



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Estimación de un Índice de Incertidumbre de Política Económica para Colombia mediante el uso de NLP y modelos de aprendizaje supervisado y no supervisado

Sergio Daniel Pedraza Quiñones

Universidad Nacional de Colombia
Facultad de Ciencias Económicas
Bogotá, Colombia

2020

Estimación de un Índice de Incertidumbre de Política Económica para Colombia mediante el uso de NLP y modelos de aprendizaje supervisado y no supervisado

Sergio Daniel Pedraza Quiñones

Tesis presentada como requisito parcial para optar al título de:
Magíster en Ciencias Económicas

Director (a):

PhD Munir Andrés Jalil Barney

Línea de Investigación:

Ciencia de Datos

Universidad Nacional de Colombia

Facultad de Ciencias Económicas

Bogotá, Colombia

2020

Resumen

El documento presenta un índice de incertidumbre de política económica (o índice EPU) para Colombia construido mediante técnicas de Procesamiento de Lenguaje Natural (NLP). La hipótesis es que existe una relación significativa entre el índice EPU y variables como el crecimiento y la inflación y que choques de incertidumbre de política económica tienen un efecto adverso en el desempeño económico del país. Así mismo, se presume que un índice construido con técnicas de NLP recoge de mejor forma la información sobre incertidumbre que un índice construido simplemente con la búsqueda de palabras clave. El uso de herramientas de NLP, así como de modelos de aprendizaje supervisado y no supervisado constituye, tan lejos como sé, la primera aplicación de este tipo de modelos para Colombia en relación con la medición de incertidumbre de política económica. Se obtiene que el índice EPU, construido con un modelo de aprendizaje supervisado, exhibe la mejor capacidad explicativa con respecto a diversos indicadores macroeconómicos. Esta investigación se desarrolla con base en los artículos del archivo del periódico El Tiempo, el único en Colombia que cuenta con una hemeroteca digital desde el año 2000 hasta 2018, periodo comprendido por el estudio.

Palabras clave: Incertidumbre, Redes Neuronales, Procesamiento de Lenguaje Natural.

Abstract

This document introduces an economic policy uncertainty index (or EPU index) for Colombia built on Natural Language Processing (NLP) techniques. The hypothesis is that there exists a significant relationship between the EPU index and indicators like economic growth and inflation, and that economic policy uncertainty shocks have a prejudicial effect over the country's economic performance. Likewise, it is supposed that an index built on NLP techniques captures more appropriately information about uncertainty than an index built just through search for keywords. The use of NLP tools, as well as of supervised and unsupervised learning models constitutes, as far as I know, the first application of this kind of models for Colombia, within the scope of the measuring economic policy uncertainty. It is obtained that EPU index, when built with a supervised learning model, exhibits the best explaining capability with respect to diverse macroeconomic indicators. This research is done by the means of extracting articles from El Tiempo newspaper, the only one in Colombia that holds a digital newspaper library from 2000 until 2018, which is the period covered by the study.

Keywords: Uncertainty, Neural Networks, Natural Language Processing.

Contenido

Introducción	6
1. Revisión de la literatura principal	7
2. Datos	12
3. Construcción de índices EPU	14
3.1 Construcción de índice mediante aprendizaje supervisado.....	14
3.2 Construcción de índice mediante aprendizaje no supervisado.....	26
4. Relación de los índices construidos con distintas variables macroeconómicas.....	30
5. Conclusiones y Recomendaciones	40
Referencias Bibliográficas	42

Introducción

A lo largo de los últimos años y especialmente luego de la crisis económica y financiera de 2008, se han publicado distintos estudios sobre la incertidumbre de política económica y el impacto que puede tener en el sector real. Por ejemplo, en su documento de Perspectivas Económicas Mundiales de abril de 2016, el Fondo Monetario Internacional menciona cómo la incertidumbre de política puede resultar un lastre para la confianza y la inversión, así como incrementar la volatilidad de los mercados financieros mundiales (International Monetary Fund, 2016). Por su parte, Leduc et al. (2016) encuentran cómo un choque de incertidumbre se asemeja a un choque de demanda agregada, y puede conllevar a mayores niveles de desempleo y menor inflación.

Se han propuesto distintas metodologías en relación con la medición de la incertidumbre de política económica. Baker et al. (2016) construyen un índice denominado EPU (Economic Policy Uncertainty) a partir de artículos de prensa, mediante la búsqueda de palabras clave. En este estudio, los autores encuentran cómo choques de incertidumbre pueden afectar distintos indicadores macroeconómicos, así como el precio de las acciones de las empresas más dependientes de la contratación estatal.

Este método ha sido ampliamente replicado para distintos países y, de hecho, Baker et al. (2016) mantienen activa una página web en la cual se puede consultar el trabajo de todos los investigadores que han desarrollado el índice para sus respectivos territorios utilizando esta metodología.¹

Este recurso web contiene el trabajo adelantado por Gil y Silva (2018), en el cual construyen el índice EPU para Colombia. No obstante, en este trabajo no se evalúa la relación del índice construido con alguna variable macroeconómica, en tanto se limita a describir su relación con el índice construido para Estados Unidos y con algunos eventos relevantes de la historia reciente en Colombia, con el fin de explicar los picos observados en el indicador.

El presente trabajo busca construir dos índices EPU para Colombia, utilizando dos aproximaciones distintas que utilizan herramientas de la disciplina conocida como Procesamiento de Lenguaje Natural (en inglés Natural Language Processing) y determinar

¹ <https://www.policyuncertainty.com/index.html>

cuál se relaciona de forma más clara con el entorno macroeconómico del país, así como compararlos con el desarrollado por Gil y Silva (2018).

Así mismo, se utilizan modelos de aprendizaje de máquina o machine learning en la construcción de los índices, basado principalmente en el trabajo de Tobbach et al. (2018) y Azqueta-Gavaldón (2017), bajo la hipótesis de que estas herramientas mejoran los resultados obtenidos al estudiar la relación del índice con distintas variables macroeconómicas.

El presente documento se divide en 5 secciones: en la primera se presenta la revisión de literatura de los principales resultados relacionados con la construcción de índices de incertidumbre de política económica en estudios que utilizan artículos de prensa. Posteriormente, en la siguiente sección se describe la fuente de los datos utilizados en la investigación y se explica el procedimiento mediante el cual se obtienen. En la tercera sección, se exponen los modelos y las herramientas usadas para la construcción de los índices EPU, y en la cuarta se muestran los resultados obtenidos respecto a la relación de los índices construidos y el desarrollado por Gil y Silva (2018) con las distintas variables macroeconómicas. Por último, en la quinta sección se resumen las principales conclusiones.

1. Revisión de la literatura principal

La construcción de un índice que permite medir la incertidumbre de política económica a partir de la prensa fue propuesta inicialmente por Baker et al. (2016). En este trabajo, los autores construyen el así llamado índice EPU (Economic Policy Uncertainty), para los Estados Unidos y otras 11 economías² utilizando algunos de los principales diarios de cada territorio.

El índice EPU se construye a partir de la frecuencia con la cual se publican artículos que tratan sobre incertidumbre de política económica. Dado el tamaño bruto del corpus documental con el que trabajan, para escoger de manera eficiente los artículos que abordan este tema sin tener que revisar uno a uno cada documento, los autores definen tres grupos de palabras clave: uno relacionado con incertidumbre, uno con política, y uno con

² Las cuales corresponden a India, Canadá, Corea del Sur, Francia, Alemania, Italia, Japón, España, el Reino Unido, China y Rusia.

economía³. Posteriormente asumen que, si un artículo contiene al menos una palabra de cada grupo, el artículo trata sobre incertidumbre de política económica.

Luego, construyen una serie de frecuencia relativa de artículos relevantes frente al total. Esta serie, con periodicidad mensual, es transformada para tener media 100 y desviación estándar 1. La elección de los grupos de palabras que representan la incertidumbre de política económica es auditada por medio de un proceso arduo, durante el cual asistentes de investigación leen y clasifican manualmente 60.000 artículos de acuerdo con un instructivo preparado por los autores.

Se comparan los resultados obtenidos entre los asistentes de investigación y el algoritmo de búsqueda de palabras. Posteriormente se ajustan las palabras seleccionadas para obtener una discrepancia tan baja como fuese posible entre ambos conjuntos de resultados.

En adición a esto, los investigadores también construyen distintos índices al tener en cuenta la palabra específica del grupo de política encontrada en el artículo. Esto es, para artículos donde la palabra del grupo de política es salud, se construye un índice de incertidumbre de política económica relacionada con salud. Para la palabra déficit se construye un índice de incertidumbre de política fiscal.

Con los índices construidos, los autores proceden a investigar su relación con distintas variables, tanto a nivel de empresa como a nivel agregado. La comparación a nivel de empresa se realiza al utilizar los índices de sectores de política específicos y, al usar un modelo VAR, y encuentran, por ejemplo, que choques de incertidumbre en el índice relacionado con el sector salud tienen un efecto de incremento en la volatilidad del precio de las acciones de las empresas que tienen contratos relacionados con salud con el gobierno norteamericano.

Por su parte, al investigar sobre la relación del índice EPU principal con algunas variables macroeconómicas estadounidenses, encuentran que choques de incertidumbre de política económica tienen un efecto adverso sobre la producción industrial y el empleo en E.E.U.U.

³ En el grupo de términos relacionados con incertidumbre se incluyen 'uncertainty' o 'uncertain'. En el grupo de economía se incluyen 'economic' o 'economy', mientras que en el grupo de política se incluyen los siguientes términos 'Congress,' 'deficit,' 'Federal Reserve,' 'legislation' 'regulation' o 'White House' entre otros. Dentro de cada grupo se incluyen sus variaciones.

Concluyen, en línea con la mayoría de los estudios sobre el tema, que los choques de incertidumbre parecen estar asociados a un pobre desempeño económico posterior, y que los altos niveles de incertidumbre observables por medio del índice EPU durante los últimos años, y especialmente después de la crisis de 2008 puede ayudar a explicar la lenta recuperación de la actividad económica.

No obstante, la metodología descrita anteriormente se ve sujeta a distintas limitaciones. La principal es que el proceso de auditoría de los artículos seleccionados es intensivo en horas de trabajo, al implicar la lectura de una gran muestra de estos documentos. Dada esta restricción, y el hecho de que el método de Baker et al. (2016) es propenso a errores Tipo I y Tipo II⁴, Tobback et al. (2018) proponen el uso de técnicas de NLP y un modelo de aprendizaje supervisado, esta vez para Bélgica, en adición a la construcción del índice al usar la metodología tradicional, mediante el uso de palabras clave.

El Procesamiento de Lenguaje Natural (o NLP, por sus siglas en inglés) explora la manera cómo los computadores analizan los lenguajes naturales, llamados así para distinguirlos de otro tipo de lenguajes no utilizados por los seres humanos naturalmente en su comunicación, como los lenguajes de programación. Por su parte, un modelo de aprendizaje supervisado es un modelo estadístico generalmente utilizado en el campo del aprendizaje de máquina, en el cual se utilizan una o más variables explicativas para estimar el comportamiento de una variable explicada. Los problemas resueltos generalmente por un modelo supervisado son regresión, cuando la variable explicada es continua, o clasificación, cuando es discreta.

