

1 Introducción

2 Objetivos y Metodología

Se propone una investigación enfocada en estudiar y evaluar la aplicación del “Procesamiento de lenguaje natural” o NLP por su sigla en inglés para construir un índice económico.

3 Desarrollo

Lo primero es definir y entender para que sirve un índice económico, desde ahora simplemente “índice”. Este se puede definir como un indicador estadístico asociado a un conjunto de instrumentos financieros, en este trabajo sólo se consideran bonos y acciones, matemáticamente hablando corresponde a una serie de tiempo $(X_t)_{t=1}^N$ típicamente indexada por día. También, se suele llamar índice al conjunto de instrumentos seleccionados para su construcción. El cálculo del índice se puede hacer de varias formas dependiendo de la naturaleza de los instrumentos que lo componen, generalmente se hace uso de un promedio ponderado de los precios de los instrumentos en el índice. Por ejemplo, suponiendo un conjunto I de instrumentos financieros, claramente finito, podemos computar el valor de nuestro índice al día t como:

$$X_t = \frac{1}{|I|} \sum_{i \in I} P(i)_t.$$

Donde $P(i)$ corresponde a la serie de precios del instrumento i . Este tipo de índices se usan como indicadores de que tan fructífera ha sido la economía, una subida en este valor se puede entender como un alza en los precios y por lo tanto, un “alza del sector económico” y viceversa. Uno de los índices más importantes es el **S&P 500** de Estados Unidos, país que al ser potencia influye en la economía del resto del mundo, de ahí su importancia. Este índice se compone de 500 acciones o empresas las cuales cotizan en la primera y segunda bolsa de valores más importantes de EE.UU, la de Nueva York o NYSE (New York Stock Exchange) y NASDAQ (National Association of Securities Dealers Automated Quotation), ubicada en la misma ciudad, correspondientemente. En el caso de Chile, existe el **IPSA** o Índice de Precio Selectivo de Acciones, este agrupa las 30 acciones con mayor presencia bursátil en el mercado y es reajustado cada año haciendo entrar y salir acciones.

3.1 Replicación IPSA

La primera parte del trabajo realizado se basó en estimar el valor del **IPSA** con los datos disponibles. La información entregada por **LVA** corresponde a los datos de transacciones diarias en la Bolsa de Santiago, agrupados por mes en distintos archivos. Estos datos se ven como se muestra en la Figura 1.

	Fecha	V	C	Cantidad	Nemo	R	Precio	L	MontoPesos	D	Folio	Hora
158255	2020-01-16	35	66	25000	SECURITY	T1	195.00	NaN	4875000	D	220598	14:54:00
187190	2020-01-14	35	35	1492	ENELAM	T1	167.94	OE	250566	D	215503	12:01:39
272243	2020-01-06	66	35	1685	CENCOSUD	T1	1015.10	NaN	1710444	D	214294	12:32:25
47906	2020-01-29	59	85	300	FALABELLA	T1	3170.00	NaN	951000	D	211146	10:47:54
40125	2020-01-29	88	88	320	CCU	T1	7099.90	OE	2271968	D	219069	12:52:06

Figure 1: Muestra de tamaño 5 de datos disponibles para estimación del IPSA, transacciones en la bolsa durante enero 2020.

Remark 3.1. *La mayoría de los datos financieros no incluyen fechas que correspondan a fines de semana o festivos. Esto se debe a que las instituciones que generan o recolectan estos datos no trabajan en tales días.*

En la Figura 1, la columna *Folio* identifica unívocamente a cada transacción, *Precio* indica el precio al cual se transó, *Cantidad* la cantidad de acciones transadas, *MontoPesos* corresponde a la multiplicación de las dos anteriores y *Nemo* es la etiqueta que se le da a la acción en la bolsa, generalmente corresponde al nombre de la empresa que emite la acción. Algunos de estos Nemos son: BSANTANDER, FALABELLA y ENELCHILE, por ejemplo. El resto de las columnas no será relevante para este estudio.

