituye una forma sencilla y transparente de resumir el contenido de las discusiones económicas en redes sociales. En las siguientes secciones evaluamos la información provista por este índice, y especificaciones alternativas, a través evaluaciones cualitativas y estimaciones de modelos estadísticos.

3. Una evaluación cualitativa

Adoptando una perspectiva de largo plazo, el índice brinda una representación inquietante de la trayectoria de la economía durante los últimos años (ver gráfico 3). El valor del índice en el año 2011 fue 28,8. Es decir, de cada 10000 palabras, 28,8 estuvieron asociadas a incertidumbre. Luego de transitar una trayectoria con tendencia ascendente, en el año 2021, el valor del indicador fue 45,9. Es decir, se produjo un aumento de un 59,6%. Vale observar que, entre el año 2011 y el año 2018, el índice muestra importantes oscilaciones. Este patrón está asociado a la "maldición de los años pares". Esta es la manera en que se denominó al patrón de años eleccionarios (impares) relativamente benévolos seguidos por años pares en los que se experimentaba un empeoramiento en las condiciones económicas.

Gráfico 3: Indice de incertidumbre económica – anual

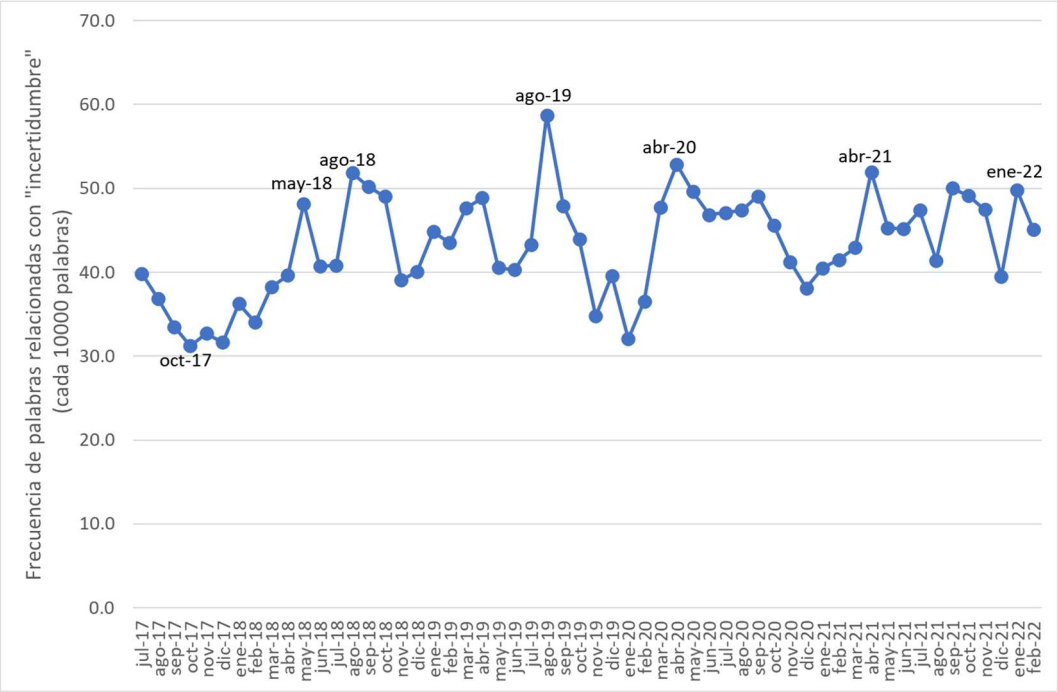


La medición mensual del estimador revela el fuerte vínculo entre el índice y eventos económicos y políticos de relevancia (ver gráfico 4). En particular, se observan incrementos en el índice durante períodos de agudas crisis. Por ejemplo, el índice muestra picos en mayo y agosto de 2018. Estos meses estuvieron marcados por fuertes devaluaciones e importantes incrementos en el diferencial de tasa de interés de la deuda del gobierno argentino. El valor más alto del índice en el período muestral corresponde a agosto 2019. En este mes se realizaron las primarias presidenciales en las que la oposición obtuvo un amplio éxito y se observó una severa respuesta negativa en los mercados financieros.