De esta forma, Tobback et al. (2018) extraen los artículos publicados por los 5 principales periódicos flamencos, de los cuales extraen una pequeña muestra (de aproximadamente 400 artículos), con el fin de ser clasificada entre dos conjuntos: si trata sobre incertidumbre de política económica, o si no. Esta clasificación es realizada por un humano de acuerdo con determinados criterios.

Con este corpus reducido, entrenan un Support Vector Machine, el cual es un modelo de clasificación que permite, a partir del contenido de un artículo, categorizarlo en uno de ambos grupos extrayendo los patrones de la codificación adelantada por los investigadores.

⁴ Esto es, a clasificar un artículo como relevante cuando no lo es y a clasificar un artículo como no relevante cuando lo es.

Este modelo es refinado de acuerdo con distintas métricas de bondad de ajuste con el fin de obtener la clasificación más certera posible.

De esta manera, logran construir un índice con un mejor control sobre los artículos que son categorizados como relevantes y que es desarrollado con base en un conjunto de patrones más amplio que un grupo predeterminado de palabras clave. De igual forma, el proceso en sí es mucho menos intensivo en horas de trabajo.

Posteriormente, los autores estudian la relación del índice construido con distintos indicadores de la economía de Bélgica, con lo cual llegan a resultados interesantes. Entre los indicadores analizados se encuentran el rendimiento de los bonos belgas de 10 años, el spread entre los bonos belgas y alemanes de 10 años, el spread de los Credit Default Swap en los bonos belgas de 5 años, el índice de confianza del consumidor, el indicador de la encuesta de negocios de Bélgica, el indicador de la demanda esperada del sector de la construcción, el indicador de demanda de casas en Bélgica, el IPC armonizado en ese país, el índice de registro de vehículos, y el rendimiento de las acciones del principal índice belga.

Los investigadores encuentran que, dentro de la muestra clasificada manualmente, el índice construido mediante NLP y el modelo de clasificación tiene mejor ajuste a la hora de categorizar un artículo como relevante en comparación con el método de palabras clave. Así mismo encuentran que, para la mayoría de los indicadores analizados, el índice presenta una buena capacidad de predicción en el corto plazo utilizando modelos de predicción rodante. Este resultado contrasta con el obtenido al utilizar el índice construido mediante el uso de palabras clave, para el cual no se encuentra evidencia de una relación significativa con alguno de los indicadores.

Por último, evidencian que el índice EPU presenta niveles más elevados en los periodos posteriores a la crisis del 2008 y plantean cómo su trabajo es un elemento que pretende cerrar la brecha entre los trabajos de economía y el aprendizaje de máquina, el cual ellos afirman sigue dominado por 'el modelamiento estadístico causal' (Tobback et al., 2018).

No obstante, esta metodología presenta igualmente algunas limitaciones. En tanto reduce la complejidad del proceso de auditoría manual de manera considerable, existe igualmente la necesidad de que un humano lea y clasifique manualmente una muestra de artículos. Dado que el tamaño de la muestra puede influir en la calidad del modelo de clasificación

construido, el argumento de que sólo es necesaria una muestra pequeña puede ser debatido. De la misma forma, puesto que el modelo clasificador de texto se entrena en función de los datos provistos por un humano, este puede 'aprender' patrones erróneos si la clasificación no se realiza con el rigor correspondiente.

Es así como, Azqueta–Gavaldón (2017) propone una metodología de aprendizaje no supervisado. En términos simples, esto quiere decir que se trabaja únicamente con variables exógenas, sin una variable de respuesta o endógena. De entrada, la gran ventaja de este tipo de modelos es que no se requiere una clasificación previa a su entrenamiento, por lo cual la auditoría humana, ya sea para escoger palabras clave o para clasificar una muestra de artículos, no es requerida.

En este caso, el modelo utilizado es conocido como Latent Dirichlet Allocation (LDA), una técnica que puede ser entendida tanto como de reducción de dimensionalidad como de clustering. Esta técnica intenta condensar el contenido de los distintos artículos en un número fijo de temas.

El principal objetivo del autor es demostrar que se puede construir un índice similar al desarrollado por Baker et al. (2016), sin la necesidad de pasar por un arduo proceso de auditoría y con la posibilidad de minimizar las horas de trabajo dedicadas. Por lo tanto, utiliza el mismo corpus documental que en el trabajo de la metodología original. Posteriormente, extrae únicamente los artículos que presentan en alguna forma las palabras 'economía' e 'incertidumbre' y ajusta el modelo LDA. Este modelo tiene como salida una matriz en la que se asocia cada uno de los artículos seleccionados con un tema dominante, que a su vez es descrito por la colección de palabras más frecuente en el mismo.

El autor encuentra que el índice construido mediante aprendizaje no supervisado y el construido por Baker et al. (2016) tienen una correlación de 0,94. Incluso, obtiene que la correlación entre los componentes ciclo de ambos índices es de 0,88, en tanto entre sus componentes de tendencia es de 0,99. Nota que las mayores diferencias se encuentran en la intensidad de la respuesta de cada uno de los índices a algunos choques asociados a eventos geopolíticos.

De esta forma, el investigador concluye que puede construirse un índice EPU, con propiedades extremadamente similares a las obtenidas mediante la metodología original,

pero con la ventaja de haber tomado unos cuantos días, en lugar de dos años (Azqueta-Gavaldon, 2017).

En el caso de Colombia, Gil y Silva (2018) utilizan la metodología propuesta por Baker et al. (2016) para construir el índice EPU usando los artículos encontrados en el archivo digital del diario El Tiempo. Así, escogen 3 conjuntos de palabras que describen la incertidumbre, la economía y la política. Adicionalmente, seleccionan un cuarto grupo de palabras relativas a Colombia. Esto con el fin de escoger artículos que traten sobre incertidumbre de política económica en el país, a diferencia de la metodología original, en la cual no se discrimina por país, bajo el argumento de que una mezcla de hechos nacionales e internacionales son los que influyen la incertidumbre de política en un territorio (Baker et al., 2016).

Gil y Silva (2018) culminan su trabajo presentando el índice construido en tanto relacionan algunos de sus picos con hechos relevantes de la historia reciente en Colombia. No obstante, no realizan una evaluación de la relación del índice construido con alguna variable macroeconómica en Colombia.

Por último, Ahir et al. (2018), construyen un índice EPU mundial basados en un corpus documental compuesto por las publicaciones del Economist Intelligence Unit, del grupo editorial The Economist. En esta ocasión, dado el contenido del corpus con el que trabajan los investigadores, basados en la metodología de Baker et al. (2016), buscan únicamente términos relacionados con la palabra incertidumbre para construir el índice. A su vez, encuentran que innovaciones en el índice WUI (World Uncertainty Index) preceden a cambios significativos en el producto económico mundial, por lo cual sugiere que este índice puede servir como una medida alternativa de actividad económica, cuando éstas no estén disponibles.

2. Datos

Los datos utilizados en esta investigación corresponden a los artículos publicados por el periódico El Tiempo en su archivo digital desde enero de 2000 hasta diciembre de 2018⁵, a semejanza de lo realizado por Gil y Silva (2018). La elección de esta fuente de datos obedece a dos grandes razones que limitan el alcance del ejercicio: en primer lugar, porque El Tiempo es el periódico en Colombia que cuenta con la hemeroteca digital más antigua,

⁵ El archivo puede ser consultado en la URL <https://www.eltiempo.com/buscar>.

y en segundo, con el fin de que los índices construidos fuesen comparables con el presentado por los autores anteriormente mencionados.

Vale la pena mencionar que la elección de esta fuente de datos lleva asociada un sesgo producto del componente ideológico de la editorial. En su trabajo, Baker et al (2016) le dan tratamiento a este problema al estimar el índice con base en 10 periódicos, entre los cuales identifican 5 de corriente liberal y 5 de corriente más conservadora, y encuentran que el resultado es similar entre ambos. No obstante, en esta investigación, la poca disponibilidad de datos en el periodo que abarca es un limitante que no permite desarrollar un ejercicio similar. Hacia el futuro, realizar esta estimación a partir de artículos de más periódicos podría enriquecer el análisis.

Para la compilación de los artículos se desarrolla un programa de scraping. El scraping es una técnica de minería de datos que consiste en consultar una página web, extraer información presentada en la misma y almacenarla. Su nombre en inglés hace referencia a cómo el minero 'raspa' contenido de una página web y procede a guardarlo para su consumo posterior. El programa, o crawler, fue desarrollado en su totalidad mediante el lenguaje de programación Python, al igual que el resto de los ejercicios de estimación y procesamiento de datos en este trabajo de investigación.

El procedimiento de extracción de artículos tuvo dos fases distintas: en primer lugar, se construye un crawler que consulta y almacena tantas páginas de resultados como se obtengan por mes al buscar todos los artículos publicados en el archivo, desde enero de 2000 hasta diciembre de 2018.

Posteriormente se itera para cada mes y para cada página de resultados y se obtienen los vínculos que dirigen a todos los artículos publicados. Estos vínculos se escriben a archivos de texto, uno por línea, y cada archivo de texto se almacena al finalizar un mes. En total, se obtienen 228 archivos de texto plano, correspondientes a los 228 meses abarcados por el estudio.

Una vez almacenadas las direcciones web de cada uno de los artículos publicados, se procede a iterar sobre estas y se extrae la información sobre el contenido de cada documento. Vale la pena aclarar que se extrae únicamente el cuerpo de cada artículo, no su título, autor ni cualquier otra metadata asociada al mismo. De la misma forma, cada

artículo es escrito a un archivo de texto titulado con el vínculo y el mes del cuál proviene. Estos archivos son almacenados en 228 carpetas diferentes, una para cada mes.

En esta primera iteración del ejercicio, se obtienen 1'660.010 archivos, cada uno correspondiente a un artículo publicado. Esta información representa aproximadamente 5,7 GB de peso. No obstante, se realiza una depuración de los artículos a utilizar dado que muchos de los archivos de texto no tienen contenido. Esto se debe principalmente a que dentro de la búsqueda se obtienen los vínculos de galerías de fotografías, caricaturas o videos, los cuales no tienen contenido escrito. Así, son separados del corpus documental utilizado en este trabajo de investigación 17.116 archivos con lo cual el total utilizado es de 1'642.894 artículos.

Por último, se utiliza información de varios indicadores macroeconómicos colombianos. Estos son extraídos de las fuentes oficiales⁶.

3. Construcción de índices EPU

3.1 Construcción de índice mediante aprendizaje supervisado

Bajo el enfoque del aprendizaje supervisado se busca ajustar un modelo que estime de la manera más acertada posible el comportamiento de una variable de respuesta Y dado un conjunto de variables explicativas X . Los problemas resueltos al utilizar este esquema suelen ser de clasificación y regresión. La diferencia entre ambos casos es que cuando se trata de clasificación, la variable de respuesta es discreta, entre tanto que en un modelo de regresión esta variable es continua.

Al utilizar como base el trabajo de Tobback et al. (2018) para construir un índice EPU mediante aprendizaje supervisado, es necesario construir un modelo de clasificación binaria que discrimine los artículos contenidos en el corpus documental. De esta forma, no existe dependencia de un conjunto preseleccionado de palabras clave, dado que el modelo es la herramienta utilizada para discriminar los artículos entre dos categorías: relevante, si trata sobre incertidumbre de política económica, y no relevante sino lo hace.

