Trabajo Final - Big Data

Predicción del riesgo país para Argentina a partir de noticias periodísticas

Gabriela Lorenzo, Ana Rocha y Eliana Uesu

13 de diciembre de 2018

Resumen

En este trabajo, se intenta desarrollar un modelo que permita predecir el riesgo país de Argentina, haciendo *text mining* de fuentes preseleccionadas de noticias internacionales. A partir de ellas, se busca construir un índice que refleje la reputación que tiene Argentina a nivel internacional, en términos de inversión. Se evaluará el poder predictivo de un modelo VAR que incluya a este índice como variable.

1. Motivación

El riego país es una de las variables más relevantes que los inversores consideran a la hora de decidir la colocación de su capital. Su importancia viene dada por su significado. Esta medida permite conocer la sobretasa de interés a la que accede cada país para financiarse en el mercado internacional. Esto le brinda a los inversores una noción de la probabilidad de que el Estado sea capaz de pagar sus obligaciones a su vencimiento.

Dada su relevancia, varios autores han intentado predecir/estimar la evolución de esta variable y hasta desarrollar modelos que permitan conocer su variación (Brown et al. (2014); Chan (1992); Wolff(2000); Erb et al. (2000)).

Es posible notar que a lo largo de los años esta variable estuvo influenciada por la reputación internacional del país. En otras palabras, este índice varía mucho de acuerdo a cómo es la imagen de Argentina a nivel internacional. Si bien no es fácil modelar el concepto que tiene el resto del mundo sobre nuestro país, se podría considerar que las noticias que circulan a nivel internacional pueden dar una idea respecto de la imagen internacional de Argentina.

Particularmente en este trabajo se analizará dos aristas de la visión internacional en torno a Argentina: ¿cuán optimistas o pesimistas son las noticias sobre el futuro de la economía? Mediante un proceso adecuado de *web scraping* y *text mining*, se buscará estimar el grado de optimismo y de pesimismo que presentan las noticias internacionales sobre la economía Argentina.

Al poder estimar esta variable a lo largo de los años, se propondrá un modelo VAR que utilice esta variable como predictor del riesgo país y se analizará la relación que existe entre ambas variables. Asimismo, se verá si este modelo propuesto es útil para predecir la variable de interés. ¿Será posible predecir el riesgo país de Argentina utilizando como variables predictivas el optimismo y el pesimismo presente en las noticias internacionales que hablan del país?

De esta forma, desarrollar un método que permita anticipar las variaciones del riesgo país sería útil para tomar, tanto decisiones de inversiones individuales como decisiones de deuda pública. Considerando la metodología propuesta, es de esperar que el índice creado sea simple de actualizar, dado que hoy en día las noticias se presentan constantemente. Esto brindaría una oportunidad de predecir el riesgo país de forma anticipada y a bajo costo.

Con el fin de hallar una respuesta a esta pregunta, en la Sección 2 se encuentra la revisión de la literatura sobre estos temas. La Sección 3 se presenta la información utilizada y la Sección 4 se expone la metodología propuesta. Finalmente, la Sección 5 se incluyen los resultados esperados y la Sección 6 las conclusiones.

2. Revisión de la literatura

A lo largo de los años, diversos autores han utilizado un sinfín de herramientas para predecir variables macroeconómicas y financieras de interés. Particularmente en la literatura reciente podemos hallar varios papers que utilizan herrameintas de *big data* y *machine learning* para predecir dichas variables.

Dentro de estos trabajos se destacan los que utilizan *scraping* como metodología. Por un lado, existen papers que utilizan este método para predecir variables macroeconómicas. Uno de ellos es Soo (2015), quien desarrolla una medida de sentimientos nacionales para 34 ciudades de Estados Unidos a través de la cuantificación del tono emocional de las noticias nacionales. Encuentra que el sentimiento de los medios locales tiene poder predictivo significativo para predecir la inflación.

De forma similar, Sharpe et al. (2017) cuantifican el optimismo y pesimismo en los

pronósticos de la Reserva Federal del mismo país para luego evaluar si esa medida tiene poder predictivo para variables macroeconómicas como la inflación, desempleo y PBI, entre otras. Luego se testea si la tonalidad de los pronósticos ayuda o no a predecir shocks de política monetaria.