El índice muestra también importantes incrementos asociados a la crisis COVID. De esta manera, en marzo y abril de 2020 se observan valores cercanos a 50. Estos son valores elevadores en base a lo observado en los años previos. Es interesante observar que, pese a la severidad de la crisis sanitaria, económica y social disparada por la pandemia, los niveles estimados de incertidumbre económica durante este período no exceden aquellos observados en momentos críticos de los años 2018 y 2019.

Por último, se puede destacar el pico de abril de 2021. Este alto valor puede ser explicado por el reporte difundido en ese mes en que se informó un sorprendentemente elevado valor del índice de inflación de marzo 2021 (4,8%).

Gráfico 4: Indice de incertidumbre económica - mensual



4. Evaluaciones formales del contenido informativo

En esta sección se reportan estimaciones de modelos empíricos que permiten caracterizar la información provista por el índice. Para estos ejercicios se consideran indicadores de actividad económica, el índice de confianza del consumidor y la volatilidad en el mercado cambiario paralelo.

Más específicamente, se estiman modelos parsimoniosos que consisten en un proceso autorregresivo de orden 1 que es extendido para incorporar el indicador de incertidumbre económica. El indicador de incertidumbre económica es expresado como la diferencia entre el valor del mes correspondiente y el valor promedio de los últimos 12 meses. Es decir, la especificación propuesta satisface:

$$\hat{I}_t = I_t - \frac{1}{12}\sum_{k=1}^{12} I_{t-k}$$

Esta transformación cumple dos objetivos. Por un lado, contempla la posible no estacionariedad de la serie. Complementariamente, suaviza la medida de cambio al comparar el valor más reciente con el valor promedio observado durante una ventana de 12 meses.

Los modelos propuestos estiman tanto asociaciones contemporáneas entre la medida de incertidumbre y variables económicas de interés como la medida en la que el índice anticipa valores futuros de estas variables económicas. Para el caso de las asociaciones contemporáneas, el modelo utilizado satisface:

$$y_t = \alpha y_{t-1} + \beta \hat{I}_t + u_t$$

Donde y_t es el valor de la variable económica de interés para el mes t, $\hat{I}_t$ es el valor del índice de incertidumbre ajustado con respecto a la historia reciente y u_t es un término aleatorio de error. El valor estimado para β indica la asociación entre el índice de incertidumbre y el valor esperado de la variable económica de interés. Similarmente, el caso de la asociación entre el índice de incertidumbre y valores futuros de la variable económica de interés es estudiado a través del siguiente modelo parsimonioso de pronóstico:

$$y_{t+1} = \alpha y_t + \beta \hat{I}_t + u_{t+1}$$

Para medir el nivel de actividad se utilizan dos variables disponibles con frecuencia mensual: el Estimador Mensual de la Actividad Económica[5] y el Índice de Producción Industrial Manufacturero[6]. En ambos casos se trabajó con la variación mensual desestacionalizada. La confianza del consumidor es evaluada a partir del Índice de Confianza del Consumidor de la Universidad Torcuato Di Tella.[7] Esta variable es expresada en términos de su variación mensual. Por último, las condiciones del mercado cambiario son medidas a partir de la volatilidad del mercado del dólar paralelo (o blue).[8] Este indicador es expresado como la raíz cuadrada retorno diario cuadrado medio.

Los modelos son estimados para el período 2013-2019. Debido a la desmesurada envergadura y origen extraeconómico del shock Covid, se eligió excluir del análisis el período posterior al año 2019. La tabla 1 muestra, para cada caso, los valores estimados para el coeficiente del índice de incertidumbre. Para brindar más información sobre estas asociaciones, en el primer par de columnas se reportan las estimaciones correspondientes al modelo en que se elimina el rezago de la variable económica respectiva.

Los coeficientes estimados sugieren que el índice de incertidumbre brinda información valiosa sobre las condiciones económicas contemporáneas. Adicionalmente, se observa que el índice anticipa información sobre las condiciones económicas en períodos futuros. Por ejemplo, considerando los pronósticos del modelo autorregresivo extendido, se encuentra que un aumento de una desviación estándar en el índice de incertidumbre anticipa una caída de 0.47% en el crecimiento del IPIM en el mes siguiente. Similarmente, se encuentra que este incremento en el índice de incertidumbre anticipa una caída promedio de 2% en el indicador de confianza del consumidor. En el caso de la volatilidad en el mercado cambiario paralelo, se encontraron asociaciones contemporáneas, pero no se encontraron asociaciones estadísticamente significativas con la volatilidad del mes siguiente.