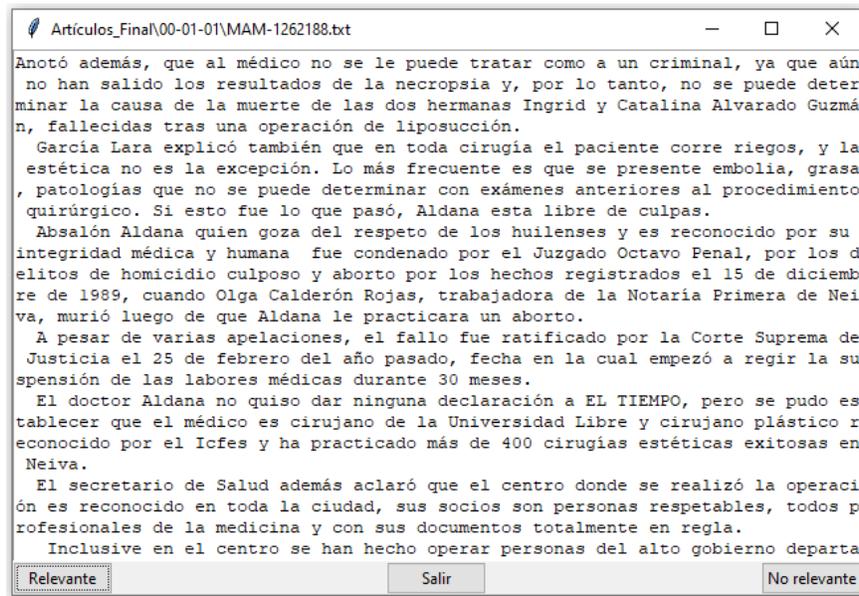
A priori, la ventaja de este método es que, dado el gran tamaño del corpus, que en este caso asciende a 1'642.894 artículos, es suficiente seleccionar una muestra a partir de la cual es entrenado el modelo que, después de ser evaluado para encontrar la mejor precisión

⁶ En particular, estas son: <https://www.dane.gov.co/> y <https://www.banrep.gov.co/>

posible en la clasificación, es utilizado para discriminar el total del corpus entre artículos relevantes y no relevantes.

En el caso del presente trabajo, se utiliza una muestra de 1.140 artículos, lo cual corresponde a 5 artículos de cada mes. Estos artículos son clasificados manualmente por el autor del documento al utilizar un programa desarrollado específicamente para este ejercicio, con su correspondiente interfaz gráfica de usuario, cuyo objetivo es mejorar la eficiencia con la cual se leen y clasifican artículos, basado igualmente en el lenguaje de programación Python. Una muestra de cómo luce el programa puede observarse a continuación:

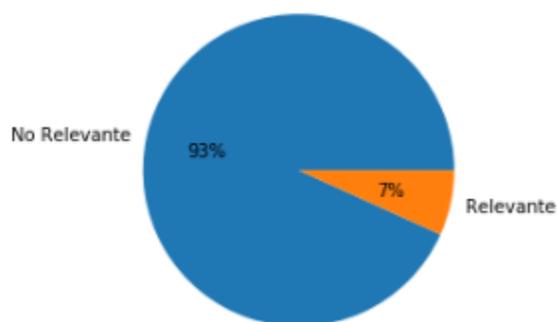
Gráfica 1. Programa de clasificación desarrollado específicamente para el trabajo de investigación.



Fuente: Código propio.

La función de este programa es mostrar en pantalla un artículo de la muestra, permitir su clasificación entre relevante y no relevante, y almacenar esta clasificación. Se obtiene que la mayoría de los artículos son no relevantes, como es de esperarse, y que sólo unos cuantos abordan el tema de la incertidumbre de política económica. En concreto, 1.063 artículos, correspondientes al 93% de la muestra son categorizados como no relevantes, en tanto 77, correspondientes al 7% lo son, como puede observarse en la siguiente gráfica:

Gráfica 2. Porcentaje de artículos clasificados como relevantes y no relevantes.



Fuente: Elaboración propia.

Respecto a los criterios bajo los cuales se realiza la clasificación, vale la pena hacer unas cuantas aclaraciones. Esta discriminación intenta seguir de la mejor forma el instructivo anexo al trabajo de Baker et al. (2016), en el cual señalan que un artículo aborda el tema de la incertidumbre de la política económica cuando cumple con las siguientes características:

- El artículo trata sobre incertidumbre económica explícitamente en relación con temas de política.
- El artículo refleja incertidumbre sobre quién toma decisiones de política que tienen efectos económicos.
- El artículo discute sobre la incertidumbre relacionada con los efectos de políticas económicas pasadas, presentes o futuras.
- El artículo trata sobre la incertidumbre económica derivada de la inacción de política.
- El artículo discute sobre la incertidumbre económica relacionada con desarrollos de política motivados por consideraciones no económicas, como por ejemplo la seguridad nacional.

En adición a esto, se realiza una segunda clasificación. La primera y ya mencionada por parte del autor del documento, y otra por un colaborador, economista y estudiante de maestría en economía en otra institución. El objetivo de este procedimiento es realizar una clasificación lo más objetiva posible. En el caso de haber inconsistencias respecto a la categoría asignada a un artículo, se llega a un consenso sobre la misma y la clasificación consensuada es la finalmente almacenada.

Es necesario realizar un preprocesamiento al corpus documental antes de utilizarlo como insumo para un modelo de machine learning que permita estimar la clase de un artículo individual. El preprocesamiento y las técnicas de NLP utilizadas en este trabajo de investigación están basadas en las expuestas en el trabajo de Weiss et al. (2015).

El preprocesamiento realizado a cada artículo de la muestra consiste en:

- Se tokeniza (es decir segmenta) por palabras.
- Se eliminan caracteres producidos por errores de codificación.
- Se eliminan las así llamadas stopwords, que son palabras que por su utilización general en el lenguaje se presuponen poco discriminativas. Como ejemplo se encuentran los artículos determinados e indeterminados, las preposiciones, algunos verbos, etc.
- Cada token, que en este caso corresponde a una palabra, se intercambia donde sea posible por su lema, procedimiento llamado lematización. Este proceso tiene como objetivo reemplazar una palabra por su variante sin flexionar, de ser posible. Como ejemplo, las palabras jugaron, jugué y jugarán quedarían todas reemplazadas por jugar.
- Se eliminan caracteres no alfabéticos, incluidos espacios.

Al final de este procedimiento, cada artículo se ve representado por un arreglo de palabras sin flexionar. Un ejemplo de cómo es transformado un documento al aplicar este proceso puede observarse en la siguiente tabla:

Tabla 1. Artículo CMS-7822869 de julio de 2010 procesado y sin procesar

Artículo sin procesar	Artículo procesado
<p>¿Cómo define al consumidor colombiano?</p> <p>El consumidor colombiano se ha convertido en un cazador de precios y promociones, y siente un menor valor en las compras que hace. Esto se suma a la presencia de tres generaciones de consumidores que tienen diferentes reacciones al precio. La generación café (personas mayores de 45 años) que son muy dados a grandes descuentos y aceptan cambios de precios, siendo consumidores muy tradicionales. La generación gris, de 25 a 45 años, es</p>	<p>'definir', 'consumidor', 'colombia', 'consumidor', 'colombia', 'convertir', 'cazador', 'precio', 'promocionar', 'sentir', 'menor', 'comprar', 'sumo', 'presenciar', 'generación', 'consumidor', 'reaccionar', 'preciar', 'generación', 'café', 'personar', 'mayor', 'año', 'dar', 'descuento', 'aceptar', 'cambio', 'precio', 'consumidor', 'tradicional', 'generación', 'gris', 'año', 'dar', 'promocionar', 'tendencia', 'dándole', 'tecnología', 'generación', 'verde', 'sensible', 'cambio', 'precio', 'tendencia', 'marcar', 'posición', 'político', 'consumir', 'pesar', 'responsabilidad', 'social', 'alto',</p>

<p>muy dada a promociones y a las tendencias, dándole gran valor a la tecnología. La generación verde, muy sensible a cambios de precios y con una tendencia muy marcada a la posición política del consumo, donde el peso de la responsabilidad social es muy alto pero no se refleja en precio sino que se considera como un estándar esperado.</p>	<p>'reflejo', 'preciar', 'estándar', 'esperar', 'diferenciar', 'marcar', 'edad', 'estrato',</p>
<p>¿Existe una diferencia marcada por edades o por estratos en las decisiones de compra?</p>	<p>'decisión', 'comprar', 'dudar', 'menor', 'nivel', 'ingresar', 'pesar', 'bien', 'durable', 'aumentar', 'medir', 'subir', 'nivel', 'ingresar', 'sensibilidad', 'preciar', 'cambiar', 'bien', 'sensible', 'descuento', 'demostrar', 'personar', 'alto', 'ingreso', 'racional', 'tomar', 'decisión', 'básico', 'bien', 'alto', 'desembolsar', 'tecnología', 'líneo', 'blanco', 'alto', 'comprar', 'tender', 'razonar', 'fundamental', 'procesar', 'diez', 'colombiano', 'vivir', 'ingresar', 'diario', 'gastar', 'diario', 'mix', 'producto', 'ofrecer', 'tienda', 'referir', 'consumir', 'diario', 'sostenimiento', 'canal', 'importancia', 'marcar', 'consumir', 'masivo', 'reaccionar', 'consumidor', 'inauguración', 'centrar', 'comercial', 'hipermercado', 'novedad', 'generar', 'interés', 'tasar', 'tráfico', 'venta', 'punto', 'comercial', 'alto', 'estabilizar', 'depender', 'ubicación', 'ofertar', 'centrar', 'bogotá', 'crecimiento', 'sostener', 'comenzar', 'mostrar', 'dinámico', 'mix', 'almacenar'</p>
<p>Sin duda, a menor nivel de ingreso el peso de bienes no durables aumenta y a medida que sube el nivel de ingreso, la sensibilidad de precio cambia en ciertos bienes, pero son más sensibles a descuentos, lo que demuestra que las personas de altos ingresos son racionales en la toma de decisiones básicas o de bienes de alto desembolso como tecnología y línea blanca.</p>	
<p>¿Por qué sigue siendo alta la compra en la tienda?</p>	
<p>Existen tres razones fundamentales en ese proceso: más o menos seis de cada diez colombianos vive de un ingreso diario y por lo tanto gasta a diario. El mix de producto ofrecido en las tiendas se refiere a consumo diario. El sostenimiento de este canal tiene una gran importancia para las marcas de consumo masivo.</p>	
<p>¿Cómo reaccionan los consumidores ante la inauguración de un centro comercial o hipermercado?</p>	
<p>Toda novedad genera interés. Por esto, las tasas de tráfico y ventas de un nuevo punto comercial son muy altas hasta cuando se estabiliza, dependiendo de la ubicación y la oferta del mismo. Un buen ejemplo de esto es Centro Mayor, de Bogotá, que ha tenido un crecimiento sostenido, pero ya comienza a mostrar dinámicas en su mix de almacenes.</p>	

vectorización. Para vectorizar una colección de documentos se construye la así llamada matriz término-documento, o matriz dt.

La matriz término-documento consiste en una matriz $n \times m$, con n documentos y m palabras distintas en la colección de documentos. Para el modelo de aprendizaje supervisado, se utiliza una técnica de vectorización conocida como term frequency – inverse document frequency o tf-idf (Weiss et al., 2015).

