

# **RELACIÓN ENTRE RIESGO PAÍS E ÍNDICES BASADOS EN INFORMACIÓN NO ESTRUCTURADA. EVIDENCIA PARA ARGENTINA**

*Llada Martín*

Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires,  
Argentina.

CONICET-Universidad de Buenos Aires. Instituto Interdisciplinario de Economía  
Política de Buenos Aires. Buenos Aires, Argentina.

[lladamartin@gmail.com](mailto:lladamartin@gmail.com)

## **RESUMEN.**

Este trabajo evalúa si ciertos indicadores construidos a partir de información no estructurada publicada en los periódicos, poseen capacidad de explicar la evolución del riesgo país de Argentina, aproximado por el índice EMBI construido por el JP Morgan. En primer lugar, se entrena el modelo Asignación Latente de Dirichlet (LDA) a fin de identificar aquellos artículos de la prensa que se asocian al tópico riesgo país. En segundo lugar, se construyen indicadores cuantitativos que resumen diferentes estados subjetivos asociados a dicho tópico, siguiendo diferentes técnicas de análisis de texto propuestas en la literatura. Por último, se realizan diversos ejercicios estadísticos, en los cuales se evidencia que los indicadores basados en información no estructurada poseen capacidad de explicar la evolución contemporánea del riesgo país, controlando por diferentes indicadores macroeconómicos que caracterizan las condiciones internas y externas de la economía.

## **1. INTRODUCCIÓN.**

La percepción de los inversores asociada a la capacidad que tiene un determinado país respecto a honrar sus compromisos externos es reflejada por la prima de riesgo soberano. La percepción pesimista asociada a estos fenómenos se refleja en alzas del riesgo país. El deterioro de este tipo de indicadores no es indiferente para los hacedores de política económica, teniendo en cuenta su potencial impacto sobre el costo marginal de financiación de su deuda externa, la estabilidad macroeconómica y sobre sus posibilidades de desarrollo sostenido.

Este trabajo estudia el comportamiento del riesgo país de Argentina, aproximado por el EMBI spread (Emerging Market Bond Index spread) computado y divulgado por el J.P. Morgan de Estados Unidos. Específicamente, a partir de ejercicios estadísticos se estudia

el grado en que un indicador cuantitativo de información económica no estructurada publicada en el periódico La Nación, tienen capacidad de explicar la dinámica de la sobretasa que pagan los activos de renta fija denominados en dólares emitido por el gobierno argentino respecto a activos similares que emite los países desarrollados. Adicionalmente se incluyen en el análisis indicadores económicos relacionados con la posición interna del país e indicadores relacionados con la posición externa del país, siguiendo trabajos previos (Edwards 1983, Rowland 2004, Uribe y Yue 2006, Azar y coautores 2007, González-Rosada y Levi Yeyati 2008, Mpapalika y Malikane 2019).

Cada día existe información estructurada y no estructurada disponible potencialmente relevante para ser utilizada por los agentes de la economía para formular expectativas y tomar decisiones, las cuales influyen sobre dinámica de las variables económicas reales y financieras. La literatura se ha concentrado en estudiar la información que capturan los indicadores económicos tradicionales, sin embargo, existe pocos trabajos que evidencian la utilidad de explotar la información no estructurada en el campo de la economía (por ejemplo, Tetlock 2007, García 2013, Loughran y McDonald 2011, Baker y coautores 2016, Aromí 2018, entre otros). En este sentido, en pos de explotar estas fuentes de información no tradicionales, se computan indicadores subjetivos basados en información publicada en la sección de economía del diario La Nación. Con este fin, en primer lugar, se entrenó un modelo de aprendizaje automático no supervisado denominado Asignación Latente Dirichlet (LDA) con el objeto de identificar todos aquellos artículos asociados al tópico riesgo país. Por último, siguiendo Aromí (2018) se computaron indicadores cuantitativos con el objeto de aproximar diferentes manifestaciones de estados subjetivos asociados a este tópico, utilizando diccionarios de palabras predefinidos por expertos (Tetlock 2007) y listas de palabras de similar significado que surgen a partir de entrenar un modelo de aprendizaje automático no supervisado de representación vectorial de palabras (Pennington y coautores 2014).

A partir de un conjunto de ejercicios empíricos, se evidencia que ciertos indicadores subjetivos construidos en base a la información no estructurada sobre el tópico riesgo país, comunicada en un diario de cobertura nacional, tienen capacidad de explicar la evolución contemporánea del índice EMBI argentino, más allá de la información que comunican los indicadores económicos utilizados en la literatura. En particular, se evidencia que un incremento en una desviación estándar en el indicador que captura

manifestaciones de temor en estos artículos genera, en promedio, un incremento del 8% en la variación de la prima de riesgo aproximada por el índice EMBI. Adicionalmente, se construyó un indicador que resume la variabilidad de estos indicadores construidos a partir de múltiples fuentes de información comunicada en los diarios La Nación y Wall Street Journal. Los resultados reportados sugieren que este indicador es estadísticamente significativo y, por tanto, también tiene capacidad de explicar la evolución del EMBI argentino. Por otro lado, respecto de los fundamentales de la economía, se evidencia que múltiples indicadores económicos tienen capacidad de explicar la evolución del EMBI. En este sentido, se encuentra que el ratio entre deuda externa y PBI, el tipo de cambio real, la tasa de interés real, la relación entre reservas internacionales y exportaciones son estadísticamente significativos, mientras que no se encuentra el signo esperado para la tasa de interés internacional.

Luego de esta introducción, en la siguiente sección se describe la metodología. La sección 3 detalla las variables y las fuentes de datos utilizadas. A continuación, en la sección 4 se discuten los resultados principales. Por último, la sección 5 resume las conclusiones.