[5] https://www.indec.gob.ar/indec/web/Nivel4-Tema-3-9-48
[6] https://www.indec.gob.ar/indec/web/Nivel4-Tema-3-6-14
[7] https://www.utdt.edu/ver_contenido.php?id_contenido=2575&id_item_menu=4982
[8] https://www.ambito.com/contenidos/dolar-informal-historico.html

Tabla 1: Asociaciones con variables económicas

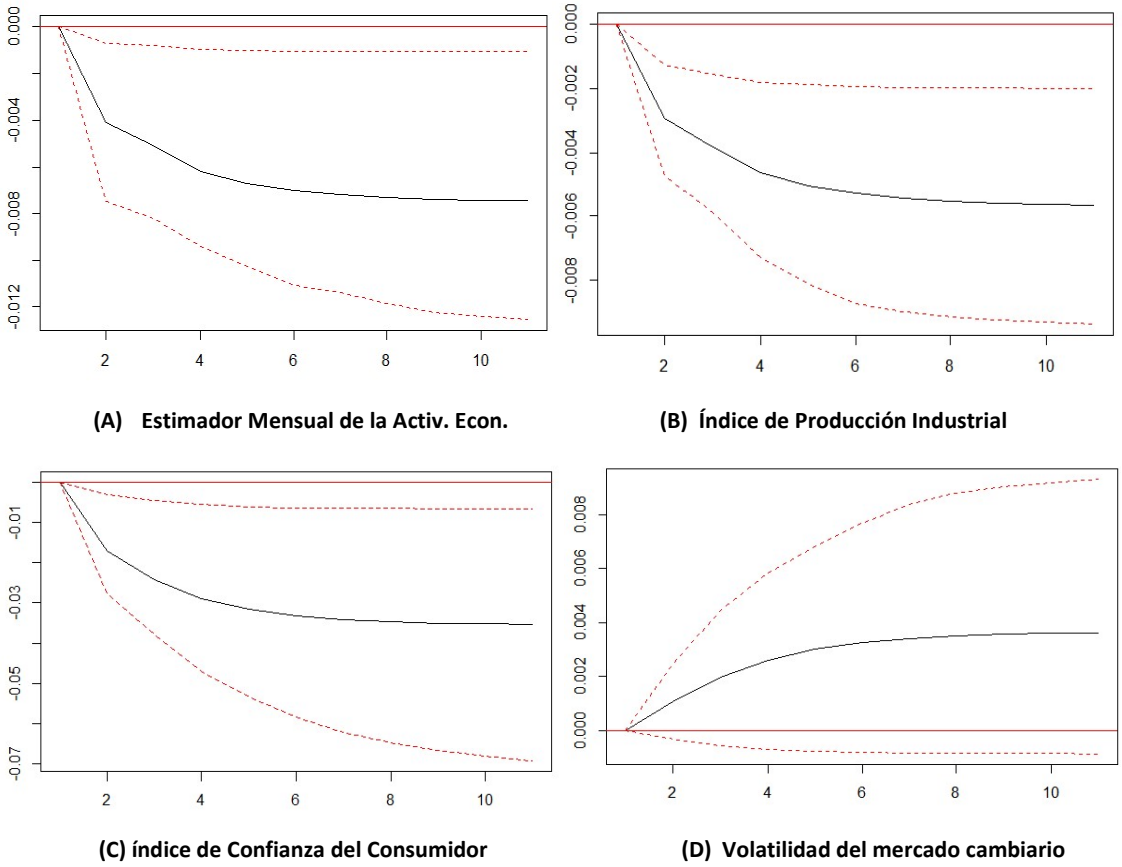
	Modelo sin rezago		AR(1) extendido	
	nowcast	**pronóstico**	**nowcast**	**pronóstico**
EMAE				
$\hat{\beta}$	-0.0023 ***	-0.0026 ***	-0.0029 ***	-0.0031 ***
t	3.1	-3.5	-2.8	-3.6
IPIM				
$\hat{\beta}$	-0.0026 *	-0.0039 **	-0.0036 **	-0.0047 **
t	2	2.4	2.3	2.6
Indice de Conf. del Cons.				
$\hat{\beta}$	-0.0214 ***	-0.0172 ***	-0.0233 ***	-0.0201 **
t	2.7	3	-2.6	2.9
Volatilidad dólar blue				
$\hat{\beta}$	0.0044 ***	0.0017	0.0042 ***	0.0008
t	3	1.6	2.9	0.07