El primer componente de esta técnica, term frequency, hace referencia a que cada elemento C_{ij} de la matriz dt corresponde a la cantidad de veces C que la palabra j aparece en el documento i. Por tanto, cada elemento de la matriz refleja la frecuencia de cada término en un documento. Posteriormente, el componente inverse document frequency modifica los valores C. Esta modificación se hace bajo la idea de que un término es poco discriminativo si aparece en la mayoría de los documentos analizados.

La transformación idf de cada elemento C de la matriz término-documento corresponde a la siguiente fórmula:

$$idf(C) = \log\left(\frac{1+n}{1+C}\right) + 1$$

Donde n corresponde al número de documentos en el corpus. Es así como, mientras que el componente tf intenta darle un valor más alto a cada palabra dentro de la matriz si aparece con más frecuencia dentro de un documento, el componente idf intenta disminuir ese valor si la palabra aparece con más frecuencia entre todos los documentos. Por último, se calcula el producto $tf \times idf$ para obtener el valor de cada componente de la matriz. Para el caso de la presente investigación, se trabaja con los 20.000 términos más frecuentes, por lo que en primera instancia se obtiene una matriz de 1.140×20.000 .

Esta matriz tiene una característica conocida como dispersión (sparsity en inglés) lo cual quiere decir que la mayoría de sus componentes son ceros. Esto conlleva algunos problemas, como que su almacenamiento o cualquier cálculo basado en ella son computacionalmente intensivos en memoria y procesamiento. De esta forma, conviene utilizar una técnica de reducción de dimensionalidad, mediante la cual se intenta aproximar una matriz de determinado tamaño con otra matriz de rango inferior. Un ejemplo de este tipo de técnicas es Análisis de Componentes Principales.

No obstante, en esta investigación es usada una técnica conocida como Análisis Semántico Latente o LSA. El LSA es el nombre utilizado en el contexto del Procesamiento de Lenguaje Natural para una técnica de reducción de dimensionalidad más conocida como Descomposición en Valores Singulares o SVD.

La Descomposición en Valores Singulares descompone una matriz X de la siguiente forma:

$$X_{n \times m} = U_{n \times n} \Sigma_{n \times m} V^*_{m \times m}$$

Donde U y V^* son matrices unitarias. Si todos los elementos de una matriz son números reales, una matriz unitaria se caracteriza porque su inversa y su transpuesta son equivalentes. Así mismo, Σ es una matriz diagonal rectangular, en la que los elementos de la diagonal son conocidos como valores singulares. Los m valores singulares son producto de combinaciones lineales de los valores de cada fila de la matriz X (Halko et al., 2011).

A partir del producto $U\Sigma_{n \times m}$, se escogen k componentes, de tal forma que se trunca la matriz Σ a sus k primeras columnas, que contienen los k primeros valores singulares. De esta manera, la matriz con la que se estima el modelo de clasificación corresponde a $U\Sigma_{n \times k}$, con $k < m$. La descomposición es generalmente aproximada mediante algoritmos numéricos, dada la complejidad del cálculo cuando se trabaja con matrices de rango elevado.

En el caso de la presente investigación, se utiliza una descomposición con 100 componentes. La elección del número de componentes no es discrecional, puesto que se escoge la cantidad que permite que el modelo entrenado alcance mejores niveles de precisión, lo cual se explica más adelante en el documento.

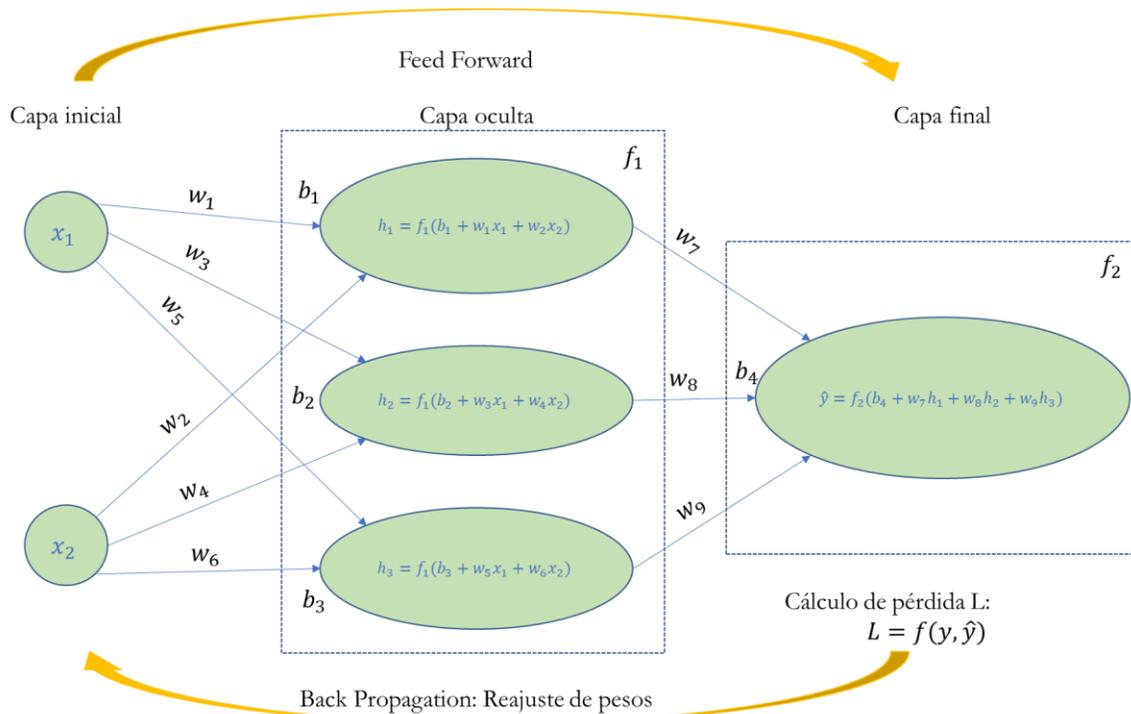
En este punto, se tienen tanto las clasificaciones como la matriz que contiene las características de cada artículo. En el primer caso, las clasificaciones corresponden a un vector de 1.140 elementos que pueden tomar el valor 1, si el artículo trata sobre incertidumbre de política económica, o 0 sino es el caso, en tanto en el segundo se tiene una matriz de dimensión 1.140×100 .

El modelo de clasificación desarrollado es una red neuronal. Una red neuronal es un modelo matemático que debe su nombre a su intención de imitar la 'arquitectura' del cerebro humano, dentro del cual la información es transmitida entre las distintas neuronas. Una red sencilla consta de una capa de entrada, una o varias capas escondidas o intermedias, y una capa de salida. A su vez, cada capa está compuesta por una o más neuronas. La

construcción de la red utilizada en esta investigación fue construida en Python por medio de la librería Keras usando como backend el software Tensorflow, desarrollado por Google. Así mismo, el modelo construido está basado en el trabajo publicado por Skansi (2018).

En términos simples, una red neuronal toma un insumo, lo transforma al interior de cada una de sus capas al operarlo con algunos parámetros y produce un valor en su capa final que corresponde a la predicción realizada. Este proceso es conocido como feed forward. A partir de esta predicción se calcula la diferencia entre el valor real y el estimado por medio de lo que se conoce como función de pérdida, y posteriormente se reajustan los parámetros con el fin de reducir esta pérdida en una siguiente iteración. Este proceso es conocido como back propagation. La siguiente gráfica muestra la estructura de una red neuronal sencilla:

Gráfica 4. Estructura de una red neuronal sencilla.



Fuente: Skansi (2018). Elaboración propia.

En la gráfica anterior, por cuestión de simplicidad, se muestra únicamente una capa oculta. En la capa de entrada, las variables x_1 y x_2 representan los vectores de características que funcionan como insumo del modelo de clasificación. Cada característica pasa por una neurona de la capa inicial, por tanto, en el caso de la red construida en el presente trabajo de investigación, esta capa cuenta con 100 neuronas.

En el caso del modelo construido, la red está densamente conectada. Esto quiere decir que cada neurona tiene una conexión con cada una de las neuronas de la capa posterior. Así mismo, cada conexión entre neuronas tiene un peso w_i asociado. Este peso representa qué tanta 'importancia' le da una neurona a la conexión que tiene con otra de una capa precedente. En adición a esto, con excepción de las pertenecientes a la capa inicial, cada neurona tiene un bias o sesgo b_j asociado. El valor de estos parámetros es inicializado aleatoriamente al inicio del proceso de estimación.

De la misma forma, todas las capas a excepción de la inicial cuentan con una función de activación f_k . Esta función se utiliza para constreñir el valor producido por cada neurona de la red en una capa. El valor h_j que produce una neurona equivale a:

$$h_j = f_k \left(b_j + \sum_{i=1}^I w_i x_i \right)$$

Donde $j = 1 \dots J$, corresponde a la cantidad de neuronas en capas distintas a la inicial, $k = 1 \dots K$ es el número de capas distintas a la inicial y $i = 1 \dots I$ corresponde al número de conexiones de una neurona con la capa precedente. Por tanto, el valor que produce una neurona corresponde al resultado de evaluar la función de activación en una combinación lineal de los productos de las neuronas de la capa precedente y los pesos de sus conexiones al adicionar el sesgo. Este procedimiento se repite para cada neurona de las capas subsecuentes.

Entre las funciones de activación más utilizadas están la sigmoide o logística, la tangente hiperbólica, la softmax y la rectificadora o relu (Skansi, 2018). El modelo construido en la presente investigación cuenta, como se menciona anteriormente, con una capa inicial de 100 neuronas, dos capas ocultas de 60 neuronas cada una con función de activación relu, y una capa final de una neurona con función de activación sigmoide.

En general, no hay restricción sobre las funciones de activación que pueden tener las capas ocultas, pero en el caso de clasificación binaria, en la capa final es necesario forzar la función de activación a una sigmoide, para obtener un output con valores entre 0 y 1. Esto se debe a que, a similitud de la regresión logística, se suele asumir que las observaciones cuya predicción, que puede interpretarse como una probabilidad, es mayor o igual a 0,5 son clasificadas dentro de la categoría 1, entre tanto las observaciones cuya predicción es menor a 0,5 son clasificadas dentro de la categoría 0.

En adición a esto, el output de cada capa oculta fue normalizado y estandarizado, procedimiento conocido como batch normalization (Ioffe et al., 2015). Este proceso permite que el entrenamiento de la red neuronal se complete a una mayor velocidad.

La elección de los parámetros de la red neuronal obedece a un proceso pseudo-aleatorio de optimización, conocido como randomized search. En resumen, este procedimiento consiste en estimar múltiples veces el modelo únicamente con un subconjunto de la muestra denominado de entrenamiento, con un cambio aleatorio en cada iteración, elegido de un conjunto de posibles cambios determinados previamente. Para cada combinación de hiperparámetros se estima el modelo mediante la técnica conocida como k-fold cross validation, en este caso con $k = 10$, y al final del proceso, se escoge el modelo que obtuvo el mejor indicador de sensibilidad.

La sensibilidad fue escogida como criterio dado que, al existir desbalanceo de clases, y por el enfoque de la investigación, es más relevante identificar correctamente la clase minoritaria. Los cambios que prueba el randomized search pueden consistir en una función de activación distinta en las capas ocultas, o una capa oculta adicional, o un número diferente de neuronas en cada capa. En el caso de este ejercicio, dentro de estos parámetros también se evaluaron distintos valores para el número de componentes en el proceso de SVD.