## **2. METODOLOGÍA.**

Cada vez es más habitual observar cómo se incrementa el volumen de datos disponibles (estructurados y no estructurados) que pueden utilizar los individuos para tomar sus decisiones económicas. Los trabajos empíricos tradicionalmente evalúan la relación entre los fundamentales económicos y el riesgo país. En este trabajo, incorporamos un atributo novedoso que intenta capturar diferentes manifestaciones subjetivas en las noticias de los diarios en pos de evaluar su capacidad para explicar la evolución de la prima de riesgo que pagan los gobiernos soberanos. En este apartado, exhibimos la metodología asociada a la construcción de estos últimos indicadores. En primer lugar, aplicamos un modelo de tópicos en pos de concentrarnos en aquellos artículos que se asocian con el tópico riesgo país. En segundo lugar, construimos los indicadores cuantitativos utilizando diccionarios de palabras previamente definidas y listas de palabras que surgen de aplicar un algoritmo denominado GloVe.

### **2.1 TOPIC MODEL.**

El modelo de tópicos es una herramienta estadística dentro del campo de aprendizaje automático desarrollada para extraer los tópicos principales que trata una colección de documento no estructurada. En este sentido, el modelo utiliza las variables observadas para descubrir automáticamente el conjunto de tópicos y asigna los documentos a estos tópicos. El modelo Asignación Latente de Dirichlet (LDA) es el más utilizado dentro de topic model, el cual fue propuesto por Blei y coatures (2003). Este modelo, descubre los principales tópicos en cada documento como así también la proporción en la cual cada artículo exhibe estos tópicos, dada una colección de documentos. Como sugieren Calvo-González y coatures 2018, una característica atractiva de este algoritmo es que el investigador debe setear, de manera arbitraria, el número de tópicos en los cuales los documentos serán clasificados. A su vez, el descubrimiento de estos tópicos se realizará utilizando las palabras de los documentos, ya que estas son las únicas variables observables del modelo. Por último, los supuestos que impone el modelo LDA son que los documentos siguen una distribución de probabilidad aleatoria sobre los tópicos, los documentos abordan múltiples tópicos con diferentes probabilidades y los tópicos contienen palabras con diferentes probabilidades. Otros supuestos que realiza LDA son que el orden de las palabras con el cual aparecen en el documento no es relevante, es decir, utiliza el supuesto de bolsas de palabra<sup>1</sup>. Además, supone que el orden del documento no es relevante y el número de tópicos es conocido y fijo (Blei, 2012).

El supuesto fundamental del LDA es que los documentos observados fueron generados a través de un proceso de probabilidad generativo. Sin embargo, los parámetros de este proceso son latentes u ocultos, lo cual caracteriza la tarea de estimación inferencial de LDA. Esta tarea se asocia con estimar la estructura latente: los tópicos y la composición de los tópicos de cada documento. Esta tarea se realiza a través de un proceso generativo en sentido inverso, es decir, este modelo utiliza las palabras de cada documento para estimar los parámetros latentes del modelo generativo que con mayor probabilidad generaron la colección de documentos observados.

Siguiendo Calvo-González y coatures (2018), el proceso que sigue LDA luego de especificar el número de tópicos  $k$ , es:

---

<sup>1</sup> Si bien ese supuesto es poco realista, parece razonable si el objetivo de este modelo es descubrir la estructura semántica de los textos.

- 1- Para cada t3pico  $k$ , se extrae la distribuci3n sobre las palabras  $\varphi_k$  acorde a la distribuci3n Dirichlet  $\sim Dir(\beta)$ , donde  $\beta$  es el par3metro prior de la distribuci3n Dirichlet de palabras por t3pico<sup>2</sup>
- 2- Para cada documento  $D$ :
  - a. Se extrae un vector de proporciones del t3pico  $\theta_d$  acorde con la distribuci3n de Dirichlet  $\sim Dir(\alpha)$ , donde  $\alpha$  es el par3metro prior de distribuci3n Dirichlet de t3pico por documento.<sup>3</sup>
  - b. Para cada una de las  $N$  palabras  $\omega_n$ :
    - i. Se extrae una asignaci3n de t3picos  $Z_n$  acorde con una distribuci3n multinomial  $\sim Multinomial(\theta)$  acorde con la proporci3n de del t3pico  $\theta_d$
    - ii. Selecci3n una palabra  $\omega_n$  de  $p(\omega_n | Z_n, \varphi)$ , una distribuci3n de probabilidad condicional multinomial sobre el t3pico  $Z_n$

Como mencionamos anteriormente, la tarea inferencial de LDA consiste en inferir la estructura latente m3s probable (la distribuci3n de los par3metros  $\theta, Z$  y  $\varphi$ )<sup>4</sup> que haya generado los documentos observados. El objetivo de LDA es computar la distribuci3n a posteriori condicional de las variables latentes dada los documentos y los par3metros a priori de Dirichlet:

$$p(\theta, Z, \varphi | \omega, \alpha, \beta) = \frac{p(\theta, Z, \varphi | \alpha, \beta)}{p(\omega | \alpha, \beta)}$$

La evaluaci3n del denominador es computacionalmente inviable, ya que implica integrar la distribuci3n sobre los par3metros latentes. Con el objeto de aproximar esta distribuci3n a posteriori, se emplea un algoritmo de inferencia denominado muestreo de Gibbs

---

<sup>2</sup> El par3metro beta representa la creencia a priori sobre la distribuci3n de las palabras por t3pico. Un valor m3s elevado de beta significa que cada t3pico es probable que este compuesto con la mayor3a de las palabras del corpus, mientras que un valor m3s peque1o indica que cada t3pico va a contener pocas palabras.

<sup>3</sup> El par3metro alfa representa la creencia a priori sobre la distribuci3n de t3picos por documento. Un valor m3s elevado de alfa representa que cada documento es probable que contenga una proporci3n elevada de t3picos, mientras que un valor de alfa bajo significa que cada documento es probable que contenga poco t3picos.