Notas: las estimaciones corresponden al período 2013-2019. El índice de incertidumbre fue estandarizado. Los errores estándar fueron computados de acuerdo a Newey-West (1987,1994). Niveles de significatividad: * 10%, ** 5%, *** 1%. En los casos del EMAE, IPIM y el Indice de Confianza, la variable explicada es la tasa de variación anual. En el caso de la volatilidad dólar, la variable es expresada en niveles.

Una descripción más expresiva de las asociaciones dinámicas puede ser provista a través de un parsimonioso modelo de vectores autoregresivos. Consideramos modelos con dos variables y un rezago. En cada caso, el primer componente del vector corresponde a la variable económica de interés y el segundo componente corresponde al indicador de incertidumbre. Este ordenamiento es usado para realizar una descomposición de Cholesky en la que se asume que los shocks en el índice de incertidumbre pueden afectar contemporáneamente la variable económica de interés pero los shocks en la variable económica impactan en el indicador de incertidumbre con un retraso de un período. Como en los ejercicios anteriores, el indicador de incertidumbre es expresado como la diferencia entre el valor actual y el valor promedio de los últimos 12 meses. El gráfico 5 muestra para cado uno de los cuatro casos la función de impulso respuesta ante un shock en el índice de incertidumbre.

Las funciones impulso-respuesta estimadas confirman los resultados previos. En el caso de las medidas de actividad económica, un shock al indicador de incertidumbre es seguido, en promedio, por una caída en los niveles de crecimiento acumulado en los siguientes meses. Los impactos son económicamente significativos. Tres meses después del shock se observa en promedio una caída acumulada de 0.6% o 0.7% en el crecimiento económico. En el caso del nivel de confianza del consumidor también se observa una caída económica y estadísticamente significativa. La excepción está dada por el mercado cambiario. En este caso, no se encuentra un vínculo estadísticamente significativo.

Gráfico 5: Funciones impulso-respuesta ante un shock en el indicador de incertidumbre



(A) Estimador Mensual de la Activ. Econ.

(B) Índice de Producción Industrial

(C) índice de Confianza del Consumidor

(D) Volatilidad del mercado cambiario

Notas: Cada función impulso respuesta corresponde a modelos bivariados en los que se asume que la variable económica de interés no responde contemporáneamente a shocks en el índice de incertidumbre. Los intervalos de confianza de 95% fueron computados por Bootstrap. En los casos del EMAE, IPIM y el Indice de Confianza, la variable explicada es la tasa de variación anual. En el caso de la volatilidad dólar, la variable es expresada en niveles.

4.1 Evaluación de especificaciones alternativas

No existe un consenso sobre cuál es la forma más conveniente de resumir las opiniones económicas en redes sociales. En parte esto se debe a que el propósito del ejercicio puede variar. Por ejemplo, el objetivo podría involucrar generar un estimador del valor contemporáneo o futuro de un indicador de opiniones pre-existente (una encuesta de consumidores). Alternativamente, el objetivo podría involucrar generar un factor que resuma información sobre el estado de la economía (un factor macroeconómico) o un predictor que reduzca la incertidumbre sobre algún resultado económico de interés.

Adicionalmente, otro motivo para la falta de consenso tiene que ver con que existen diversas estrategias metodológicas cuyas implicancias son desconocidas. Por ejemplo, existen estrategias que computan indicadores a partir de un número pequeño de palabras clave mientras que otros indicadores son construidos a partir de grandes diccionarios de palabras.