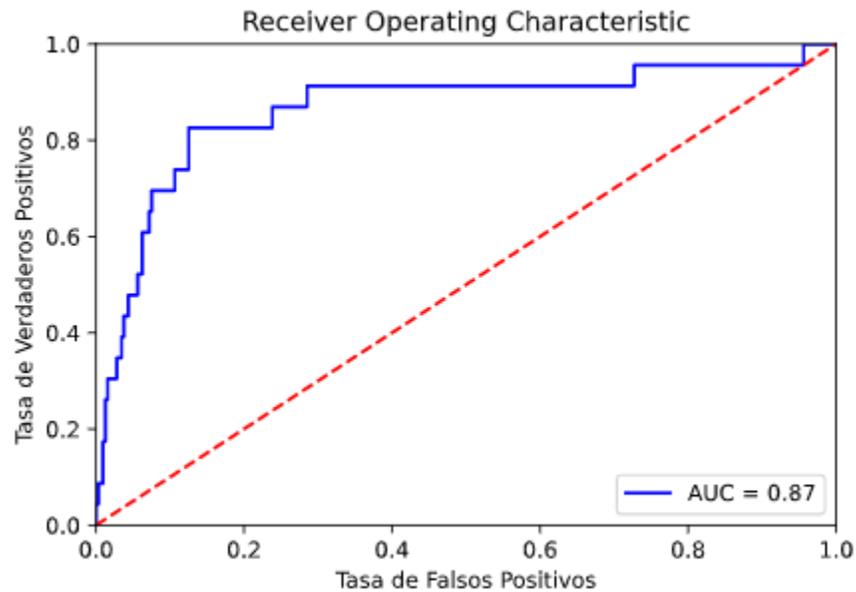
Posteriormente, con el modelo entrenado se realiza la predicción sobre el subconjunto de la muestra denominado de prueba, para el cual se obtienen las métricas que pueden observarse en la Tabla 2. Así mismo, en la Gráfica 5 puede verse la curva ROC obtenida:

Tabla 2. Métricas de bondad de ajuste del modelo entrenado con datos de entrenamiento (70%) sobre datos de prueba (30% restante).

Accuracy	Sensibilidad	Especificidad	AUC
90%	70%	91%	87%

Fuente: Elaboración propia.

Gráfica 5. Curva ROC del modelo entrenado.



Fuente: Elaboración propia.

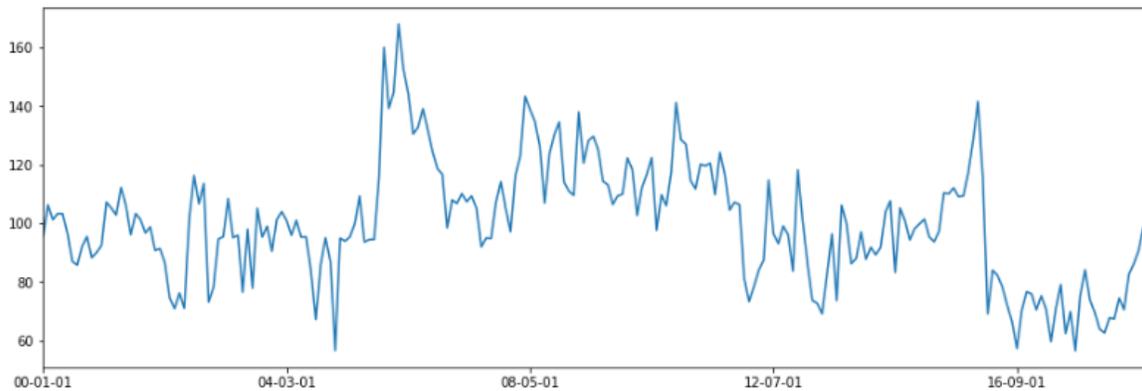
Se puede apreciar que la red neuronal presenta en general un buen ajuste, respecto a la predicción de la clase positiva. En adición a esto, vale la pena mencionar que este criterio de selección es utilizado para comparar distintos modelos cuyo desempeño no resulta mejor. De esta forma, se evalúan algoritmos como Naive Bayes (Raschka, 2014), Support Vector Machine (Tobback et al., 2018) y Random Forest (Breiman, 2001).

El modelo es finalmente utilizado para predecir sobre el total de los 1'642.894 artículos analizados. A partir de la cantidad de artículos clasificados como relevantes para cada mes, se utiliza el siguiente procedimiento para construir el índice EPU (Baker et al., 2016):

1. Se toma la cantidad de artículos relevantes y se divide sobre el total de artículos para cada mes.
2. Se divide la serie obtenida entre su desviación estándar.
3. Por último, la serie se multiplica por $100/\mu$ donde μ corresponde a la media del proceso.

El índice obtenido puede observarse en la siguiente gráfica:

Gráfica 6. Índice EPU construido mediante aprendizaje supervisado.



Fuente: Baker et al. (2016). Elaboración propia.

3.2 Construcción de índice mediante aprendizaje no supervisado

Para construir el índice mediante aprendizaje no supervisado se sigue la metodología propuesta por Azqueta-Gavaldón (2017). En este caso el autor utiliza el modelo llamado Latent Dirichlet Allocation o LDA para discriminar entre artículos que abordan el tema de la incertidumbre de política económica y los que no.

Esta metodología parte del conjunto de artículos que contienen las palabras economía o incertidumbre, en cualquiera de sus flexiones. Por tanto, comparte con la metodología de Baker et al. (2016) el inconveniente de depender de un conjunto de palabras preseleccionado. No obstante, esta dependencia es menor, dado que excluye la restricción de un conjunto de palabras asociado a la política. La hipótesis es que un artículo no aborda necesariamente el tema de incertidumbre de política económica, a pesar de incluir términos relacionados con economía e incertidumbre. Al truncar el corpus mediante este conjunto de palabras, la cantidad de artículos se reduce a 10.328.

La nube de palabras que representa los términos más comunes dentro de esta colección se puede observar a continuación:

El algoritmo LDA es un modelo probabilístico bayesiano no supervisado para colecciones de documentos que permite modelar cada elemento de la colección como la mezcla de un conjunto latente de temas. De la misma forma, cada tema es modelado como una mezcla de los términos que lo componen (Blei et al., 2003). El modelo LDA parte de una matriz dt sin normalización por IDF, de forma que la matriz está compuesta por la frecuencia de los términos en los documentos.

Este modelo aprende dos distribuciones latentes: la distribución de palabras que definen un tema y la distribución de temas que definen un artículo. Se estiman los parámetros de las distribuciones que maximizan la probabilidad de que cada palabra aparezca en cada artículo, supeditado a una cantidad previamente establecida de temas K (Azqueta-Gavaldon, 2017). Es así como, la probabilidad de que una palabra w_i aparezca en un artículo se calcula como:

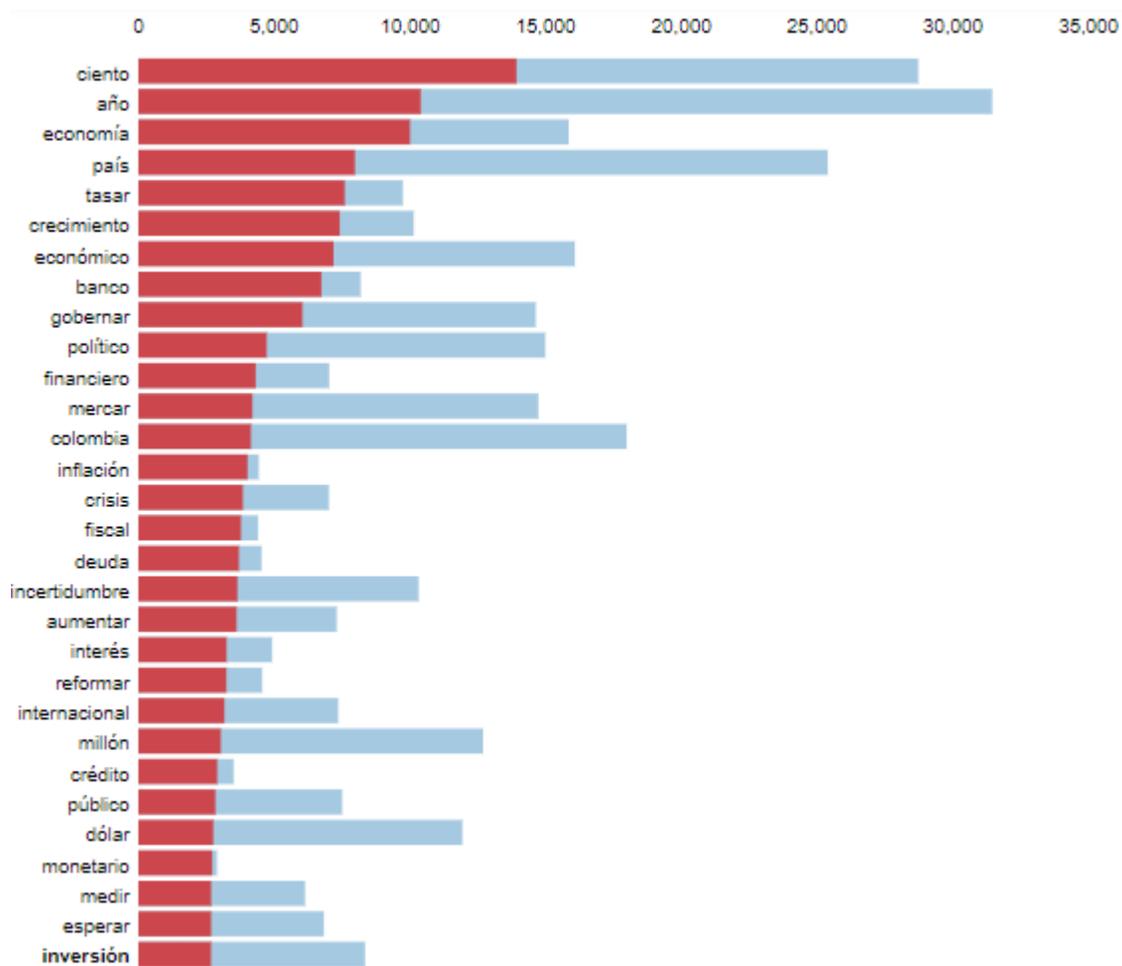
$$P(w_i) = \sum_{j=1}^K P(w_i|z_i = j) P(z_i = j)$$

Donde z_i es una variable latente que representa el tema del que es extraída la palabra i . El objetivo del modelo es maximizar $P(w_i|z_i = j)$ y $P(z_i = j)$. La distribución que rige estas probabilidades es la conocida como Dirichlet o distribución beta multivariada y la estimación del modelo se realiza mediante máxima verosimilitud. Se requieren tres parámetros previos a la estimación del modelo: el número de temas a extraer denotado como K, que corresponde a la dimensión de la distribución multivariada, y los parámetros α y β , que denotan los priores de las distribuciones de temas y palabras respectivamente. Estos últimos se ajustan a $1/K$.

El número de temas se escoge mediante una búsqueda de rejilla, i.e. grid search. En este tipo de algoritmo se estima el modelo múltiples veces con diferentes valores preestablecidos para un parámetro y se escoge el que demuestra mejor métrica de ajuste. En este caso, el indicador de ajuste corresponde al valor de la función de log-verosimilitud, que tuvo su máximo cuando el modelo fue estimado con 4 temas.

Las palabras que describen el tema elegido como relevante se muestran a continuación:

Gráfica 8. Tema extraído del modelo LDA.