Por otro lado existen papers que realizan *scraping* para predecir en el área de las finanzas. Entre estos, se destacan los siguientes: Loughran y McDonald (2011) desarrollan una lista de palabras de connotación negativa junto con otras cinco listas, que mejor reflejan la esencia de los textos financieros. Luego se utilizan dichas listas para predecir variables macro financieras como volatilidad de los retornos de activos, fraude y ganancias inesperadas, entre otras.

Aromí (2017) construye un índice a través de representaciones vectoriales de palabras para aproximar la percepción de la incertidumbre de Argentina con data de la prensa económica entre 1900 y 2017. Este índice permite predecir niveles futuros de volatilidad en mercados de activos y presenta una correlación con medidas alternativas de incertidumbre. Nyman et al. (2018) utilizan *scraping* de noticias para predecir el riesgo sistémico en los mercados financieros explotando el rol de las emociones sobre la evolución de variables financieras.

Otro conjunto de papers utilizan otras metodologías de *big data* para realizar predicciones de variables macro financieras. Entre ellos Tetlock (2011) busca testear si los inversores en mercados de activos distinguen correctamente entre información nueva y vieja sobre las firmas, donde una historia es considerada como nueva si su contenido es similar a las 10 historias pasadas de la firma. Se encuentra que los inversores individuales sobre reaccionan a la información nueva llevando a variaciones temporales en los precios de los activos. Choi y Varian (2012) utilizan *google trends* para predecir indicadores económicos a corto plazo, entre ellos: ventas de automóviles, desempleo, confianza del consumidor, etc.

Barsky y Sims (2012) utilizan inferencia indirecta para estudiar la respuesta a impulsos a innovaciones de confianza en un modelo neo keynesiano. Se obtiene que las noticias y los *animal spirits* contribuyen a la innovación de la confianza, y que el efecto de la confianza sobre la actividad se debe enteramente al componente de las noticias.

Finalmente Brown, Cavusgil y Lord (2014) crean un índice de riesgo país, el Robinson Country Risk Index (RCRI), basado en información provista por datos de Internet. El índice suma cuatro dimensiones a las habituales: Gobierno, Economía, Operaciones y Sociedad. Además, utiliza 70 sub-dimensiones, 126 países y 8 años de data.

Considerando la literatura previamente citada, este trabajo si bien propone el uso de una metodología ya utilizada (el *text mining* de noticias), su aplicación se realiza sobre una variable no estudiada en profundidad, como es el riesgo país, y para la Argentina, que es un país difícil de estudiar dada la volatilidad de los indicadores.

3. Data

3.1. Formación del EMBI+ original y su historia

Para los inversores, el *Emerging Market Bond Index* (EMBI+), construido por J&P Morgan, es el índice que mejor refleja el riesgo país. El EMBI+ de Argentina está construido a partir de una canasta de títulos de dueda soberanos de Argentina en USD. El peso de cada instrumento de deuda en el EMBI+ viene dado por el peso de su emisión como un porcentaje del total de instrumentos en el índice.

Este índice muestra la diferencia de rendimiento que existe entre la canasta de títulos correspondiente a Argentina y la de títulos del Tesoro de los Estados Unidos. El EMBI+ se calcula todos los días hábiles, de acuerdo con el calendario del mercado de bonos de Estados Unidos. La base (un número índice igual a 100) es el 31 de diciembre de 1993, cuando comenzó el cálculo EMBI+.

3.2. Recolección de datos

Con el fin de predecir cambios en el EMBI+ de Argentina se utilizan distintas fuentes que resultarán útiles. Primero para testear nuestra predicción se utiliza la base original de EMBI+ Argentina extraída del diario Ámbito Financiero[1] desde 11/12/1998 hasta 3/12/2018 con valores diarios (el índice opera con el calendario norteamericano, por lo que omite fines de semana y feriados nacionales de aquel país). La figura 1 muestra la serie, por lo que se puede observar que a lo largo de los años el índice fue cambiando considerablemente su valor, por lo que hace nuestro estudio más rico.