Otro aspecto a tener en cuenta es la utilización de diccionarios pre-establecidos versus el caso en que se identifica un diccionario de palabras a partir de propiedades del conjunto de documentos a ser procesados. Adicionalmente, existen múltiples atributos o aspectos de las opiniones que podrían ser objetivo de estimación. Una opción tradicional en este sentido tiene que ver con la estimación de la polaridad (positividad o negatividad) expresada por un conjunto de textos. Alternativamente se pueden considerar atributos más específicos como la incertidumbre o emociones transmitidas por los mensajes.

La metodología aquí propuesta define un conjunto relativamente grande de palabras clave. Estas palabras son identificadas a través de un modelo que es entrenado utilizando un conjunto de documentos que coincide con los textos que son procesados para generar la medida de incertidumbre. Esta coincidencia puede ser valiosa ya que permite identificar el significado de la palabra en el contexto relevante. Finalmente, en este documento se propone poner el foco en la medición del nivel de incertidumbre transmitida por los mensajes. Considerando que es de interés documentar el impacto de estas elecciones metodológicas, en esta sección se evalúan los resultados ante distintos cambios en la forma en que se construye el índice de opiniones económicas. Estos cambios involucran el tamaño del diccionario, la selección de diccionarios de alternativos y métodos alternativos de agregar la frecuencia de palabras individuales.

En primer lugar, se considera el número de palabras utilizadas para medir la incertidumbre. En vez de elegir las 100 palabras más cercanas, se evalúan los resultados cuando se elige un número más chico (50) y un número más grande (200) de palabras. Estas evaluaciones sugieren que 100 palabras es una cantidad adecuado y que los resultados son robustos ante cambios en este aspecto metodológico. La tabla 2 muestra que los resultados no cambian en forma apreciable cuando el número de palabras asociadas a incertidumbre es duplicado o reducido a la mitad.

Tabla 2: Análisis de robustez – Número alternativo de palabras

	50 palabras		200 palabras	
	nowcast	**pronóstico**	**nowcast**	**pronóstico**
EMAE				
$\hat{\beta}$	-0.0021 **	-0.31 ***	-0.0029 ***	-0.003 ***
t	2.6	3.1	3.3	3.6
IPIM				
$\hat{\beta}$	-0.0018	-0.0048 **	-0.0036 **	-0.0042 **
t	0.9	2.4	2.2	2.3
Indice de Conf. del Cons.				
$\hat{\beta}$	-0.0183 ***	-0.0128 **	-0.0253 ***	-0.0251 ***
t	3.1	2	3.1	3.3
Volatilidad dólar blue				
$\hat{\beta}$	0.0047 ***	0.0001	-0.004 **	0.0009
t	4.1	0.1	2.5	0.7

Notas: las estimaciones corresponden al período 2013-2019. El índice de incertidumbre fue estandarizado. Los errores estándar fueron computados de acuerdo a Newey-West (1987,1994). Niveles de significatividad: * 10%, ** 5%, *** 1%. En los casos del EMAE, IPIM y el Indice de Confianza, la variable explicada es la tasa de variación anual. En el caso de la volatilidad dólar, la variable es expresada en niveles.

En segundo lugar, consideramos desviaciones más importantes con respecto a la metodología anteriormente propuesta. En vez de entrenar un modelo de lenguaje natural para identificar palabras clave se pueden elegir diccionarios pre-definidos que capturen algún aspecto de las opiniones. A continuación evaluamos dos listas de palabras que miden la polaridad de los mensajes. En primer lugar consideramos una versión traducida al español de la tradicional lista de palabras negativas utilizadas, por ejemplo, por Tetlock (2007) en su trabajo seminal sobre contenido de los diarios y retornos en el mercado de capitales. También consideramos la lista de palabras positivas y negativas provista por Molina-Gonzalez et. al. (2013). Por último se consideró una alternativa en que se utiliza un conjunto pequeño de palabras asociadas a incertidumbre. Esta estrategia vincula a la propuesta implementada en el influyente trabajo de Barker y otros (2016). Este índice que sólo computa la frecuencia de un conjunto pequeño de palabras que empiezan con “incertidumbre” o “inciert”.