Posteriormente, el valor del IPSA, $(X_t)_t$, lo aproximamos con la siguiente fórmula,

$$X_t = base * \sum_{i \in I} w_i P(i)_t \quad (1)$$

$$w_i = \frac{C_i}{\sum_{i \in I} C_i p(i)_0} \quad (2)$$

donde I es el conjunto de instrumentos que componen el IPSA el año 2020 ¹, t hace referencia a días, $P(i)$ es la serie de precios diarios del instrumento i , $base$ corresponde al valor que se le quiera dar a la serie en tiempo inicial y C_i es el número de acciones suscritas pagadas del instrumento i . Este último término se define como la parte de las acciones disponibles para la venta, y por ende para la obtención de fondos por parte de la empresa o sociedad que las emite, que efectivamente han sido suscritas y pagadas. En otras palabras, nos habla del tamaño de cada emisor en el mercado bursátil chileno.

Por otra parte, notar que los “pesos” w_i no suman 1, sino que son tales que $X_0 = base$. La Fórmula 1 no es la forma exacta del IPSA, esta busca aproximar el valor real usando los datos entregados por LVA.

Remark 3.2. *El análisis no estuvo exento de daños en los datos. Las primeras aproximaciones al IPSA se caracterizaban por peaks severamente pronunciados o fechas en las cuales los valores obtenidos simplemente no tenían sentido. Estos problemas se deben a que en el caso de no tener el dato, este era sustituido por -1000 y además, existe un rango de fechas, específicamente en el mes de marzo del 2020, en las cuales los datos no fueron recolectados correctamente (ver Figura 2). Esta situación fue confirmada por el supervisor y se decide simplemente eliminar estas filas problemáticas.*

	Fecha	V	C	Cantidad	Nemo	R	Precio	L	MontoPesos	D	Folio	Hora
425626	2020-03-13	35	35	85.782	ITAUCORP	T1	3055.000	OE	262.064	D	220805	11:36:18
348279	2020-03-17	35	86	9.988	BSANTANDER	T1	30610.000	NaN	305.733	D	225103	12:15:24
309202	2020-03-18	35	58	452.000	CHILE	T1	62690.000	NaN	28.336	D	218817	12:01:57
411736	2020-03-13	70	90	10.102	RIPLEY	T1	275000.000	NaN	-1000.000	D	234759	13:36:25
411639	2020-03-13	70	58	2.511	CMPC	T1	1.515	NaN	-1000.000	D	234856	13:37:40

Figure 2: Muestra de datos dañados. Podemos ver los -1000 y que, por ejemplo, en la fila del 18 de marzo del 2020 la columna *MontoPesos* claramente no corresponde a la multiplicación entre *Cantidad* y *Precio*.

Lo primero que se necesita es estimar el precio de una acción en un día cualquiera. Para esto nos fijamos en una acción (o Nemo) y fecha cualquiera, a modo de ejemplo tomamos **CCU** y el 15 de abril del 2020. Para este par existen 456 transacciones (ver Figura 3).

¹Lista de empresas componentes del IPSA 2020 y 2021.

	Folio	Precio	MontoPesos	Cantidad	Precio * Cantidad
229	224374	6369.4	1273880.0	200.0	1273880.0
55	231417	6397.0	767640.0	120.0	767640.0
211	225283	6399.0	236763.0	37.0	236763.0
107	228925	6368.4	312052.0	49.0	312051.6
302	221209	6445.0	1031200.0	160.0	1031200.0

Figure 3: Muestra de tamaño 5 de las transacciones asociadas a **CCU** el 15 de abril del 2020.

Se decide calcular el precio del día como un promedio de los valores *Precio* ponderando por los *MontoPesos*. Es decir, para un instrumento i ,

$$P(i)_t = \frac{1}{\sum_{k \in T(i,t)} MontoPesos_k} \sum_{k \in T(i,t)} Precio_k * MontoPesos_k$$

donde $T(i,t)$ son las transacciones del instrumento i en el día t . Lo anterior nos entrega una estimación del precio del día. Por otro lado, la cantidad de acciones suscritas pagadas la obtenemos de la página web de la *Bolsa de Santiago*². Finalmente, en la Figura 4 se expone el resultado de este procedimiento.