<sup>4</sup> Donde  $Z$  representa la asignaci3n de t3pico por palabra,  $\theta$  representa la distribuci3n de t3picos por documento, la cual indica en que magnitud cada documento pertenece a un t3pico y  $\varphi$  representa la dsitribucion de las palabras por t3pico  $k$ , la cual es utilizada para definir la estructura sem3ntica del contenido de cada t3pico.

colapsado propuesto por Griffiths y Steyvers (2004) para estimar  $\beta$ ,  $\theta$  y  $z$ , los parámetros latentes.

Con el fin de estimar los parámetros que rigen los pesos de los temas y las distribuciones de palabras del corpus de texto observado, el proceso generativo descrito anteriormente debe invertirse. La distribución posterior de los parámetros latentes condicionales en el corpus de texto se puede formar dividiendo la densidad por la probabilidad de observar ese corpus.

### **2.1.1 IMPLEMENTACIÓN DE LDA.**

Para estimar el modelo LDA utilizando el enfoque descrito arriba, en primer lugar, se debe transformar el vector de artículos en una representación de espacio vectorial que capture la frecuencia de las palabras. Para esto, previamente se realiza una tokenización (tarea que se asocia a romper el texto en palabras individuales) de los artículos y se realizan diferentes tareas de limpieza estándar en la práctica de procesamiento de lenguaje natural asociadas a eliminar signos de puntuación, números, mayúsculas y “*stops words*” (términos como artículos, preposiciones, pronombres y otras palabras que no tienen un valor semántico relevante para la definición de los tópicos). Una vez realizada la reducción del vocabulario, sólo nos concentramos en aquellas palabras que ocurren al menos 30 veces de manera de reducir el espacio vectorial de palabras.

El modelo LDA se aplica sobre la sección de economía del diario La Nación con el objeto de identificar aquellos artículos que traten el tópico riesgo país. Los artículos de este diario asociados a dicha sección, fue construida a partir de técnicas de web scraping. A su vez, seteamos parámetros necesarios para el algoritmo Gibbs alfa (distribución de términos por tópico) y beta (distribución de tópicos por documento) en  $50/k$  (número de tópicos) y 0.1, respectivamente. El número de tópicos se fija en 60 y utilizamos el parámetro que define la proporción de tópico que exhibe el documento ( $\theta$ ) para seleccionar aquellos artículos cuyo parámetro de proporción asociado al tópico riesgo país sea máximo.

El modelo LDA fue entrenado con todos los artículos que cubre el periodo 1996-2004 en pos de evitar el sobreajuste del modelo. Los temas seleccionados asociados a riesgo país seleccionados son el tópico 7 (cuyas 10 palabras asociadas a él con mayor probabilidad relativa son: fmi, fondo, argentina, acuerdo, internacional, monetario, organismo, programa, washington y banco), el tópico 47 (deuda, bonos, acreedores, argentina,

gobierno, canje, reestructuración, default, títulos, oferta ), el tópico 48 (plazo, tasa, tasas, interés, anual, crédito, riesgo, créditos, financiamiento) y el tópico 57 (mercado, mercados, inversores, semana, analistas, tasas, bonos, local, interés, acciones).

## **2.2 ÍNDICE BASADOS EN INFORMACIÓN NO ESTRUCTURADA.**

Una vez construido el corpus de documentos asociados al tópico de interés, procedemos a computar los índices cuantitativos basados en información no estructurada divulgada en la prensa. En esta sección presentamos la metodología seguida en pos de computar los índices de prensa.

### **2.2.1 REPRESENTACIÓN VECTORIAL DE PALABRAS.**

En este trabajo, se utilizan modelos para aprender estructuras de significado de palabras utilizando vectores de palabras. En particular, se utiliza el modelo GloVe (Global Vectors) propuesto por Pennington y coautores (2014), el cual produce estructuras lineales de significado. Este modelo representa las palabras a través de vectores y captura las estadísticas globales del corpus, las cuales utiliza para entrenar dichos vectores y aprender el significado semántico de las palabras, a partir de los vectores que las representan. En otras palabras, permite identificar relaciones entre las palabras a través de computar la distancia entre sus vectores representativos, mediante simples operaciones algebraicas. Esta técnica de aprendizaje automático no supervisado, permite generar indicadores cuantitativos que resumen información de la prensa. A su vez, este modelo combina las ventajas de los modelos word2vec y skip-gram. Además, como sugiere Aromi (2018), el modelo GloVe posee buenos resultados en tareas como: i) resolución de la ambigüedad en el significado de una palabra, ii) identificación de entidades a partir de la composición del vector y iii) identificación de palabras que indican tono o tópico.

Múltiples índices que capturan manifestaciones de diferentes estados subjetivos se puede computar a partir de este modelo. Para dicha tarea, en primer lugar, se entrena los vectores de palabras utilizando un corpus mediante el cual se construye una estructura de significado. Es decir, para cada palabra se obtiene un vector de números que representa el significado de esta. En segundo lugar, se escoge una palabra clave o set de palabras que manifieste el estado subjetivo que se desea representar. Luego, el indicador que resume la información de la prensa es computado utilizando el vector de palabras correspondiente a ésta palabra clave previamente seleccionada como así también, las palabras fuertemente asociadas a esta las cuales son identificadas computando la distancia

entre sus respectivos vectores representativos. Por último, el indicador está dado por la frecuencia con la cual dichas palabras aparecen en un segundo corpus. Esto último, se realiza para evitar sesgo hacia el futuro. La capacidad con la cual los vectores de palabras capturan el significado de una palabra, depende de cuan informativo sea el corpus de entrenamiento y la efectividad del modelo de aprendizaje.