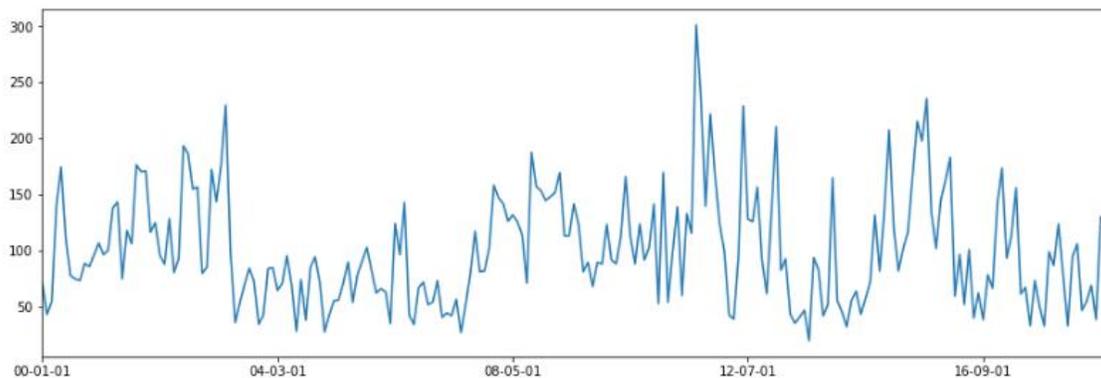


Fuente: Elaboración Propia.

En rojo puede observarse la frecuencia de cada término dentro del tema y en azul puede observarse la frecuencia del término dentro del corpus utilizado. Este tema es el que se categoriza como relevante, dado que los términos que lo conforman están más relacionados con la incertidumbre de política económica que los términos del resto de temas encontrados.

Este modelo asigna a todos los artículos la probabilidad de pertenecer a cada uno de los 4 temas elegidos como parámetros. De esta forma, las probabilidades de cada artículo de pertenecer a uno de los 4 temas se encuentran en el intervalo (0, 1) y la suma de estas 4 probabilidades da como resultado 1. Así, se plantea que los artículos para los cuales la probabilidad más alta sea la que está asociada con el tema descrito por la gráfica 8, son los entendidos como los relevantes a la hora de construir el índice. El índice construido puede observarse a continuación:

Gráfica 9. Índice EPU construido mediante aprendizaje no supervisado.



Fuente: Baker et al. (2016). Elaboración propia.

4. Relación de los índices construidos con distintas variables macroeconómicas

Las dos metodologías utilizadas y la existente abordan el problema de medir la incertidumbre de política económica desde distintos enfoques. En primer lugar, el índice de aprendizaje supervisado intenta detectar, a partir de una pequeña muestra los patrones existentes en el discurso que permiten identificar cuándo un artículo aborda este tema. Así mismo, esta metodología permite más control al investigador, al ser quien etiqueta manualmente cuándo un artículo es o no relevante.

No obstante, este método también está sujeto al riesgo de exponer al modelo de clasificación a cualquier sesgo que pueda poseer el investigador. Esta es una de las razones por las que Baker et al. (2016) desarrollan un profundo proceso de auditoría. A pesar de que no está construido para establecer un conjunto de datos inicial que permita construir un modelo predictivo, crea de todas formas un documento que guía el proceso de etiquetado.

De igual manera, el etiquetado cruzado, en el que distintas personas con formación en el área de economía clasifican de acuerdo con su interpretación de esta guía cada artículo está pensado para minimizar la posibilidad de que cualquier sesgo permee el ejercicio. Es cierto que durante el desarrollo de esta investigación no se cuenta con el recurso humano para realizar una validación exhaustiva de este proceso, pero la participación de un ayudante pretende mitigar esta problemática.

En todo caso la motivación económica detrás de la clasificación manual de los artículos seleccionados (aleatoriamente) se basa completamente en el trabajo de Baker et al (2016). De su criterio como expertos en el tema de la incertidumbre depende la calidad de este ejercicio.

Desafortunadamente, tampoco es una posibilidad hacer un análisis profundo de los artículos que el modelo encuentra como relevantes. En esta ocasión, el limitante son los recursos computacionales, que impiden retener en memoria la totalidad de los artículos y su contenido con el fin de, por ejemplo, graficar una nube de palabras. Como una pequeña muestra extraída aleatoriamente, evidentemente no representativa, se muestran dos artículos del año 2006 que la red neuronal encontró como relevantes:

Tabla 3. Muestra de artículos clasificados como relevantes por modelo de aprendizaje supervisado.

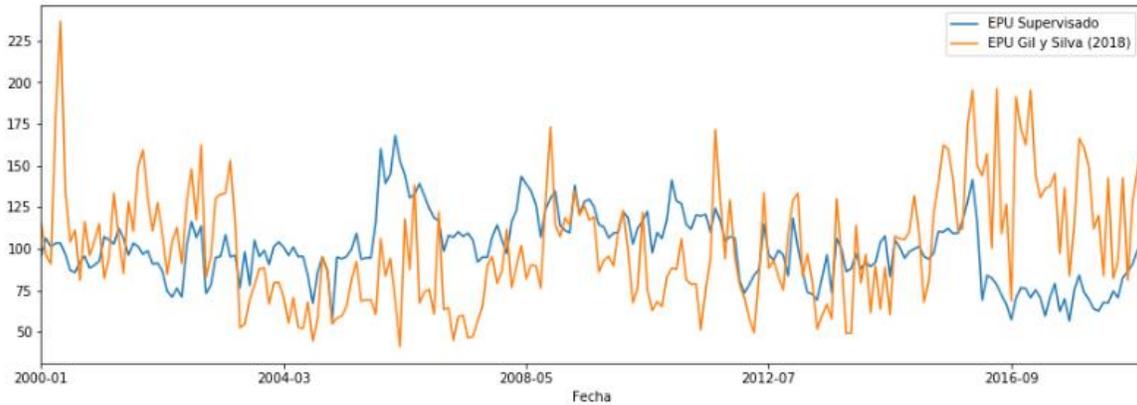
Artículo MAM-2247793	Artículo MAM-2196068
<p>Hoy, la cerveza paga un impuesto por el líquido, pero no incluye el empaque, la etiqueta, la marca, el envase y otros aspectos. Según las licoreras, de esa forma la base del impuesto no es el 100 por ciento del valor del producto, sino alrededor del 50 por ciento. El gobernador de Antioquia, Aníbal Gaviria Correa, afirmó que en este tema lo justo es que quien vende el doble de alcohol, pague el doble de los impuestos. "O todos en la cama o todos en el suelo, con las mismas condiciones", dice Gaviria. "Si las cosas se equiparan existiría una alta competitividad entre las dos industrias, que a la fecha no existe, pues las cerveceras aprovechan todo lo que se ahorran para destinarlos a una publicidad avasalladora", explica Gaviria Correa. La cervecera Bavaria dice que no rechaza la propuesta de tributar según los grados de alcohol. Sin embargo, precisa que las condiciones para hacerlo deben ser las mismas que a la fecha existen, es decir, rango de 2,5 grados a 15 grados, 134 pesos; de 15 grados a 35 grados, 219 pesos, y mayores a 35 grados 330 pesos. Sus directivos afirman que no conciben que la cerveza, el vino y algunos aperitivos se encuentren en el mismo rango que el aguardiente y el ron, pues la propuesta de las licoreras es que se deje una escala de 2,5 grados a 35 grados con un valor aproximado de 180 pesos. También indican que otra alternativa es mantener la tabla que está vigente, pero no con cobro en pesos, sino en porcentaje, pues recogería todo el valor del producto.</p> <p>'DECAE DEMANDA DE LICORERAS': ACIL</p>	<p>Esta es una meta fundamental que hay que convertirla no sólo en una fría estadística del Banco de la República sino en un gran propósito nacional. Solamente con inflaciones moderadas y predecibles las economías logran asignar eficientemente el ahorro hacia inversiones rentables y socialmente productivas. Una carestía moderada es, igualmente, condición necesaria (aunque no suficiente) para que haya una mejor distribución del ingreso y de la riqueza en el país. La meta de inflación para el 2006 es la de mantener el aumento de precios de la economía circunscrito a un margen entre el 4 y el 5 por ciento. Este objetivo debe cumplirse para continuar con la buena tendencia que se trae desde hace siete años y para preservar -lo que es fundamental- la credibilidad del Banco de la República entre los agentes del mercado. Además, las condiciones están dadas para que dicha meta se pueda cumplir perfectamente. Por ello, está bien que las autoridades monetarias sean cautelosas de que la meta se esté cumpliendo a lo largo del año. Y para tomar los correctivos que sean necesarios para asegurarse de que así suceda. Esto es lo que acaba de hacer el Banco de la República ante un leve cabeceo que presentó la inflación en el mes de agosto. En el lenguaje sibilino que suelen utilizar los bancos centrales en sus comunicados, el Banco de la República dijo en el suyo del pasado 4 de septiembre: "El aumento de la inflación en agosto (4,72 por ciento en los últimos doce meses) fue el resultado principalmente de choques de oferta en los precios de los alimentos y de los regulados. A pesar de que el equipo técnico del Banco de la República tenía previsto este repunte, no esperaba que se</p>

<p>El presidente de la junta de la Asociación Colombiana de Industrias Licoreras (Acil), Manuel Alberto Soto, manifestó que el problema data desde 1929, pero que a la fecha ya no es sólo una inconformidad, sino un perjuicio, pues las licoreras y los municipios han visto reducidas sus ventas y por ende sus ingresos. "En los últimos cinco años hemos dejado de vender 25 millones de botellas por el efecto de los altos costos del aguardiente frente a los bajos de la cerveza", dijo Soto. En las reuniones que Acil ha sostenido con el Gobierno le ha propuesto que para que exista mayor igualdad toda la industria tribute según los grados de alcohol con lo que se podría tener un impuesto que vaya de 20 a 35 grados de 180 pesos. "Con esta propuesta solucionaríamos parte del déficit del Gobierno y habría mayor equidad en los tratamientos tributarios. También hemos conversado con SAB Miller y ellos afirman que la tributación en Colombia es alta, por ello temen un nuevo incremento que implique una reducción en la demanda de sus productos", expresó Soto.</p>	<p>sucediera antes de fin de año. La Junta Directiva del Banco de la República ha señalado que en las actuales condiciones la economía no requiere del mismo estímulo monetario de antes para operar satisfactoriamente; en consecuencia, ha aumentado las tasas de interés de intervención de las operaciones a través de las cuales el banco otorga o recoge liquidez del mercado en 0,75 por ciento en lo que va corrido del año". En buen romance lo que está advirtiendo el Banco Emisor es que mantendrá un monitoreo muy estricto de la oferta monetaria (que ya no necesita crecer a los ritmos a los que venía haciéndolo), y que si observa algún desfase con relación a las metas de inflación no le temblará el pulso para seguir subiendo las tasas de interés. Está bien que el Emisor mantenga a raya la inflación, evitando en lo posible elevar el costo del dinero (para no golpear al crecimiento económico y la generación de empleo), pero sin dejar de hacerlo si es absolutamente indispensable para contener las presiones inflacionarias. "Sería ideal no elevar el costo del dinero, pero debe hacerse si es indispensable para contener el alza de precios".</p>
--	---

Fuente: Periódico El Tiempo.