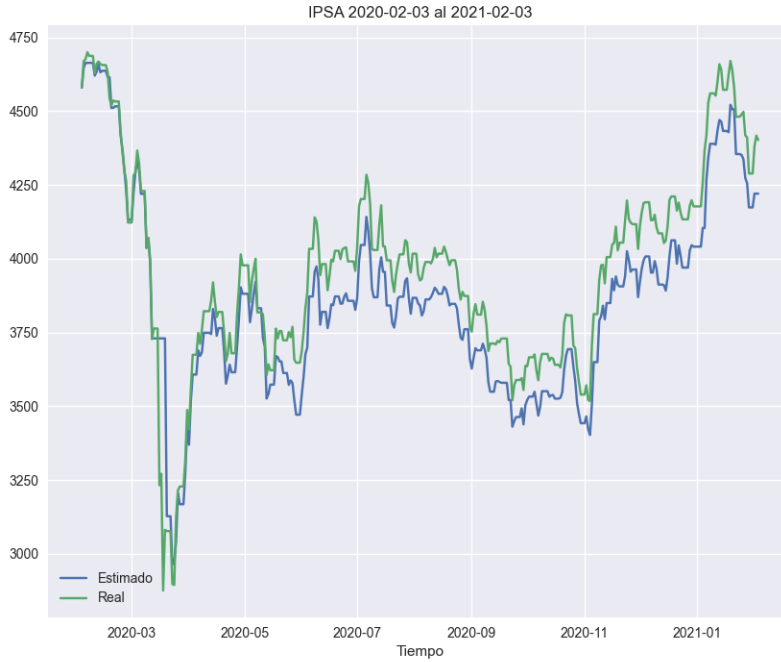


Figure 4: Replicación del IPSA.

La curva estimada captura bastante bien el comportamiento de la real, las subidas y bajadas del **IPSA** son logradas con buenos resultados. No obstante, es evidente el desfase que desde aproximadamente mediados de marzo comienza a hacerse notar y sigue la misma tendencia durante todo el periodo restante. Esta desviación se adjudica a la cantidad de datos dañados durante el periodo mencionado,

²Bolsa de Santiago.

ocasionando la recta horizontal en el punto donde el desfase comienza a hacerse más notorio. Por otro lado, es clara a la vista la caída que sufre el índice, tanto el real como el estimado, durante el mes de marzo. Esto refleja el esperado retroceso económico generado por la pandemia y no solamente en índices nacionales, la mayoría de los indicadores económicos mundiales sufrieron una caída similar.

3.2 Monto óptimo

La segunda tarea requerida puede entenderse de forma independiente al objetivo principal de la práctica. Se cuenta con dos fuentes de datos, la primera es relativa a fondos mutuos y cada fila de esta tabla se interpreta como inversiones que realizan los distintos fondos en ciertas acciones y bonos (ver Figura 5). La segunda corresponde a las transacciones que se realizan en la bolsa y da cuenta de cómo se mueve el mercado bursátil (ver Figura 1).

	NemoLVA	Run	ValorizadoCLP
13862	BSWTR-I	8287.0	1.000660e+08
39667	BSECZ31217	9727.0	1.090540e+08
35299	FNITA-030321	9106.0	9.995850e+08
22153	FNBNS-300321	8618.0	4.995846e+09
13374	BAMAI-F	8287.0	1.267460e+08

Figure 5: Muestra de tamaño 5 de los datos de fondos mutuos.

Se busca generar un gráfico donde x e y representan la cantidad de transacciones e inversiones respectivamente que cumplan cierto criterio parametrizado por un monto m . Para el eje Y , se agrupa la data de FFMM por run y luego por nemo para después sumar la columna “ValorizadoCLP” que tengan igual nemo dentro de cada grupo con el mismo run. Luego, dado un monto m , se cuenta la cantidad de inversiones resultantes (o instrumentos/nemos) tales que su monto sea menor a m obteniendo un número y . Por otro lado, para el eje X , se agrupa la data por día y luego por nemo para después contar las inversiones que tengan igual nemo dentro de cada fecha. Después, para cada día se calcula la cantidad de transacciones que superen el monto m para finalmente promediar este valor sobre un rango de días obteniendo la coordenada x .