Como muestra la tabla 3, en los tres casos se encuentra que los resultados empeoran al considerar estas listas alternativas de palabras. Es decir, la evidencia sugiere que el uso de diccionarios identificados a través de modelos entrenados con el conjunto de documentos a procesar y el foco en el concepto de incertidumbre son atributos convenientes a la hora de construir un índice de opiniones económicas.

Tabla 3: Evaluación de diversos diccionarios pre-definidos

	Negatividad (General Inquirer)		ISOL (Pos - Neg.)		Incertidumbre/inciert…	
	nowcast	pronóstico	nowcast	pronóstico	nowcast	pronóstico
EMAE						
$\hat{\beta}$	-0.0016 **	-0.0003	0.0009	0.15	-0.0012	-0.0014
t	2.5	0.3	1.5	1.5	1.1	1
IPIM						
$\hat{\beta}$	-0.0033 *	0.0005	0.0041 **	0.0027 *	-0.0029 *	-0.0013
t	1.9	0.3	2.4	1.9	1.7	1
Indice de Conf. del Cons.						
$\hat{\beta}$	0.0027	0.0069	-0.0056	0.0015	-0.0144 ***	-0.0105
t	0.5	0.9	0.6	0.3	-2.9	1.1
Volatilidad dólar blue						
$\hat{\beta}$	0.0006	0.0016	-0.0013	-0.0024 ***	0.0051 ***	-0.0017
t	0.6	1.9	1.5	2.9	3.8	1.5

Notas: las estimaciones corresponden al período 2013-2019. El índice de incertidumbre fue estandarizado. Los errores estándar fueron computados de acuerdo a Newey-West (1987,1994). Niveles de significatividad: * 10%, ** 5%, *** 1%. En los casos del EMAE, IPIM y el Indice de Confianza, la variable explicada es la tasa de variación anual. En el caso de la volatilidad dólar, la variable es expresada en niveles.

Finalmente, consideramos el impacto de formas alternativas de agregar la información provista por la frecuencia de las distintas palabras en el diccionario. Según la especificación presentada, el índice es igual a la suma de las frecuencias de cada palabra. Esta especificación es particularmente conveniente en términos de interpretación. Es decir, de acuerdo a esta definición, el índice reporta la cantidad de palabras relacionadas con incertidumbre como proporción del total de palabras. Sin embargo, es posible que esta especificación le asigne excesivo peso a algunas palabras más frecuentes. Adicionalmente, este tipo de agregación es potencialmente frágil frente a valores extremos que puedan tomar algunas palabras debido a motivos accidentales no relacionados por las condiciones macroeconómicas que se intenta aproximar.

En última instancia, el impacto de distintas formas de agregación es una cuestión empírica que debe ser evaluada. Tres formas alternativas de agregar la frecuencia de palabras son consideradas. Una de estas formas de agregación consiste en computar la mediana. Esta opción podría ser ventajosa ya que elimina valores extremos e ignora las palabras más frecuentes. Otra opción que podría proveer los mismos beneficios con un mayor nivel de expresividad involucra, simplemente, eliminar las palabras más frecuentes. Está claro que estas estrategias también involucran pérdida de información por lo que el desempeño resultante es, en principio, incierto. Finalmente, una tercera alternativa considerada involucra computar la suma del logaritmo de la frecuencia como proporción del total.

La tabla 4 muestra el desempeño de los indicadores asociados a cada una de las formas alternativas de agregan la información a nivel de palabras. Se observa que, la tercera opción, la suma del logaritmo de la frecuencia, resulta la opción que consistentemente muestra una mayor capacidad para capturar información contemporánea y futura sobre los niveles de actividad, la confianza del consumidor y el estado de mercado cambiario.