En el modelo GloVe, los vectores de palabras son entrenados para capturar información sobre la co-ocurrencia de palabras en el corpus de entrenamiento. Este método es global ya que todos los vectores son computados mediante un ejercicio de optimización individual. Utiliza un modelo de mínimos cuadrado (modelo de regresión log-bilineal). Sea  $W$  el tamaño del vocabulario<sup>5</sup> y sea  $X_{ij}$  el número de veces que ocurre la palabra  $i$  en el contexto de la palabra  $j$ . La función de pérdida del modelo viene dada por

$$\sum_{i,j=1}^W f(x_{ij})[v_i^T * \tilde{v}_j + b_i + \tilde{b}_j - \log(x_{ij})]^2$$

donde  $v_i$  y  $\tilde{v}_j$  son vectores de palabras,  $f(X_{ij})$  es una función de ponderación creciente y cóncava cuya función es ponderar las co-ocurrencias más frecuentes,<sup>6</sup>  $b_i$  y  $\tilde{b}_j$  son sesgos de palabras. La representación de vectores se entrena utilizando un gradiente descendiente estocástico (Duchi y coautores, 2011). El objetivo de entrenamiento de GloVe es aprender vectores de palabras de tal forma que el producto escalar sea igual al logaritmo de la probabilidad de co-ocurrencia de las palabras. Más detalles pueden encontrarse en Pennington y coautores (2014).

Siguiendo los parámetros que están en línea con aquellos utilizados en la literatura de representación de lenguaje natural, la dimensión del vector es 100 y el tamaño de la ventana utilizado para computar la matriz de co-ocurrencia es 5. El vocabulario utilizado en la implementación está dado por palabras con una frecuencia de 100 o más en el corpus de entrenamiento. El análisis de robustez indica que los resultados no son sensibles a variaciones en los valores de estos parámetros. La representación vectorial de palabras es realizada utilizando la librería `text2vec` en la plataforma R, utilizando como corpus de

---

<sup>5</sup> Cuyo tamaño se expondrá más adelante.

<sup>6</sup> Más específicamente, siguiendo Pennington y coautores (2014), la función de ponderación es  $f(x) = \left(\frac{x}{100}\right)^\alpha$  (a la *alpha*), si  $x < 100$ , caso contrario  $f(x) = 1$



entrenamiento los artículos de la sección de economía del diario La Nación desde 1996-2004.

### 2.2.2 ÍNDICES CUANTITATIVOS.

Como mencionamos anteriormente, en este trabajo se evalúa la capacidad explicativa de indicadores de prensa basado en información no estructurada siguiendo diversas técnicas del procesamiento del lenguaje natural. En particular, en este apartado se exponen un conjunto de índices construidos a partir de computar la frecuencia de un conjunto de palabras que pertenece a una lista de palabras o diccionarios predefinidos en el conjunto de artículos que abordan el tópico riesgo país.

En pos de construir estos indicadores del tono de las noticias en la prensa, se identifica el conjunto de palabras perteneciente a la lista de palabras en el subconjunto de textos muestrales asociados al tópico riesgo país seleccionados previamente. Mas formalmente, dada una palabra  $k$  e  $W$ , siendo  $W$  la lista de palabras y  $T_t$  el total de palabras del texto seleccionado  $C$ . El índice computado para el subconjunto de texto seleccionado, está dado por:

$$I_C^k = \frac{\sum_{w \in W} c_w}{\sum_{w \in T} c_w}$$

Donde  $c_w$  indica el número de veces que la palabra  $w$  aparece en el conjunto de texto seleccionado  $C$ . El conjunto de texto selección en este ejercicio está dado por el contenido de la prensa económica para una ventana de tiempo específica.

Este trabajo se propone analizar la capacidad de explicar la evolución del EMBI que poseen diferentes índices de tono construidos siguiendo las metodologías propuesta por diferentes trabajos de la literatura especializada. Siguiendo la contribución de Tetlock (2007), la lista de palabras utilizada pertenece a la categoría “negativa” del diccionario Harvard IV.<sup>7</sup> Por otro lado, se construye una serie de indicadores a partir de la lista de palabras construidas mediante la implementación del modelo GloVe. Estos últimos indicadores se computan a partir del siguiente procedimiento: en primer lugar, se debe identificar una palabra que represente una característica de interés. En segundo lugar, se debe hallar el set de  $K$  palabras estrechamente asociadas al termino elegido en el primer paso, mediante la distancia coseno entre los respectivos vectores. Mas formalmente, dada una palabra clave  $k$  e  $W$ , el conjunto de  $K$  palabras estrechamente asociadas a  $k$  se

---

<sup>7</sup> La lista puede ser descargada del sitio web <http://www.wjh.harvard.edu/inquirer/homecat.htm>

identifica computando la distancia coseno:  $\frac{v_w v_j}{\|v_w\| \|v_j\|}$ .<sup>8</sup> Por último, el indicador es definido como la frecuencia de las palabras seleccionadas como se mencionó previamente.

### 3. DATA.

El spread soberano se mide a través del índice *Emerging Markets Bond Index*<sup>9</sup> (EMBI+) que computa y divulga JP Mogan. El modelo que se pretende estimar plantea que el EMBI computado para Argentina (*embiarg*) depende de variables macroeconómicas asociadas al frente doméstico y externo de la economía, así como variables económicas subjetivas, que explotan la información no estructurada divulgada en la prensa. Con este fin, se construyó una base de datos con frecuencia trimestral que cubre el periodo 1994:Q1-2019:Q1.

Con respecto a los índices subjetivos, se presentan una serie de indicadores computados a partir de datos no estructurados. El procedimiento realizado para computar los mismos se ha detallado en las subsecciones anteriores. Los indicadores computados son *tetlock* (siguiendo Tetlock, 2007) y *temor*. Este último índice corresponde a lista de palabras construidas a partir del modelo GloVe, utilizando las palabras más fuertemente asociadas a la palabra *temor*. En pos de capturar una medida más suavizada de este indicador, se promedió el índice de *temor* construidos para 50, 500 y 1000 palabras.

Adicionalmente, se utilizan un set de variables que intentan capturar la situación económica interna y externa del país, como así también sus condiciones de liquidez y solvencia, a saber: ratio entre reservas internacionales<sup>10</sup> y exportaciones<sup>11</sup> (*reservas\_x*), relación entre deuda externa<sup>12</sup> y PBI<sup>13</sup> (*deuda\_pbi*), tasa de interés real doméstica<sup>14</sup> (*ireal*), tasa de interés internacional<sup>15</sup> (*tasaint*), tipo de cambio real<sup>16</sup> (*tcr*) y EMBI+<sup>17</sup> (*embiemer*).