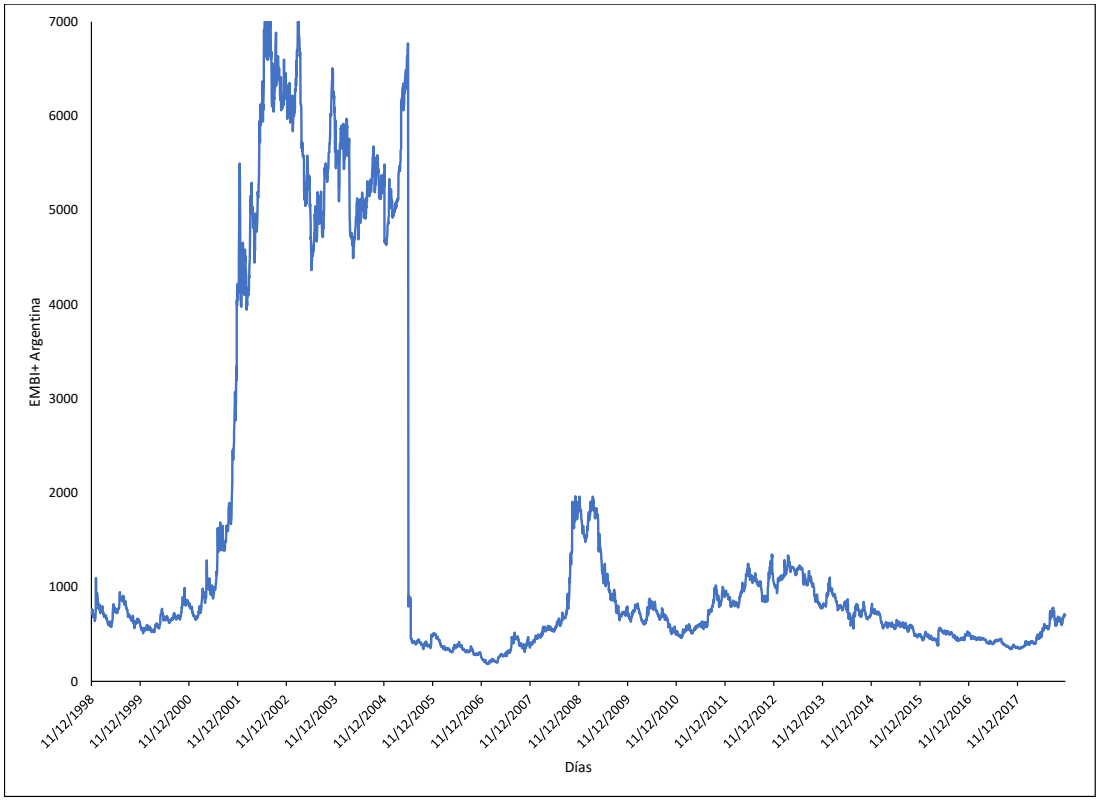

Figura 1: Dispersión de EMBI+ Argentina

El máximo cambio intra-diario se observa el 10/06/2005 (viernes) al 13/6/2005, día en que se incorporaron nuevos bonos a la canasta, producto del pago de la deuda argentina, pasó de 6606 a 794 puntos (una caída de 87,98 %). El segundo mayor cambio intradiario fue del 29/06/2005 al 30/06/2005, día en que se re-actualiza la ponderación de los bonos en pesos y dolares de la cartera. Por lo que se espera que nuestro predictor resulte útil para predecir a futuro, cambios de tal magnitud, teniendo en cuenta las primicias internacionales. Claramente, el curso de este índice está marcado por hechos relevantes en la economía argentina que posiblemente han circulado previamente en las noticias internacionales.

En otro orden, para realizar el análisis de *text mining* se utilizarán diversas fuentes de noticias financieras: *Bloomberg, Thompson Reuters, The Economist*. Una cuestión importante de aclarar es que las noticias están en inglés, dado que internacionalmente es el

[1] https://www.ambito.com/contenidos/riesgo-pais-historico.html

idioma más utilizado en el ámbito financiero.