Tabla 4: Especificaciones alternativas de la función de agregación

	Mediana		Eliminar las 5 más frecuentes		Suma del log de la frecuencia	
	nowcast	**pronóstico**	**nowcast**	**pronóstico**	**nowcast**	**pronóstico**
EMAE						
$\hat{\beta}$	-0.003 ***	-0.002 **	-0.0025 ***	-0.0034 ***	-0.004 ***	-0.0023 **
t	5.3	-2.6	3	3	4.3	2.6
IPIM						
$\hat{\beta}$	-0.003 **	-0.004 **	-0.0043 **	-0.0057 ***	-0.004 ***	-0.0058 ***
t	2.2	2.3	2.6	2.9	2.7	3.2
Indice de Conf. del Cons.						
$\hat{\beta}$	-0.009	-0.012	-0.0167 **	-0.0082	-0.015 **	-0.0096
t	1.4	1.6	2.5	1.2	2.5	1.6
Volatilidad dólar blue						
$\hat{\beta}$	0.0034 ***	0.014	0.0042 ***	0.0006	0.0039 ***	0.0008 ***
t	2.9	1.6	2.8	0.5	3.8	0.4

Notas: las estimaciones corresponden al período 2013-2019. El índice de incertidumbre fue estandarizado. Los errores estándar fueron computados de acuerdo a Newey-West (1987,1994). Niveles de significatividad: * 10%, ** 5%, *** 1%. En los casos del EMAE, IPIM y el Indice de Confianza, la variable explicada es la tasa de variación anual. En el caso de la volatilidad dólar, la variable es expresada en niveles.

5. Discusión

Este trabajo propone un método para resumir las opiniones económicas de usuarios de la red social Twitter. La estrategia se concentra en la medición de la incertidumbre y utiliza herramientas de procesamiento de lenguaje natural. Una característica importante del índice es su fácil interpretación. La evidencia sugiere que el indicador es informativo de las opiniones económicas y del estado de la economía. Esta conclusión se desprende de análisis cualitativos y cuantitativos considerando distintas frecuencias y distintos indicadores económicos. De esta forma, este indicador emerge como una novedosa forma de monitorear opiniones que se complementan con otros indicadores tradicionales como las encuestas de opinión, los pronósticos de profesionales y los precios de activos financieros.

En análisis extendidos se muestra que tanto el foco puesto en el concepto de incertidumbre como la utilización de herramientas de procesamiento de lenguaje natural constituyen elementos importantes para el desempeño satisfactorio del indicador.

El ejercicio aquí desarrollado tiene, inevitablemente, un carácter preliminar. Es decir, las innovaciones en el campo del procesamiento de lenguaje natural, la disponibilidad de nuevos datos digitalizados y los cambios en las prácticas de comunicación implican que constantemente surgen oportunidades para refinar los métodos. Entre las extensiones que puede ser contempladas en lo inmediato, una alternativa involucra la detectar de comunidades de usuarios en Twitter y utilizar esta información para computar índices que resuman las opiniones observadas en distintas comunidades de usuarios. Otro ejercicio de interés tiene que ver con implementar técnicas de aprendizaje supervisado. Para ello es necesario en primer lugar clasificar el contenido de un número importante de mensajes que serían luego utilizados para entrenar modelos.

Referencias:

Aromí, J. D. (2017). Measuring uncertainty through word vector representations. Económica, 63, 135-156.

Aromi, J. D. (2020). Linking words in economic discourse: Implications for macroeconomic forecasts. International Journal of Forecasting, 36(4), 1517-1530.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. The quarterly journal of economics, 131(4), 1593-1636.

Baker, S. R., Bloom, N., Davis, S. J., & Renault, T. (2021). Twitter-derived measures of economic uncertainty.

Becerra, J. S., & Sagner, A. (2020). Twitter-Based Economic Policy Uncertainty Index for Chile. Central Bank of Chile, (883), 1-25.

Bernanke, B. S. (1983). Irreversibility, uncertainty, and cyclical investment. The quarterly journal of economics, 98(1), 85-106.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

Bloom, N. (2009). The impact of uncertainty shocks. econometrica, 77(3), 623-685.

Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2020). The structure of economic news (No. w26648). National Bureau of Economic Research.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. Journal of Economic Literature, 57(3), 535-74.

Istat (2022). SOCIAL MOOD ON ECONOMY INDEX: A daily measure of the Italian sentiment on the economy based on Twitter data. https://www.istat.it/it/files//2018/07/Methodological_Note_social-mood.pdf

Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M. T., & Perea-Ortega, J. M. (2013). Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, 40(18), 7250-7257.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica: Journal of the Econometric Society*, 703-708.

Newey, W. K., & West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, *61*(4), 631-653.