Entre tanto, para el índice construido a partir del modelo de aprendizaje no supervisado, se obtiene que los artículos designados como relevantes, se pueden representar a partir de la siguiente nube de palabras:

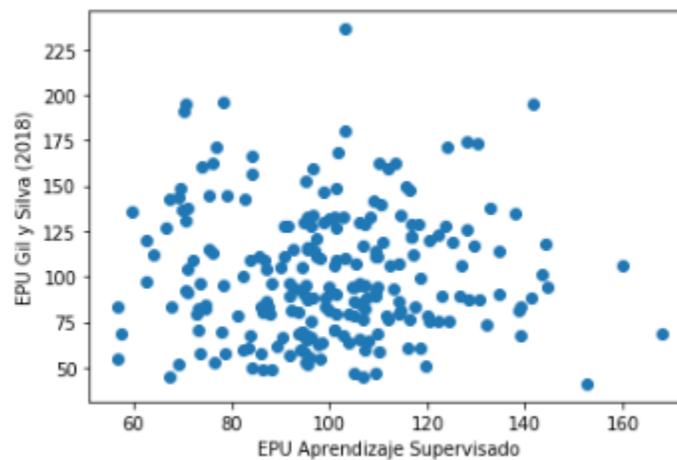
Gráfica 11. Índice EPU construido mediante aprendizaje supervisado vs índice EPU construido por Gil y Silva (2018)



Fuente: Gil y Silva (2018). Elaboración propia.

Puede observarse que no existe una correlación clara entre los indicadores. De hecho, el índice de correlación lineal entre ambos es de $-0,002$ con lo cual puede decirse que no presentan relación el uno con el otro. Esto puede apreciarse más claramente mediante un diagrama de dispersión:

Gráfica 12. Diagrama de dispersión entre índice EPU e índice EPU construido por Gil y Silva (2018)



Fuente: Gil y Silva (2018). Elaboración propia.

En la ilustración anterior no sólo no es posible observar una relación lineal entre ambas variables, sino que tampoco es claro que exista una relación polinomial en otro grado. Podría decirse que ambos índices tienen poco en común, lo cual es un resultado interesante dado que intentan medir el mismo fenómeno.

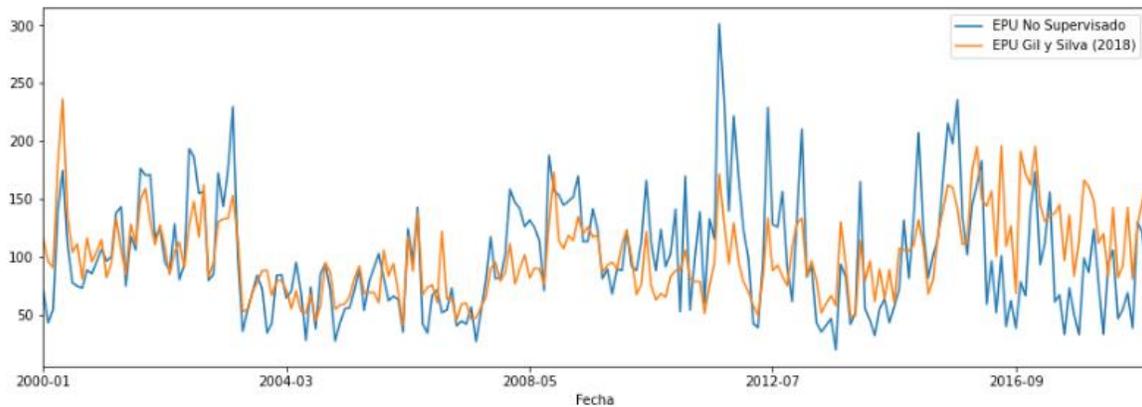
Al respecto, hay algunos datos que pueden arrojar luz sobre la razón de esta diferencia. Durante el desarrollo de esta investigación no se accede a la fuente de datos original utilizada por Gil y Silva (2018), en tanto que en la página web que funciona como recurso para acceder a su investigación únicamente se expone el índice estimado y no las variables en función de las cuáles se calcula, como son el total de artículos con los que trabajan y la cantidad entre estos que encuentran como relevantes, es decir, que abordan el tema de incertidumbre de política económica dado que contienen las palabras claves definidas en su metodología. Tampoco mencionan el procedimiento utilizado para obtener la información.

De esta forma, es difícil saber si se tiene como base el mismo corpus documental. No obstante, es claro que los autores trabajan con un corpus de menor tamaño que el utilizado para calcular el índice EPU mediante aprendizaje no supervisado, puesto que este comprende artículos que incluyen palabras relacionadas con economía, incertidumbre y política, en tanto que el utilizado en esta investigación, como se menciona en la sección anterior, no incluye palabras asociadas a política, por lo que comprende una selección menos restrictiva. Este último subconjunto comprende 10.328 artículos.

Entre tanto, del modelo de aprendizaje supervisado se utiliza para clasificar 1'642.894 artículos, de los cuales estima que 189.440 abordan el tema de la incertidumbre de política económica. En contraste, la metodología que enmarca la utilización del modelo no supervisado encuentra como relevantes 2.960 artículos sobre el mismo total general, una proporción notablemente menor. El hecho de que esta diferencia de magnitudes no sea tan obvia al momento de visualizar los índices se debe a que las transformaciones realizadas sobre las series garantizan que sea cual sea la cantidad de artículos seleccionados, éstas se muevan sobre la misma escala.

Por otra parte, la evolución del índice construido mediante aprendizaje no supervisado en comparación con el desarrollado mediante la metodología tradicional puede verse en la siguiente gráfica:

Gráfica 13. Índice EPU construido mediante aprendizaje no supervisado vs índice EPU construido por Gil y Silva (2018)

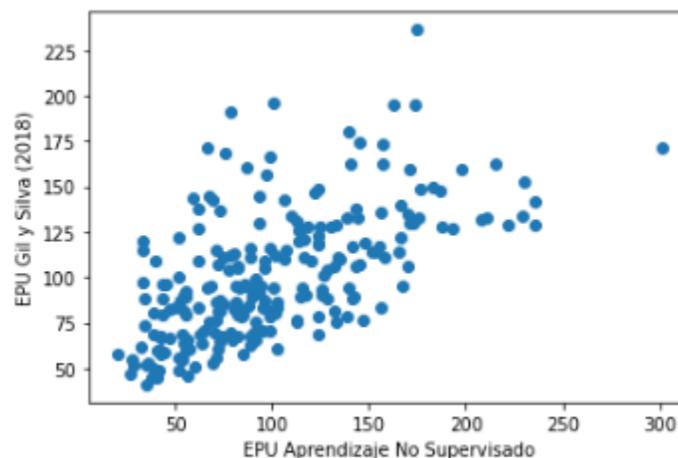


Fuente: Gil y Silva (2018). Elaboración propia.

En este caso, parece que existe una relación mucho más clara entre ambos índices. En principio, esto tiene sentido, dado el hecho de que ambos se construyen a partir de un conjunto de artículos similar. El coeficiente de correlación lineal, que para estos dos indicadores asciende a 0,602, parece apoyar esta idea. No obstante, a simple vista puede decirse también que los índices aparentan tener una respuesta diferente durante los periodos en los cuales se producen choques de incertidumbre. Así, el índice EPU construido mediante aprendizaje no supervisado parece reaccionar mucho más fuerte en estos periodos.

Al construir un diagrama de dispersión, esta idea se ve reforzada:

Gráfica 14. Diagrama de dispersión entre índice EPU e índice EPU construido por Gil y Silva (2018)



Fuente: Gil y Silva (2018). Elaboración propia.

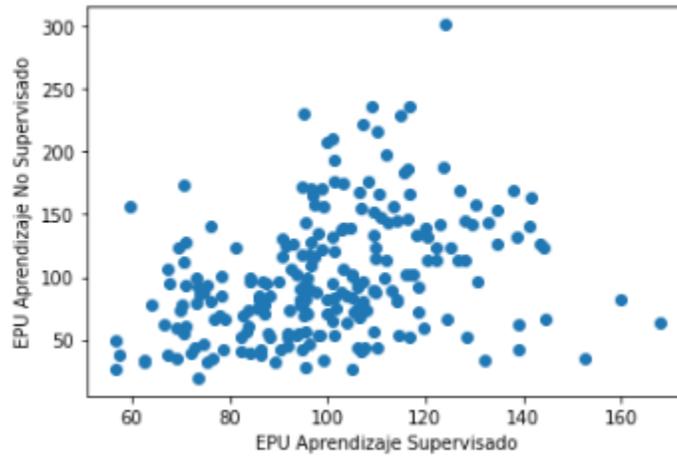
En línea con lo anteriormente mencionado, en esta gráfica puede observarse cómo para los niveles más bajos de ambos indicadores existe una relación lineal mucho más clara, entre tanto, para valores más altos se produce una dispersión más alta. Nuevamente, esto puede sugerir que para niveles más moderados de incertidumbre los indicadores reaccionan de manera similar, entre tanto, para niveles más elevados la respuesta de ambos indicadores empieza a exhibir un comportamiento disímil.

Este hallazgo es interesante dado que Azqueta-Gavaldón (2017) específicamente propone esta metodología como una herramienta para hallar un índice que se comporte de manera lo suficientemente similar al desarrollado mediante palabras clave. Su hipótesis se ve confirmada, como se menciona en la sección 1 de este documento, dado que encuentra un índice con una correlación lineal de 0,94. En adición a esto menciona que la correlación entre los componentes de ciclo es del 0,88 y entre los de tendencia es de 0,99.

Este resultado no es coherente con lo observado en el caso colombiano al utilizar un filtro de Hodrick y Prescott (Hodrick & Prescott, 1997). En esta ocasión, la correlación lineal entre los componentes de tendencia de los índices es de 0,004, por lo cual es casi nula, entretanto entre los componentes de ciclo es de 0,71, con lo que se evidencia una relación mucho más clara. De esta forma, puede sugerirse que los índices suelen reaccionar a los choques de incertidumbre en la misma dirección, aunque en el caso del índice construido mediante aprendizaje no supervisado la magnitud del choque sea mayor, en cambio su evolución tendencial a lo largo del tiempo es disímil.

Por último, vale la pena exponer el comportamiento conjunto de los índices construidos mediante aprendizaje supervisado y no supervisado, de igual forma a partir de un diagrama de dispersión:

Gráfica 15. Diagrama de dispersión entre el índice EPU construido a partir de aprendizaje supervisado y el construido a partir de aprendizaje no supervisado.