3.3 Datos de Twitter

Necesitamos analizar texto que represente una crítica o reseña sobre una compañía. Estos textos vendrán en forma de *tweets* y queremos que hayan sido creados en cierto rango de fechas y que tengan ciertas palabras claves. Se investigaron variadas formas de hacer esto en `Python` pero se prefirió la librería `twint`³ dada su facilidad de uso y porque cuenta con la opción de manejar los tweets con DataFrames de `pandas`. Para obtener tweets con esta librería se debe especificar un intervalo de fechas y una request, esto último consiste en palabras claves o ciertas características que deban cumplir los tweets a buscar, ver Figura 6.

³Repositorio de `twint`.

	tweetcreatedts	text	language	hashtags	nreplies	nretweets
2111	2020-09-23 18:18:34	@banmedica todos los meses me llega un mail do...	es	[]	1	0
52	2021-02-03 09:50:25	@banmedica aún no me pagan el reembolso, en la...	es	[]	1	0
3729	2020-06-02 20:47:09	@banmedica Hola para la cobertura CAEC hay alg...	es	[]	1	0

Figure 6: Muestra de tamaño 3 de *tweets* obtenidos con *twint* con la request “@banmedica -from:banmedica”, esto significa que *twint* buscará tweets que etiqueten a Banmedica y que no vengan de la misma cuenta de Banmedica, más adelante se justificará esta request. La librería nos entrega un DataFrame con más columnas pero estas no fueron consideradas en este análisis.

Se procede entonces a obtener tweets relativos a cada nemo o empresa del IPSA durante el periodo del 28 de febrero del 2020 y el 30 de diciembre del mismo año, cabe señalar que esto tomó una cantidad de tiempo considerable obteniendo aproximadamente 1GB de datos en archivos *.csv*. Se decide que la request para cada nemo sea de la forma “[cuenta(s) de la empresa] -from [cuenta(s) de la empresa]”, esto quiere decir que *twint* buscará tweets que etiqueten a la empresa en cuestión, con la idea de buscar reseñas o opiniones sobre esta, pero no buscará tweets que vengan de la empresa, esto para evitar posibles sesgos que vengan de la compañía.

Remark 3.3. *Dado que algunas de las empresas asociadas a los nemos tenían una presencia relativamente baja en twitter, se decide por eliminarlas y trabajar solamente con las empresas cuyo promedio de tweets diarios sea mayor a 10. El haber o no eliminado estos nemos no tiene efecto en el resultado final, veremos que las empresas escogidas para entrar al índice son siempre del grupo que tiene mayor presencia en la red social.*

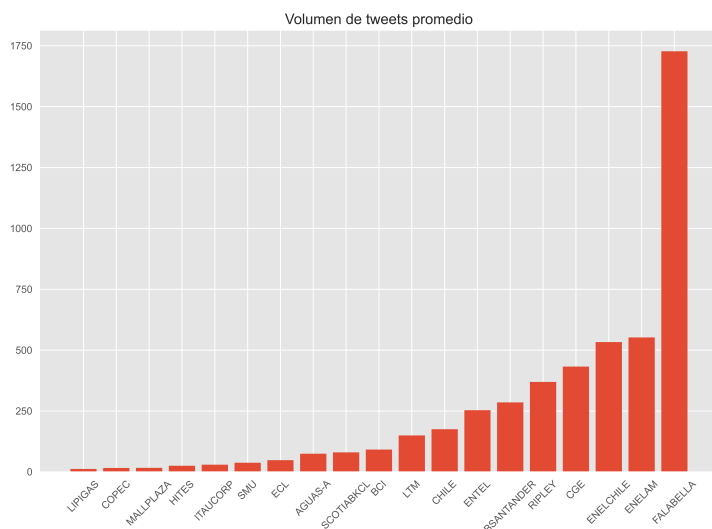


Figure 7: Promedio de tweets diarios relativos a las empresas con mayor presencia en la red social