---

<sup>8</sup> La similitud entre dos vectores de términos se puede calcular mediante el ángulo que forman. En otras palabras, cuanto más próximo están los vectores (menor ángulo) mayor es la similitud entre ellos.

<sup>9</sup> Promedio trimestral de valores mensuales del EMBI+ computado para Argentina.

<sup>10</sup> Promedio trimestral de valores mensuales de los saldos a fin de mes (en millones de dólares).

<sup>11</sup> Exportaciones (en millones de dólares corrientes).

<sup>12</sup> Total de deuda externa del sector público no financiero y Banco Central a partir de una estimación de la deuda externa bruta por sector residente a valor de mercado. Promedio trimestral de valores mensuales de los saldos a fin de mes (en millones de dólares)

<sup>13</sup> Producto Bruto Interno (PBI) corriente interpolado (en dólares).

<sup>14</sup> Tasa de interés real mensual, depósitos plazo fijo 30-60 días (promedio geométrico de tasa reales mensuales: calculada como el ratio entre la tasa de interés mensual en el mes  $t$  y la variación del nivel de precios entre los mes  $t$  y  $t+1$ ).

<sup>15</sup> Tasa de interés de las letras del tesoro de Estados Unidos a 10 años (promedio trimestral de valores mensuales)

<sup>16</sup> Tipo de cambio real (promedio trimestral de valores mensuales)

<sup>17</sup> Promedio trimestral de valores mensuales del EMBI+.

La información proviene de múltiples fuentes de datos: Banco Central de la República Argentina (BCRA)<sup>18</sup>, Ámbito financiero<sup>19</sup>, Secretaria de Modernización<sup>20</sup> y Reserva Federal de Saint Louis<sup>21</sup>.

Una vez recolectada la información, se procedió a analizar el comportamiento estadístico de la serie. Se realizan tests de Dickey Fuller Aumentado para constatar la existencia o no de raíz unitaria para cada una de las series seleccionadas. Luego de realizar dichas pruebas, se procedió a obtener serie estacionarias para todos aquellos indicadores que exhibían raíz unitaria. Teniendo en cuenta esto último, se computa la diferencia en logaritmo del EMBI+ computado para Argentina (*dif\_embiarg*), EMBI+ (*dif\_embiemer*), tipo de cambio real (*dif\_tcr*), tasa de interés real (*dif\_ireal*) y tasa de interés internacional (*dif\_tasaint*).

En la tabla 1 se pueden observar las estadísticas descriptivas correspondientes a las variables utilizadas en este trabajo. En primer lugar, se puede visualizar que el índice EMBI computado para Argentina (*embiarg*), en promedio, es 1.7 veces mayor que aquel computado para los países emergentes (*embiemer*). A su vez, el máximo valor de la serie del EMBI para Argentina se observa en el tercer trimestre del 2002, el cual coincide con el periodo de crisis que atravesaba el país y el postdefault de la deuda soberana.

En segundo lugar, se pueden observar las variables que caracterizan la economía del país. Por un lado, el periodo se caracterizó por mantener una tasa de interés real negativa, mientras que la tasa de interés internacional (nominal) se ubicó en el orden del 4%. Por otro lado, el ratio entre deuda y PBI exhibe que el país durante el periodo bajo análisis, en promedio, cuenta con una relativa baja exposición, ya que su deuda externa representa 28% del PBI (el valor máximo de exposición externa se observó en el tercer cuatrimestre del 2002), mientras que la relación reservas y PBI demuestra un ratio de liquidez considerable, el cual, en promedio, se ubica en torno al 9%. Por último, el tipo de cambio real en promedio se ubica en 128 puntos, reflejando la competitividad vía tipo de cambio alto del país y, además, dicha variable se caracteriza por un elevado desvío estándar como se puede observar en la tabla 1. A su vez, el valor máximo de esta variable también coincide con el tercer trimestre del 2002.

---

<sup>18</sup> <http://www.bcra.gov.ar/>

<sup>19</sup> <https://www.ambito.com/>

<sup>20</sup> <https://www.argentina.gob.ar/modernizacion>

<sup>21</sup> <https://www.stlouisfed.org/>

En tercer lugar, observando las estadísticas referidas a los indicadores económicos subjetivos que se basan en información no estructurada, se puede observar que la información que comunican es similar independiente de la técnica que se utilizó para su computo. En particular, se observa que la media del indicador *tetlock* y *temor\_500* es del 4%, similar al valor de la mediana, denotando cierta simetría en su distribución. A su vez, el valor máximo del indicador de *temor* se registra en 2011:Q3 periodo que se caracteriza por las elecciones presidenciales, estancamiento de la economía y el agravamiento del conflicto con los *Holdouts*, mientras que el máximo del indicador *tetlock* se registra en 1998:Q1 el cual responde a las severas crisis que sufrieron ciertos países emergentes, no solo de la región.

TABLA 1: ESTADÍSTICAS DESCRIPTIVAS

Variable	Media	Mediana	Desvio Estándar	Q1	Q3	Mínimo	Máximo
embiarg	1396.47	719.62	1691.30	490.34	1091.04	208.38	6638.92
embiemer	520.18	398.03	292.39	313.45	702.88	158.22	1384.84
tasaint	4.12	4.17	1.62	2.73	5.22	1.56	7.84
tetlock	0.04	0.04	0.00	0.03	0.04	0.03	0.05
temor_500	0.04	0.04	0.01	0.04	0.05	0.03	0.06
reservas_pbi	0.09	0.09	0.03	0.07	0.11	0.04	0.15
deuda_pbi	0.28	0.24	0.19	0.15	0.28	0.09	0.89
i_real	-0.00	-0.00	0.01	-0.01	0.01	-0.05	0.04
tcr	128.81	106.21	47.65	95.02	148.38	83.22	269.64

Fuente: elaboración propia.