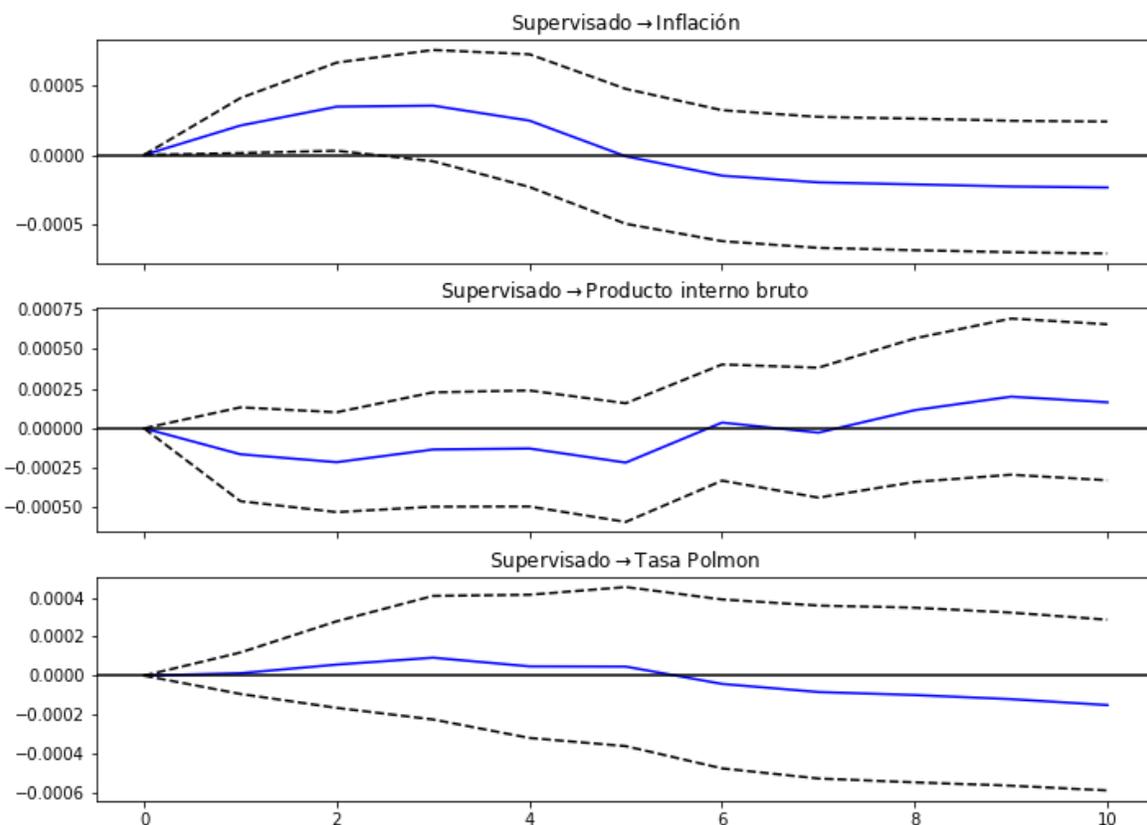
Fuente: Elaboración propia.

En este caso, tampoco puede apreciarse una relación lineal clara, con un coeficiente de correlación lineal de 0,317. A simple vista, tampoco parece intuitivo que ambos indicadores intentan reflejar el mismo fenómeno.

Con cada uno de los índices analizados se construye un modelo VAR (Luetkepohl, 2005), que analiza su relación con tres indicadores líderes de la economía colombiana: la inflación, la tasa de política monetaria del Banco de la República y el crecimiento económico. Las variables elegidas corresponden a un modelo simple que suele ser utilizado para representar las dinámicas macroeconómicas (Orphanides & Wei, 2012). Se prueba un máximo de seis rezagos para cada modelo, y se elige la cantidad de rezagos que mejor criterio de información de Akaike presente. Por último, entre el mejor modelo para cada uno de los tres índices se escoge el más apropiado, con base tanto en el criterio de Akaike como en el valor de la función de log-verosimilitud.

En este caso, el modelo construido a partir del índice EPU desarrollado mediante aprendizaje supervisado resulta ser el que mejores métricas presenta. La gráfica de las funciones impulso respuesta, cuando la perturbación se produce en la variable que representa el índice EPU, pueden observarse a continuación:

Gráfica 16. Gráfica de las funciones impulso respuesta del modelo seleccionado, con el índice EPU supervisado como variable de interés



Fuente: Gil y Silva (2018). Elaboración propia.

En las visualizaciones anteriores se puede apreciar cómo un choque de incertidumbre sugiere un aumento en el corto plazo de la inflación y la tasa de política monetaria, en tanto que parece tener un efecto inverso en el crecimiento económico. Esto sugiere que la hipótesis planteada anteriormente es acertada, dado que un choque de incertidumbre de política monetaria muestra un efecto adverso en el desempeño económico del país. A su vez, esto pone de relevancia la importancia que tiene la existencia de un canal de comunicaciones claro y honesto entre las autoridades de política y el público.

Respecto al comportamiento de este índice EPU y su relación con eventos económicos, como puede verse en la gráfica 6, el pico del índice se encuentra en febrero de 2006. Al revisar el contenido de los artículos seleccionados como relevantes para ese mes, se encuentra que la mayoría aborda el tema del TLC con Estados Unidos, cuyas negociaciones concluyeron dicho mes, a pesar de haber sido aprobado por los congresos de ambos países algunos años después. Así mismo, puede notarse un pico durante principios de 2016, periodo durante el cual se debate la venta de Isagén, los precios del

petróleo son bajos y la inflación aumenta y se encuentra por fuera del rango meta del Banco de la República.

Por otra parte, es notorio el decrecimiento en los niveles medidos de incertidumbre a partir de los últimos meses de 2016 hasta el final del periodo abordado por la investigación, en 2018. El inicio de este intervalo corresponde a la época durante la cual se discute y se vota el plebiscito por el acuerdo de paz. Al revisar los artículos clasificados como relevantes durante este periodo, se encuentra que se abordan temas como la inminente reforma tributaria y la elección presidencial estadounidense, por lo que se puede afirmar que el modelo no excluye estos fenómenos.

No caben dudas de que el TLC con los Estados Unidos ha sido uno de los acontecimientos económicos más relevantes en la historia reciente de Colombia, tanto a nivel político como económico. De esta forma, es coherente observar el efecto que tuvo sobre la incertidumbre de política económica en el país. Dados los resultados presentados en la gráfica 16, es importante reconocer como la proposición de una política económica puede derivar en consecuencias adversas para el país, sea por los efectos de un mal contenido o por la polarización.

5. Conclusiones y Recomendaciones

En la literatura, se suele establecer que un entorno de alta incertidumbre tiene consecuencias económicas desfavorables para un país. De tal forma, encontrar un indicador que permite medir la incertidumbre respecto a la política económica nacional facilita el análisis del entorno macroeconómico del país, y puede ayudar a explicar por qué en algunas situaciones decisiones de inversión y/o de consumo se ven postergadas.

En el presente trabajo de investigación, se concluye que un índice construido a partir de un modelo de aprendizaje supervisado, a pesar de que su cálculo es más costoso en términos de tiempo en comparación con uno construido a partir de un modelo de aprendizaje no supervisado, evidencia una mejor relación con distintos indicadores macroeconómicos, al igual que frente al construido a partir de la búsqueda de palabras clave como el desarrollado por Baker et al (2016) y replicado para Colombia por Gil y Silva (2018).

Así mismo, se encuentra que choques de incertidumbre de política económica tienen un efecto adverso en el desempeño macroeconómico colombiano, dado que un incremento en el índice EPU influye en un aumento tanto en la inflación como en la tasa de interés del

Banco de la República en el corto plazo, mientras que está asociado a una caída en el crecimiento del Producto Interno Bruto. Este resultado pone de relevancia la existencia de un canal de comunicación claro y honesto entre las autoridades de política monetaria y el público.

Una limitación en la construcción del IPE es que el origen de los artículos utilizados en el presente trabajo sea de una única fuente, que, como todos los medios de comunicación, se ve sujeta al sesgo político de su agenda. No obstante, la poca disponibilidad en Colombia de un archivo digital de prensa confiable durante el periodo abarcado por el estudio (2000–2018), pone de manifiesto la necesidad de utilizar esta fuente, debido a la metodología propuesta.

Entre los temas por revisar en futuras investigaciones, se podrían utilizar otras metodologías que refinen la precisión con la cual el índice calculado mide la incertidumbre de política económica. Una posibilidad sería usar metodologías que involucren n-grams, o unidades de texto compuestas por más de una palabra, con el fin de extraer patrones que involucren expresiones más complejas.

En adición a esto, el componente ideológico editorial al que se ve sujeta la fuente de datos de la presente investigación puede tener un efecto no observado en este trabajo sobre el resultado. A futuro, se puede esperar que una mayor diversidad y disponibilidad de datos permita que se pueda capturar el discurso desde múltiples aristas del espectro.

Por su parte, podría etiquetarse cada artículo con más detalle que si trata o no sobre incertidumbre de política económica, como si además está relacionado con el sector salud o fiscal, de acuerdo con lo propuesto por Baker et al (2016) y construir índices más específicos. De esta manera, se podría realizar un análisis más profundo, con el objetivo de observar cómo la incertidumbre afecta distintos sectores de manera diferenciada.

Adicionalmente, se abre campo a la construcción de índices que permitan aproximar la medición de otro tipo de fenómenos y evaluar su efecto en el desempeño económico del país. Para el caso de Colombia, en particular, un índice que permita medir la incertidumbre provocada por el conflicto armado puede resultar muy útil. No obstante, esta metodología también permite abordar fenómenos como el cambio climático o la migración, tanto interna como externa.

Referencias Bibliográficas

- Azqueta-Gavaldon, A. (2017). Developing News-Based Economic Policy Uncertainty Index with Unsupervised Machine Learning. *Economics Letters*, 158, 47–50. Retrieved from <http://www.sciencedirect.com/science/journal/01651765>
- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *Quarterly Journal of Economics*, 131(4), 1593–1636. Retrieved from <https://academic.oup.com/qje/issue>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. In *Journal of Machine Learning Research* (Vol. 3). <https://doi.org/10.1016/b978-0-12-411519-4.00006-9>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Gil, M., Silva D. (2018). Economic Policy Uncertainty Index for Colombia. Retrieved from <https://www.policyuncertainty.com/colombia.html>
- Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2), 217–288. <https://doi.org/10.1137/090771806>
- Hodrick, R. J., & Prescott, E. C. (1997). Postwar U.S. Business Cycles: An Empirical Investigation. *Journal of Money, Credit and Banking*, 29(1), 1–16. <https://doi.org/10.2307/2953682>
- International Monetary Fund. (2016). *World Economic Outlook: Too Slow for Too Long*. Washington, April 2016. Retrieved from <https://www.imf.org/en/Publications/WEO/Issues/2016/12/31/Too-Slow-for-Too-Long>
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, 1, 448–456. International Machine Learning Society (IMLS).
- Luetkepohl, H. (2005). The New Introduction to Multiple Time Series Analysis. In *New Introduction to Multiple Time Series Analysis*. <https://doi.org/10.1007/978-3-540-27752-1>
- Orphanides, A., & Wei, M. (2012). Evolving Macroeconomic Perceptions and the Term Structure of Interest Rates. *Journal of Economic Dynamics and Control*, 36(2), 239–254. Retrieved from

[http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1278070&lang=es
&site=ehost-live](http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1278070&lang=es&site=ehost-live)

Raschka, S. (2014). Naive Bayes and Text Classification {I} - Introduction and Theory.

CoRR, *abs/1410.5*. Retrieved from <http://arxiv.org/abs/1410.5329>

Skansi, S. (2018). *Introduction to Deep Learning* (1st ed.). <https://doi.org/10.1007/978-3-319-73004-2>

Tobback, E., Naudts, H., Daelemans, W., Fortuny, E. J. de, & Martens, D. (2018). Belgian Economic Policy Uncertainty Index: Improvement through Text Mining. *International Journal of Forecasting*, *34*(2), 355–365. Retrieved from

[http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1722032&lang=es
&site=ehost-live](http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1722032&lang=es&site=ehost-live)

Weiss, S., Indurkha, N., & Zhang, T. (2015). *Fundamentals of Predictive Text Mining*.

<https://doi.org/10.1007/978-1-4471-6750-1>