Bloomberg[2]

Bloomberg es una compañía financiera que provee servicios de datos y noticias al instante. La compañía no se encarga de generar contenido sino de almacenar y distribuir. Actualmente posee un tercio de la proporción de mercado, por lo que lo hace un jugador influyente en la toma de decisiones de los inversores.

Thomson Reuters[3]

Thomson Reuters es una compañía de similares características que la anterior, sin embargo esta produce sus propios datos y noticias. En la actualidad posee 23 % del mercado, por lo que también es un actor influyente en la toma de decisiones de los inversores.

The Economist[4]

The Economist es el diario económico más famoso del mundo, por lo que se publican constantemente tanto noticias de la economía nacional como internacional, por lo que es una buena fuente de referencia.

La forma de seleccionar las palabras a utilizar es a criterio del autor. Aromí (2017) utiliza la frecuencia de palabras, luego en base a modelos de *Ridge* y componentes principales las selecciona. En este trabajo se seguirá la metodología de Nyman et al. (2018) por lo que las palabras (en el apéndice mostramos algunas de ellas) son clasificadas en optimismo y pesimismo de acuerdo a un conjunto de palabras previamente seleccionadas por Strauss (2013) en un contexto experimental.

Una vez realizado el *scraping* de noticias de las fuentes seleccionadas, se procesan todos los documentos, creando con las palabras *bags-of-words* en donde el orden de ellas y sus terminaciones (por ejemplo, los verbos se toman sin su conjugación) son ignorados, usando el algoritmo *Porter's stemming* (Porter, 1980). Luego, se computa una matriz de frecuencia de palabras, donde las palabras son filas y los documentos son columnas (cada entrada ij es la frecuencia de la palabra i en el documento j).

4. Metodología

Tomando como base la metodología utilizada en el trabajo de Nyman et al. (2018), se decidió plantear un esquema similar para ver si las noticias pueden predecir el riesgo país.

Con este fin, considerando los dos grupos de palabras que corresponden a las emociones de optimismo y pesimismo, tal y como se presentó en la sección previa, se propone construir un índice que mida el optimismo/pesimismo relativo en la narrativa de las noticias de las fuentes previamente presentadas. Este índice permitirá calcular el "sentimiento" que presenta la noticia de la siguiente forma:

$$sentimiento_{ij} = \frac{N_{ij}}{N_j}$$

donde para cada sentimiento ($i \in \{optimismo, pesimismo\}$) y para cada noticia j se construye este índice que se calcula a partir de la cantidad de palabras de tipo i presentes

[2] https://www.bloomberg.com/news/articles

[3] (http://financial.thomsonreuters.com/content/dam/openweb/documents/pdf/financial/ultra-low-latency-news.pdf)

[4] https://www.economist.com/

en la noticia j (N_{ij}) y el tamaño total del texto medido como la cantidad de palabras totales (N_j). A la hora de contar la cantidad de palabras correspondientes a cada sentimiento se decidió excluir aquellas que estén precedidas por expresiones de negación como: *"no"*, *"not"*, *"none"*, *"neither"*, *"nobody"*, tal y como se propone en Nyman et al. (2018).

Una vez obtenido el índice de *sentimiento* para cada noticia se construye una medida agregada, para optimismo y para pesimismo, que será el promedio de cada uno de estos. La medida agregada será:

$$opt_t = \frac{\sum_{j=1}^{J_t} sentimiento_{opt,j}}{new_t}$$

$$pes_t = \frac{\sum_{j=1}^{J_t} sentimiento_{pes,j}}{new_t}$$

También se decidió considerar como variable relevante para la estimación el caudal de noticias sobre la Argentina, es decir, la cantidad de noticias diarias presentes en la fuentes seleccionadas que hablan de Argentina. De esta forma tendríamos por día, el EMBI+ ($cntyrisk_t$), medidas agregadas de sentimiento (opt_t y pes_t) y la cantidad de noticias sobre el país ($new_t = J_t$).