#### 4. ESTIMACIÓN Y RESULTADOS.

En esta sección, se evalúa la capacidad explicativa de los diferentes indicadores basados en el contenido de la prensa, descriptos en las secciones anterior, sobre la evolución del indicador de riesgo país, controlando por múltiples indicadores económicos que captura la situación económica del país. Como se mencionó anteriormente, se propone un modelo MCO simple. La ecuación que se estima, está dada por:

$$\Delta r_t = \alpha + \beta_1 I_t + \beta_2 X_t + \mu_t$$

donde la variable dependiente  $\Delta r$  es la diferencia en logaritmos del EMBI para Argentina (*dif\_embiarg*),  $\alpha$ ,  $\beta_1$  y  $\beta_2$  son parámetros a estimar por el modelo,  $\mu_t$  es el término de error. En este modelo, el indicador del riesgo país es una función lineal del valor del índice de prensa y un vector de indicadores económicos: ratio entre reservas

internacionales y exportaciones (*reservas\_x*), relación entre deuda externa y PBI (*deuda\_pbi*), diferencia en logaritmos de la tasa de interés real doméstica (*dif\_ireal*), diferencia en logaritmos de la tasa de interés internacional (*dif\_tasaint*), diferencia en logaritmos del tipo de cambio real (*dif\_tcr*). El parámetro de interés en este trabajo es  $\beta_1$ , el cual indica en qué medida el indicador de prensa puede ser utilizado para explicar la evolución del indicador de riesgo país. Bajo esta especificación, se espera que incrementos en estos indicadores impacten de manera positiva a sobre la prima de riesgo soberano. Para facilitar las comparaciones de los coeficientes estimados, todas las variables explicativas han sido estandarizadas.

La tabla 2 muestra los resultados obtenidos que surgen de la estimación del modelo que intenta caracterizar la relación entre el indicador de riesgo país e indicadores de la prensa, controlando por una serie de fundamentales macroeconómicos. Al observar los signos de los coeficientes estimados, se puede apreciar una asociación clara: incrementos en los niveles de los indicadores de prensa poseen un efecto positivo sobre la prima de riesgo soberano. Este resultado sugiere una asociación estadísticamente significativa y económicamente no despreciable.

En la columna 1 se exponen los resultados que surgen de la estimación del modelo expuesto anteriormente evaluando la capacidad explicativa del indicador de prensa computados a partir de lista de palabras, siguiendo las metodologías propuesta por Tetlock (2007). Como se puede observar, el indicador *tetlock* es estadísticamente significativo y tiene capacidad para explicar la evolución del indicador del riesgo país. En particular, se evidencia que un incremento en un desvío estándar dicho indicador, en promedio, está asociado a un aumento de 9% en el indicador de variación del EMBI.

Como se desprende de la columna 2, el parámetro estimado correspondiente al indicador de prensa computado a partir de representaciones vectoriales de palabras, es positivo y estadísticamente significativo. En particular, el coeficiente refleja capacidad para explicar la evolución del cambio en el EMBI se asocia a un índice que aproxima manifestaciones de temor. La estimación sugiere que un incremento en una desviación estándar en el

indicador de prensa asociado a temor se asocia a un incremento de 14% en la esperanza de la diferencia del EMBI, sugirieron que el efecto económico no es despreciable.<sup>22</sup>

Los resultados expuestos en la columna 1 y 2 permiten inferir que los indicadores económicos tradicionales habitualmente utilizados en la literatura explican la evolución del cambio en el indicador de riesgo país. En este sentido se puede observar que los signos de todas las variables poseen el signo esperado. En particular, el ratio reservas internacionales sobre exportaciones posee un signo negativo, denotando que incrementos asociados a dicho ratio se asocian con disminuciones en el indicador de riesgo país. Teniendo en cuenta que este ratio se asocia con la liquidez que posee un país para repagar o servir sus obligaciones, el signo es consistente con lo esperado. Sin embargo, este ratio parece no ser relevante para explicar la evolución del riesgo país. Por otro lado, otro indicador que describe la posición externa del país es el ratio deuda externa sobre PBI. Esta variable posee el signo esperado y, en particular, un incremento en una desviación estándar en la relación deuda externa - PBI se asocia a un incremento de 8% en la esperanza de la variación del EMBI, siendo estadísticamente significativo en la especificación del modelo exhibido en la columna 2 de la tabla 2, sugiriendo además que el efecto económico no es despreciable. Este indicador posee el signo esperando indicando que conforme se incrementa este ratio, se genera mayor presión sobre la liquidez disponible, lo que implica un incremento en el riesgo país. Este resultado se alinea con los resultados exhibidos en Edwards (1983), Mpapalika y Malikane (2019).

Continuando con los indicadores fundamentales de la economía, la tasa de devaluación del tipo de cambio real es estadísticamente significativa en las especificaciones expuestas en la columna 1 y 2. El signo de dicha variable es positivo, indicando que devaluaciones de tipo de cambio real se asocian con incrementos en el riesgo país, lo cual puede estar asociado al efecto negativo de dicha variable puede ejercer sobre el nivel de apalancamiento y endeudamiento externo del país. A su vez, el cambio en la tasa de interés real también posee el signo esperado, pero sólo es estadísticamente significativo bajo la especificación exhibida en la columna 2. En particular, un incremento en una desviación estándar en el cambio de la tasa de interés real se asocia a un incremento de 3% en la esperanza de la diferencia del EMBI. Este coeficiente parece indicar que

---

<sup>22</sup> Se regresó el modelo utilizando la misma especificación, pero en este caso utilizando índices que capturan manifestaciones de incertidumbre y pesimismo. En ambos casos, el signo de estos índices de información subjetiva no estructurada es positivo, pero no significativo estadísticamente.

incremento en la tasa de interés real doméstica atrae capitales internacionales especulativos o de corto plazo, los cuales pueden generar periodos de inestabilidad o turbulencia macroeconómica cuando se invierte dicho flujo.