La Figura 7 nos confirma que las empresas de retail, bancos, compañías de energía y agua son las más mencionadas en Twitter. Por el lado de las compañías de servicios básicos, cada vez que sucede un corte de agua o luz, las cuentas de estas empresas son mencionadas aumentando su volumen de tweets. Los bancos por otro lado, mantienen cuentas de tipo servicio al cliente y por lo tanto son

constantemente mencionadas, tanto para dar críticas como para realmente preguntar por sus servicios. Finalmente, el caso del retail lo tenemos representado en gran parte por Falabella, esta es una de las compañías más grande en Chile (confirmado en parte por la Figura 7) y además, debido a la pandemia, han aumentado la cantidad de compras online aumentando a su vez su presencia en redes sociales.

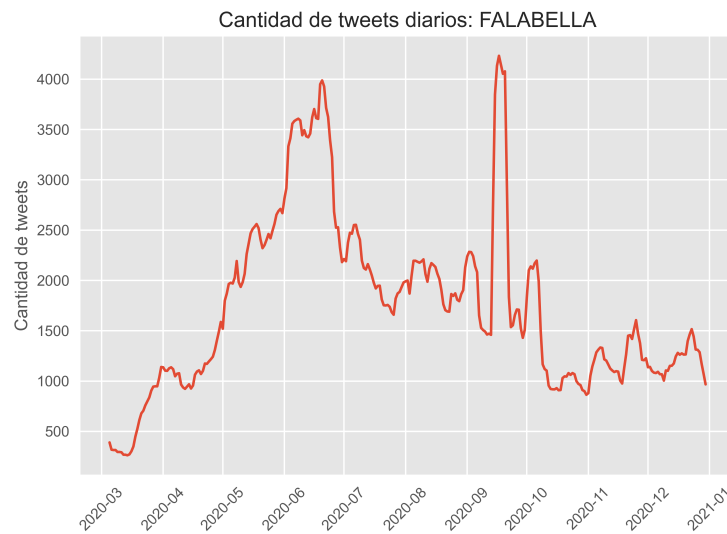


Figure 8: Cantidad de tweets diarios mencionando a Falabella. Se toma un promedio móvil con ventana de 7 días.

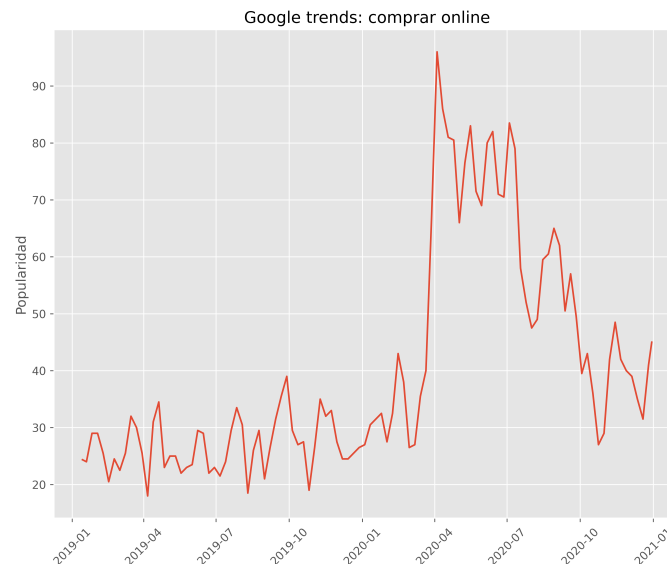


Figure 9: Popularidad del término de búsqueda “comprar online” en Google trends en Chile. Google define la popularidad como “el interés de búsqueda en relación con el valor máximo de un gráfico en una región y un periodo determinados”

En las Figuras 8 y 9 se observa el comportamiento sugerido anteriormente, la presencia de Falabella en Twitter comienza a crecer durante los meses de marzo y abril, meses que coinciden con el inicio de la pandemia y además, como es de esperarse, el término “comprar online” aumenta su popularidad durante mismo periodo.