Orlik, A., & Veldkamp, L. (2014). Understanding uncertainty shocks and the role of black swans (No. w20445). National bureau of economic research.

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2020). Measuring news sentiment. *Journal of econometrics*.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. The American economic review, 71(3), 393-410.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, *62*(3), 1139-1168.

Anexo 1: Lista de palabras clave económicas

' actividad ',' ahorrar ',' ajuste ',' alimentos ',' alquiler ', ' alquileres ',' bancos ',' barato ', ' baratos ',' billete ', ' billetes ', ' billetera ', ' billeteras ',' boleto ',' boletos ',' bolsillo ',' bono ',' cobra ',' cobran ',' comercia', ' comerciamos', ' comercian', ' comercial ', ' comerciales',' comercio ', ' comercios ',' caro ',' compra ', ' compramos ', ' compran ', ' compras ',' compró ',' consumo ' , ' consumos ' ,' costo ',' crédito ', ' créditos ',' cuota ', ' cuotas ',' desarrollo ',' desempleo ',' deuda ',' dinero ',' dolar', ' dolares ',' dólar ', ' dólares ','economía ', 'economías ',' economia ', ' economías ' ,' económica ', ' económicas ', 'económica','economica','economico',' efectivo ',' empleados ',' empleo ',' empresa ',' empresas ',' estabilidad ',' estable',' factura ',' fmi ',' fondos ',' fortuna',' fortunas ',' gas ',' gastar',' gasto', ' gastos',' guita ', ' guitas ', ' impuesto ', ' impuestos ',' industria ', ' industrias ', ' industrial ', ' industriales ',' inflación ', ' inflación ', ' inflacionario ',' ingreso ', ' ingresos ',' inversión ', ' inversion ', ' inversiones ', ' jubilados ',' laboral ',' mercado ', ' mercados ', ' millón ', ' millones ',' ministro ', ' ministerio ',' moneda ', ' monedas ', ' nafta ' ,' negocio ',' negocios ',' oferta ', ' ofertas ',' paga ',' pagamos ','pagan ',' pagando ',' pagar ',' pago ',' pagos ',' pague',' pasaje ', ' pasajes ',' pesos ',' pobres ','pobreza ',' precio ',' precios ',' presupuesto ', ' presupuesto ',' privada ', ' producción ',' producto ',' productos ',' públicos ',' pyme ', ' pymes ',' recursos ',' ricos ','salario ', 'salarios ',' servicio ',' servicios ',' sueldo ',' sueldos ',' supermercado',' supermercado ',' usd ',' tarjeta ', ' tarjetas ',' trabaja ', ' trabajamos ', ' trabajan ', ' trabajas ',' trabajador ', ' trabajadores ',' trabajo ', ' trabajos ',' valores ',' vende', ' venden ', ' vendemos ',' vendiendo '

Anexo 2: Lista de palabras asociadas a incertidumbre

"incertidumbre" "inestabilidad" "crisis" "económica" "genera" "desconfianza" "contexto" "angustia" "cambiaria" "coyuntura" "recesión" "desesperación" "temor" "constante" "desigualdad" "volatilidad" "impacta" "política" "actualidad" "institucional" "tristeza" "panorama" "agrava" "impacto" "generalizada" "solvencia" "impotencia" "verdadera" "situación" "ante" "monetaria" "galopante" "optimismo" "financiera" "atraviesa" "profunda" "conflictividad" "destrucción" "especulación" "clima" "concentración" "global" "indignación" "desaceleración" "irresponsabilidad" "recuperación" "generó" "sequía" "pandemia" "inflación" "imposibilidad" "proyección" "expectativa" "escenario" "plena" "injusticia" "demanda" "calma" "tregua" "inseguridad" "generando" "devaluación" "frustración" "pérdida" "actual" "expectativas" "gregorossello" "confianza" "restricciones" "crecimiento" "seguida" "reactivación" "energética" "causa" "deforestación" "libertad" "tremenda" "inflacion" "creciente" "escasez" "socioeconómica" "economica" "intensidad" "momentos" "miedo" "debacle" "desocupación" "repercusión" "corrupción" "vulnerabilidad" "torno" "mucha" "profundiza" "tanta" "preocupación" "caída" "altísima" "generan" "depresión" "interna"