Con respecto al cambio en la tasa de interés internacional, dicha variable no posee el signo esperado. Este resultado contrasta con los resultados exhibidos por González-Rozada y Levi Yeyati (2008). A partir del ejercicio exhibido en las columnas 1 y 2, se observa una relación negativa y estadísticamente significativa entre la tasa de interés internacional y el indicador del EMBI. Sin embargo, incrementos en la tasa de interés de las letras que pagan los bonos de EE.UU a 10 años se asocian con menor liquidez internacional y, por tanto, mayores indicadores de riesgo país para nuestra economía.

Teniendo en cuenta el fenómeno abordado en este trabajo y el periodo bajo análisis, Argentina se vio sumergida durante varios trimestres en default, es decir, en cesación de pago de sus compromisos internacionales, en los cuales el riesgo país, en promedio, fue de 5265 puntos básicos (2001:Q4 – 2005:Q2). Considerando esto último, en las columnas 3 y 4 se regresa el mismo modelo, pero excluyendo las observaciones atípicas asociadas al cambio en el EMBI<sup>23</sup>. Como puede observarse, las conclusiones exhibidas anteriormente se mantienen con excepción del cambio en la tasa de interés internacional la cual parece no ser relevante para explicar la evolución del cambio en el indicador de riesgo país.

Teniendo en cuenta que los agentes pueden utilizar múltiples fuentes de información no estructurada en pos de formular sus opiniones y expectativas económicas, se construyó una variable a partir de computar un análisis de componente principal entre el indicador de basado en la lista de palabras siguiendo Tetlock (2007) en base a los artículos asociados al tópico del riesgo país del diario La Nación, aquellos artículos asociados a toda la sección de economía del diario La Nación y los artículos de la sección de economía asociados a Argentina del diario Wall Street Journal, un diario con cobertura mundial. En este sentido, la columna 5 incorpora la variable *PCI* la cual representa la primera componente principal (la cual explica un 41% de la variabilidad de todos los índices computados a partir de las múltiples fuentes de información). El signo de esta componente es positivo, conforme a lo esperado, indicando que mayores niveles de pesimismo asociados a estas fuentes de

---

<sup>23</sup> Eliminando aquellas observaciones que se encuentran 1.5 veces por encima de la diferencia intercuartil.

información se asocian con mayores niveles de riesgo país. Los resultados no varían cuando se controla, además, por la segunda componente principal. Por otro lado, los resultados no varían significativamente cuando se incorporan las componentes principales resultantes de dicho análisis teniendo en cuenta de forma conjunta el índice Tetlock y temor para las múltiples fuentes de información.

En la columna 6, se exhibe una nueva especificación del modelo controlando esta vez por la variación del índice EMBI construido para todos los países emergentes (*dif\_embiemer*), con el objeto de capturar efecto contagios a los cuales puede estar sujeta la prima de riesgo que paga Argentina. El signo de dicha variable es positivo, el cual se alinea a lo esperado. En este sentido, un incremento en el EMBI para los países emergentes produce un incremento en el riesgo país argentino. A su vez, las conclusiones expuestas anteriormente referidas al conjunto de covariables utilizadas bajo esta especificación se mantienen, sin embargo, la magnitud de los coeficientes se invierte. Respecto de esto último, los coeficientes asociados a aquellas variables económicas fundamentales de la economía son mayores respecto de que aquel coeficiente asociado al indicador subjetivo construido a partir de información no estructurada.

Por último, continuando con el análisis de los estadísticos expuestos en la tabla 2 se arriba a similares conclusiones expuestas anteriormente. En este sentido, se puede inferir a partir del test-F de significatividad conjunta que estos indicadores tienen capacidad explicativa sobre cómo evoluciona el indicador de riesgo país. Por otro lado, observando el  $R^2$  – *Ajustado* se puede apreciar que dichos índices tienen una considerable capacidad explicativa sobre la variabilidad del cambio en el indicador de riesgo país, teniendo en cuenta las características del fenómeno que se desea abordar.

A partir de lo expuesto en esta sección, se puede concluir que ciertos indicadores de prensa basada en representación vectorial de palabras y en lista de palabras poseen capacidad para explicar la evolución del cambio en el EMBI, controlando por variables macroeconómicas habitualmente utilizadas en la literatura, que modelan la situación interna y externa de la economía. Adicionalmente, se arriba a similares conclusiones cuando se incorpora diferentes especificaciones del indicador basado en información no estructurada. Por otro lado, el modelo propuesto posee la capacidad de explicar una considerable fracción de la variabilidad del cambio en el EMBI.



TABLA 2: MODELO ESTIMADO

	(1)	(2)	(3)	(4)	(5)	(6)
tetlock	0.088*** (0.029)		0.064*** (0.024)			
temor		0.139*** (0.044)		0.082*** (0.022)		0.044** (0.018)
PC1_ao					0.040*** (0.015)	
dif_embierner						0.093*** (0.021)
reservas_x	-0.006 (0.025)	0.001 (0.016)	-0.020 (0.018)	-0.013 (0.012)	-0.002 (0.012)	-0.006 (0.015)
dif_ter	0.044* (0.024)	0.056*** (0.021)	0.040** (0.020)	0.050*** (0.016)	0.038*** (0.014)	0.054*** (0.015)
dif_jreal	0.018 (0.012)	0.027** (0.014)	0.021** (0.009)	0.027** (0.012)	0.022* (0.012)	0.005 (0.011)
dif_tasaint	-0.039** (0.017)	-0.042** (0.019)	-0.017 (0.013)	-0.021* (0.012)	-0.027* (0.015)	-0.006 (0.013)
deuda_pbi	0.014 (0.014)	0.085*** (0.027)	0.019* (0.011)	0.060*** (0.015)	0.057*** (0.017)	0.050*** (0.016)
Constante	-0.007 (0.032)	-0.007 (0.028)	-0.001 (0.019)	-0.003 (0.018)	-0.009 (0.021)	-0.003 (0.015)
Observaciones	92	92	88	88	81	88
R <sup>2</sup>	0.116	0.161	0.236	0.260	0.171	0.481
Adjusted R <sup>2</sup>	0.054	0.101	0.179	0.205	0.104	0.436
Estadístico F	1.866*	2.712**	4.167***	4.737***	2.549**	10.601***

Nota: entre paréntesis se encuentra el error estándar. Los estadísticos-t son computados utilizando errores estándar corregidos por la matriz de covarianza siguiendo Newey & West (1987, 1994).