Una vez obtenidos los datos necesarios, se va a estudiar la relación que existe entre estas variables a través de un modelo de vectores autoregresivos (VAR) de 4 variables ($cntyrisk_t$, opt_t, pes_t, new_t). Para esto, primero sería necesario que las variables sean estacionarias, lo cual puede ser evaluado con los test de Dickey-Fuller, el de Phillips-Perron y el de Kwiatkowski-Phillips-Schmidt-Shin. De ser necesario, se puede trabajar con las primeras diferencias de las variables para que éstas sean integradas de orden 0 (I(0)).

En segunda instancia es necesario considerar cuál es la longitud óptima de rezagos a incluir. Para esto se utilizarán los criterios información de Akaike (AIC), de Schwarz/Bayesiano (SBIC) y de Hannan and Quinn (HQIC). Esto permitirá conocer el orden de rezagos a incluir en el modelo VAR.

El modelo VAR quedaría definido de la siguiente forma, considerando un número p de rezagos significativos:

$$y_t = \delta + \phi_1\, y_{t-1} + \phi_2\, y_{t-2} + ... + \phi_p\, y_{t-p} + \epsilon_t$$

donde δ es un vector de constantes, ϕ_{t-h} una matriz de 4x4 que se corresponde con las relaciones del rezago $t-h$ sobre las variables contemporáneas, y el vector y_t es de la siguiente forma:

$$y_t = \begin{pmatrix} cntyrisk_t \\ opt_t \\ pes_t \\ new_t \end{pmatrix}$$

Posteriormente, será necesario realizar el test para autocorrelación multivariada de Portmanteau para ver que los residuos son ruido blanco. También sería útil aplicar un test de Causalidad de Granger, que no es una prueba de causalidad, sino que tan solo indica precedencia temporal de una variable respecto de otra.

Para evaluar la performance predictiva de este modelo se considerará una ventana *out of sample* correspondiente a un número de T períodos que será definido de acuerdo a las necesidades de predicción y a las disponibilidades de datos. La predicción se hará bajo un esquema recursivo que permita que el origen del pronóstico se actualice sucesivamente. Como *benchmark* de comparación se puede utilizar un modelo autoregresivo univariado a definir de acuerdo a las características de la información.

5. Resultados esperados

Utilizando la metodología propuesta se espera, por un lado que los coeficientes que relacionan las variables sean significativos para poder establecer que existe una correlación entre las variables creadas opt_t y pes_t, principalmente. Por lo tanto, sería interesante que mediante el test de Causalidad de Granger podamos establecer que las noticias anteceden temporalmente a las variaciones del índice y, de esta forma, poder concluir que esas variables son relevantes para predecir el riesgo país.

Finalmente se espera que el modelo VAR propuesto tenga una performance predictiva mejor que el modelo autorregresivo univariado utilizado como *benchmark*. Sería ideal que mediante este índice, que refleja la imagen internacional de Argentina, se pueda anticipar caídas y crecimientos relevantes en el riesgo país.

6. Conclusiones

Partiendo de los últimos avances en la aplicación de herramientas de *big data* y *machine learning* al campo de las finanzas como motivación, se propuso la predicción del riesgo país.

Se tomó como base la metodología utilizada en el trabajo de Nyman et al. (2018) y se planteó un esquema similar para estudiar si las noticias pueden predecir dicho indicador.

Con este fin, se propuso construir un índice para medir el optimismo/pesimismo de las noticias, considerando los dos grupos de palabras que corresponden a las emociones de optimismo y pesimismo.

Para realizar el análisis de *text mining* se propuso utilizar las siguientes fuentes de noticias financieras: *Bloomberg, Thomson Reuters, The Economist.*

Una vez obtenidos los datos necesarios, se propuso estudiar la relación existente entre ellos a través de un modelo de vectores autoregresivos (VAR) de 4 variables.

Posteriormente, a los fines de evaluar la capacidad predictiva del modelo se planteó la utilización de una ventana *out of sample* correspondiente a un número de T períodos definida de acuerdo a las necesidades de predicción y a las disponibilidades de datos. Finalmente se propuso utilizar un modelo autoregresivo univariado como *benchmark* de comparación.