3.4 Análisis de sentimiento

El análisis de sentimientos corresponde a una parte del **NLP** (Natural Language Processing), la tarea de este es asignar a un trozo de texto un número que indique el puntaje del texto en cada sentimiento, esto son: negativo, neutro y positivo. La primera dificultad es que la mayoría de las librerías utilizadas para esta tarea están hechas para analizar textos en inglés, al investigar la existencia de análogos en español, se llega a la librería **pysentimiento**, ver Figura 10.

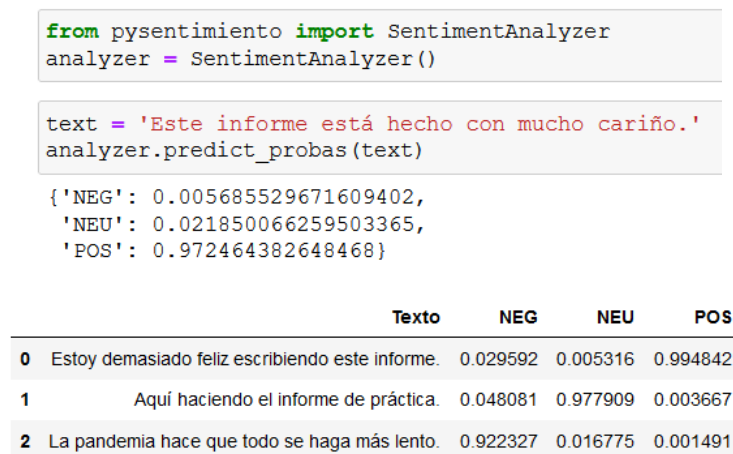


Figure 10: Ejemplos de uso de la librería **pysentimiento**.

3.5 Creación del índice

En esta subsección se mezclan las herramientas de las partes anteriores para implementar un programa que obtenga tweets asociados a cada nemo, analice sus sentimientos y maneje toda la información con DataFrames de **pandas**. A grandes rasgos, la lógica del código se hizo en base a una jerarquía de clases simple, esta consiste en una clase madre **Instrument** que representa un instrumento financiero genérico y clases **InstrumentRV** e **InstrumentRF** que heredan de la anterior y representan respectivamente a acciones y bonos. El tratamiento de bonos resulta un poco más complicado y será reportado más adelante en el informe.

Figure 11: Resumen diagrama UML del código.

El input del programa consiste en un diccionario, digamos **dict**, tal que **dict["RV"]** y **dict["RF"]** entreguen, correspondientemente, la lista de nemos de acciones y bonos considerados para trabajar en el índice. Idealmente este diccionario guarda todos los instrumentos del mercado chileno, pero dado que no todos estos tienen presencia destacable en redes sociales, se trabaja sólo con un subconjunto del total de instrumentos. Las acciones consideradas son un subconjunto de las acciones que conforman el IPSA y los bonos son los que tienen registro en las bases de datos de LVA y que sean

emitidos por compañías con presencia destacable en twitter, esto último hace que existan empresas relacionadas tanto a bonos como acciones y nemos de bonos asociados a la misma empresa. Para solucionar este enredo se crea un diccionario auxiliar que relaciona cada emisor de bonos a una lista de nemos que este emite, luego, se recorre el diccionario de emisores creando los objetos correspondiente como `InstrumentRF(nemo, emisor)`. El caso de las acciones es más simple dado que, por lo menos en este contexto, hay una relación uno a uno entre empresas y nemos de acciones. En esta parte del programa se cuenta con una lista de objetos tipo `Instrument` a los cuales se les “setean” sus datos con el método `setData()`. Dado que por cada día se tienen varios tweets, el valor de las columnas neg, neu y pos en el dataframe de la Figura 12 corresponde al promedio de estos valores en los tweets del día.

	num_tweets	tweetcreatedts	neg	neu	pos	Financial Variable	Return-1D
37	113	2020-04-05	0.605347	0.324006	0.070647	223.465541	0.000000
213	577	2020-09-28	0.564668	0.359067	0.076264	230.009518	-0.026040
244	145	2020-10-29	0.628995	0.291832	0.079174	199.924670	-0.025873
15	37	2020-03-14	0.541205	0.355911	0.102884	295.299402	0.000000
159	849	2020-08-05	0.605586	0.331345	0.063069	253.205868	-0.014425

Figure 12: Dataframe asociado al `InstrumentRV` que representa a COPEC.