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## 5. CONCLUSIONES.

Este trabajo se propone estudiar los determinantes de la prima de riesgo que debe pagar Argentina, con el objeto de analizar su asociación con factores económicos domésticos y factores externos que pueden influir sobre la evolución de dicha prima. Argentina a lo largo de su historia se caracterizó por transitar múltiples episodios de estrés económicos y financieros cuyas fuentes respondieron al marco institucional del país, la política económica aplicada, desviaciones de los parámetros fundamentales respecto de sus niveles de equilibrio, choques de liquidez externos asociado a cambios en la tasa de interés internacional, los cuales han tenido un impacto relevante sobre la sobretasa que pagan los bonos emitidos en el país. Estos episodios de estrés y su impacto sobre el riesgo

país pueden potencialmente comprometer la política de desarrollo del país, el sostenimiento de su deuda y estabilidad macroeconómica. Por otro lado, este trabajo se propone analizar la capacidad de ciertos indicadores construidos en base a información no estructurada comunicada en la prensa para explicar la evolución de ciertos indicadores económicos. En particular, teniendo en cuenta el incremento notable de múltiples fuentes relevantes de información disponibles para los agentes, se proponen indicadores novedosos que capturen en parte esta información y se evalúa su capacidad para explicar la evolución del riesgo país.

A partir de diversos ejercicios empíricos, se evidencia que ciertos indicadores construidos utilizando diferentes técnicas de procesamiento de lenguaje natural contribuyen a explicar la evolución contemporánea de la prima de riesgo que debe pagar Argentina por encima de la tasa de libre de riesgo que afrontan un bono de similares características emitido por un país desarrollado. Estos resultados se mantienen cuando se controla por diferentes indicadores macroeconómicos que capturan el estado del frente económico interno y externo del país. En particular, se encuentra que un incremento en una desviación estándar en el indicador que captura manifestaciones de temor en los artículos asociados al tópico riesgo país produce un incremento de 14% en la esperanza de la variación del EMBI. A su vez, estos resultados se mantienen cuando se incorpora un indicador que captura la mayor variabilidad de múltiples indicadores basados en diferentes fuentes de información económica comunicada en la prensa, cuyo efecto sobre la variación en el EMBI continúa siendo económica no despreciable (4%).

A la luz de los resultados obtenidos, la economía argentina puede reducir su exposición a la variabilidad de la prima de riesgo a partir de mejorar no solo sus condiciones de liquidez y solvencia, sino que también incorporar en el diseño de política económica indicadores que capturen diferentes manifestaciones de estados subjetivos en la prensa asociados a temor, pesimismo e incertidumbre.

Una posible extensión de este trabajo se asocia con incorporar otros países en el análisis y evaluar si las regularidades reportadas en este trabajo se mantienen. A su vez, otro posible ejercicio se vincula con evaluar si los efectos de este conjunto de controles poseen el mismo efecto sobre el riesgo país a lo largo de la distribución del EMBI.

## **REFERENCIA.**

- Aromí, J. D. (2018). "Linking words in Economic Discourse: Implications for Macroeconomic Forecasts", mimeo, IIEP-Baires UBA-CONICET.
- Azar, K., Oreiro, C., Tramontín, F. y Adler, G. (2007). "Determinantes del Riesgo Soberano en Uruguay", Banco Central del Uruguay, documento de trabajo N°1.
- Blei, D. M., Ng, A. Y. y Jordan, M. J. (2003). "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3, 993-1022.
- Blei, D. M. (2012). "Probabilistic Topic Models", *Communication of the ACM*, Vol. 55, No. 4
- Baker, S. R., Bloom, N., y Davis, S. J. (2016). "Measuring economic policy uncertainty", *The Quarterly Journal of Economics*, 131 (4), 1593-1636.
- Calvo-González, O., Eizmendi, A., Reyes, G. (2018). "Winners Never Quit, Quitters Never Grow: Using Text Mining to Measure Policy Volatility and Its Link with Long-Term Growth in Latin America", World Bank, Policy Research Working Paper, 8310.
- González-Rozada, M. y Levi-Yeyati, E. (2008). "Global Factors and Emerging Market Spreads", *The Economic Journal*, 118, 1917-1936.
- Garcia, D. (2013). "Sentiment during recessions". *The Journal of Finance*, 68(3), 1267-1300.
- Loughran, T., y McDonald, B. (2011). "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks", *The Journal of Finance*, 66 (1), 35-65.
- Mpapalika, J. y Malikane, C. (2019). "The Determinants of Sovereign Risk Premium in African Countries", *Journal of Risk and Financial Management*, 12, 29.
- Pennington, J., Socher, R., y Manning, C. D. (2014). "Glove: Global vectors for word representation". In *EMNLP* (Vol. 14, pp 1532-1543).
- Rowland, P. (2004). "The Colombian Sovereign Spread and its Determinants", Banco de la República de Colombia, Borradores de Economía 315.
- Tetlock, Paul, (2007). "Giving content to investor sentiment: The role of media in the stock market", *Journal of Finance*, 1139-68.
- Uribe, M. y Yue, V. Z. (2006). "Country spreads and emerging countries: Who drives whom?", *Journal of International Economics*, 69, 6 – 36.