Se espera obtener coeficientes significativos que relacionan las variables para poder establecer que existe una correlación entre las variables creadas opt_t y pes_t. Finalmente se espera que el modelo VAR propuesto tenga una performance predictiva mejor que el modelo autorregresivo univariado utilizado como *benchmark.*

Apéndice

Cuadro 1: Selección de palabras indicando optimismo y pesimismo.

Pesimismo	Pesimismo	Optimismo	Optimismo
Jitter	Terrors	Excited	Excels
Threatening	Worries	Incredible	Impressively
Distrusted	Panics	Ideal	Encouraging
Jeopardized	Eroding	Attract	Impress
Jitters	Terrifying	Tremendous	Favoured
Hurdles	Doubt	Satisfactorily	Enjoy
Fears	Traumatised	Brilliant	Pleasures
Feared	Panic	Meritorious	Positive
Traumatic	Imperils	Superbly	Unique
Fail	Mistrusts	Satisfied	Impressed
Erodes	Failings	Perfect	Enhances
Uneasy	Nervousness	Win	Delighted
Distressed	Conflicted	Amazes	Energise
Unease	Reject	Energizing	Spectacular
Disquieted	Doubting	Gush	Enjoyed
Perils	Fearing	Wonderful	Enthusiastic
Traumas	Dreads	Attracts	Inspiration
Alarm	Distrust	Enthusiastically	Galvanized
Distrusting	Disquiet	Exceptionally	Amaze
Doubtable	Questioned	Encouraged	Excelling

Nota: las palabras son puestas en inglés, ya que las noticias utilizadas también lo estarán. Asimismo el algorítmico de Porter también utiliza el mismo idioma. Fuente de las palabas: Strauss (2013)

Referencias

[1] Aromí, J. D. (2017). Measuring uncertainty through word vector representations. Económica.

[2] Barsky, R. B., & Sims, E. R. (2012). Information, animal spirits, and the meaning of innovations in consumer confidence. American Economic Review, 102(4), 1343-77.

[3] Brown, C. L., Cavusgil, S. T., & Lord, A. W. (2015). Country-risk measurement and analysis: A new conceptualization and managerial tool. International Business Review, 24(2), 246-265.

[4] Chan, K. C., Karolyi, G. A., & Stulz, R. M. (1992). Global financial markets and the risk premium on US equity (No. w4074). National Bureau of Economic Research.

[5] Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. Economic Record, 88, 2-9.

[6] Country Risk calculation for Brazil. Banco Central do Brasil (2016).

[7] Erb, C. B., Harvey, C. R., & Viskanta, T. E. (1996). Expected returns and volatility in 135 countries.

[8] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10ks. The Journal of Finance, 66(1), 35-65.

[9] M.F. Porter, (1980) .ᴬn algorithm for suffix stripping", Program, Vol. 14 Issue: 3, pp.130-137

[10] Nyman, R., Kapadia, S., Tuckett, D., Gregory, D., Ormerod, P., & Smith, R. (2018). News and narratives in financial systems: exploiting big data for systemic risk assessment.

[11] Sharpe, S., Sinha, N., & Hollrah, C. (2017). What's the Story? A New Perspective on the Value of Economic Forecasts.

[12] Soo, C. K. (2015). Quantifying animal spirits: news media and sentiment in the housing market.

[13] Strauss, Viktor, M. (2013). Emotional Values of Words in Finance: Anxiety about Losses and Excitement about Gains. M.Sc. thesis in Social Cognition, University College London.

[14] Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information?. The Review of Financial Studies, 24(5), 1481-1512.

[15] Tuckett, D., Nyman, R., Ormerod, P., & Smith, R. (2014). Big data and economic forecasting: a top-down approach using directed algorithmic text analysis. In ECB Workshop on Big Data for Forecasting and Statistics.

[16] Wolff, C. C. (2000). Measuring the forward foreign exchange risk premium: multi-country evidence from unobserved components models. Journal of International Financial Markets, Institutions and Money, 10(1), 1-8.