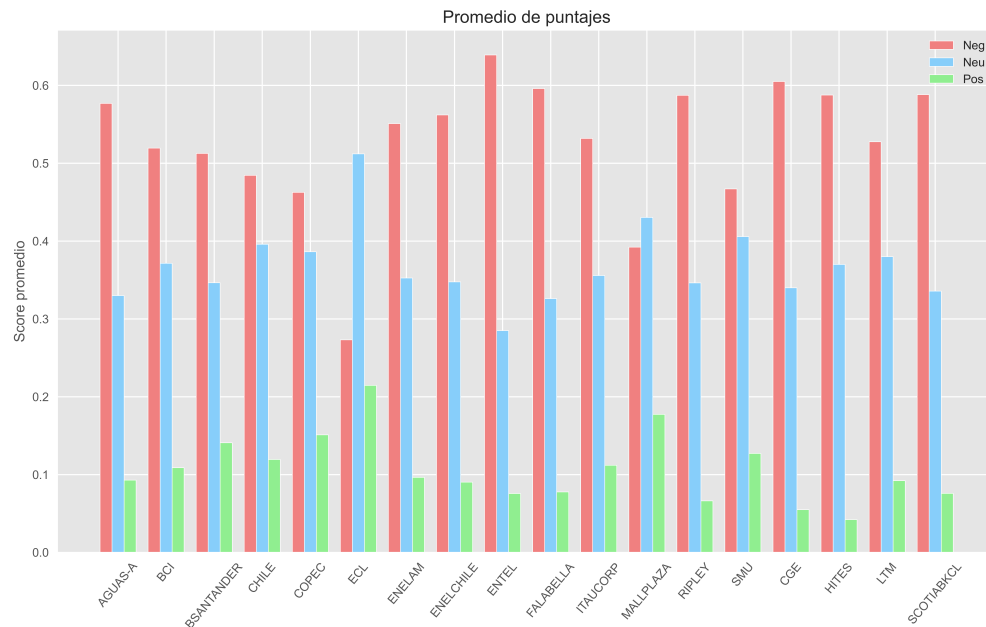


Figure 13: Promedio de los score en cada sentimiento para las acciones con mayor presencia en Twitter.

La Figura 13 nos ratifica la sospecha que se tenía desde un principio y es que la mayoría de los tweets tienen carácter negativo con un score mayor a 0.5.

4 Punto de que hablar

1. Calculo de índices: IPSA. Explicar el cálculo del índice y por qué se calcula este tipo de indicadores.
2. Análisis de sentimientos con twitter: Hablar de la cantidad de data por nemo y poner gráfico de barras.
3. Hablar de la decisión de dejar retweets porque representan el mismo comentarios pero de otra persona.
4. Comentar que hay tweets en otros idiomas y por lo tanto se hizo necesario trabajar con APIs de idioma.
5. Series de tiempo en economía.
6. Correlación entre dos variables.
7. Test de correlación.
8. Explicar el cálculo del sentimiento saturado, shifted y todo el preproceso que se le hace a los valores pos, neu y neg para transformarlos en un número representativo del sentimiento de cada tweet
9. Explicar que el sentimiento del día asociado a un instrumento se calcula como un promedio de los score de tweets diarios.
10. Explicar que se busca parámetros que maximicen la correlación entre las variables adecuadas: **retorno** del sentimiento, sentimiento promedio en el periodo **sent days**, promedio de los retornos diarios, etc contra el retorno del precio de la acción o menos la tasa de interés para el caso de bonos.
11. Mostrar los gráficos del índice obtenido y los nemos que participan del índice.