UNIVERSIDAD NACIONAL DE TUCUMAN
FACULTAD DE CIENCIAS ECONÓMICAS

UNIVERSIDAD NACIONAL DE TUCUMAN

FACULTAD DE CIENCIAS ECONÓMICAS

ESPECIALIZACIÓN EN FINANZAS

CONSTRUCCIÓN DE UN ÍNDICE DE SENTIMIENTOS CON TWITTER PARA EL MERCADO ARGENTINO

Carreño Giscafré Francisco Salvador

Curso 2019 - 2020

Resumen

Este trabajo de investigación surge, de la idea de entender mejor el fenómeno social detrás del mercado argentino, en un contexto actual donde las redes sociales toman un papel muy importante para la vida de la sociedad moderna y funcionan como depósito de los sentimientos de las personas. Se intentará encontrar si es que existe una relación entre las emociones plasmadas en comentarios de twitter, y las variaciones en las tendencias de los precios de los valores, creando un índice de sentimientos de mercado en base a la clasificación de los comentarios recopilados de enero a septiembre de 2020.

La utilización de redes sociales no ha hecho más que incrementar a lo largo de los años, por lo que los datos generados son cada vez más masivos y representan un gran valor para el análisis en todos los campos, por detrás de este contexto, es que florecen estudios que se valen en la utilización del “Big Data” y de la inteligencia artificial que ayudan a entender y resolver problemas de distinta índole.

Este trabajo intentará ser, además, un puntapié e inspiración para futuros estudios en materia de Big Data y Machine learning aplicados en el análisis del mercado argentino de valores, pudiendo ampliarse sus criterios, y mejorar los procedimientos aplicados en el análisis de la lingüística de las redes sociales para incluso intentar predecir las tendencias del mercado y así, entender de una manera mejor y más completa el fenómeno social detrás de las cotizaciones bursátiles.

Tabla de Contenidos

Capítulo 1

Motivación, Objetivos e Hipótesis

Motivación

Con el incremento en la utilización de las redes sociales, y particularmente de twitter para publicar y compartir opiniones, material y estudios, y tomando como inspiración trabajos de índole similar realizados en otros mercados, surgió la pregunta y la motivación de conocer si es que ¿Se puede realizar una medición del componente sentimental expresado en las redes y posteriormente trazar si es que existe una relación con el comportamiento del mercado argentino de acciones (MERVAL)?.

Otra importante motivación surge de la disyuntiva entre los distintos modelos teóricos que intentan describir los mercados desde puntos de vista totalmente opuestos como lo son, la hipótesis del mercado eficiente y la de la economía conductual, e intentar, de los resultados obtenidos, llegar a una conclusión en concordancia o no, con alguna postura.

La mayor motivación de este trabajo es crear un punto de inicio o puntapié para una rama profundamente estudiada, pero, poco explorada en Argentina como lo es el análisis de sentimientos de mercado, e inspirar a realizar estudios cada vez más completos y con mejores técnicas a futuro.

Objetivos

Objetivo General

Conocer si existe alguna relación entre los sentimientos manifestados en twitter con los movimientos en los precios de las cotizaciones bursátiles del mercado argentino representado por el índice MERVAL utilizando técnicas de procesamiento de lenguaje natural sobre el texto de los comentarios de tweets relacionados al tópico en la red social.

Objetivos Específicos

Para poder responder a la pregunta planteada, en este trabajo se intentó cumplir con los siguientes objetivos específicos:

1. Lograr extraer datos suficientes y aptos para un análisis conciso desde el periodo que va del 01/01/2020 al 30/09/2020.
2. Clasificar el contenido de los datos utilizando Inteligencia artificial.
3. Elaborar un índice de Market feelings en una línea temporal.
4. Estudiar si existe alguna relación con el movimiento del mercado en el periodo analizado.

Este trabajo además tiene por objetivo, documentar todo el proceso en detalle para futuros avances en trabajos de investigación relacionados para buscar mejorar la técnica y los resultados obtenidos.

Hipótesis Planteada

La hipótesis que se intentará evaluar en los siguientes capítulos, va de la mano de la teoría de las finanzas conductuales, y plantea que existe una relación entre los sentimientos que la gente expresa en redes sociales con los ciclos que experimentan los precios de los mercados. Esta relación implicaría que precios de los valores se mueven a través del tiempo en consonancia o con cierta similaridad con los sentimientos de las personas medidos mediante una clasificación de polaridad. De comprobarse esta hipótesis, podría funcionar como una evidencia o un aporte más, en defensa de la teoría de las finanzas o economía conductual, y en disonancia con la hipótesis del mercado eficiente, aunque no se busca ni refutar ni afirmar ninguna de estas importantísimas teorías que dan origen al entendimiento que tenemos hoy en día de los mercados. La interpretación de los resultados no puede ser tomada como absoluta, ya que el periodo bajo análisis es acotado, y solamente incluye información del mercado argentino.

Capítulo 2

Introducción

Para empezar, resulta necesario encuadrar y definir los campos de estudio que abarca el presente tema, el estudio de los “Market Feeling” pertenece a ramas de la sociología, psicología y la lingüística además de las finanzas y mercado. Este trabajo además está especialmente vinculado al área de las ciencias computacionales, sirviendo estas últimas como herramienta fundamental para el procesamiento de la información.

El tema principal sobre el cual gira alrededor el presente estudio y al cual va dirigido este aporte, es el entendimiento del funcionamiento de los mercados, su comportamiento, ciclos, tendencias, y evolución. Y como pueden o no, relacionarse con las cotizaciones, el factor social, la masa de personas, que tienen opinión y sentimientos. Se intentará contestar a la pregunta ya plantada en el capítulo anterior midiendo ese sentimiento expresado en la red.

Es importante lograr conclusiones y estudios diversos sobre esta temática ya que cualquier aporte puede ayudar a comprender mejor y profundizar los conocimientos que se tienen sobre los mercados. Encontrar relación entre el mercado con las interacciones sociales, es entender mejor la naturaleza del ser humano y como el comercio y la economía son ejes fundamentales de nuestra vida pasada, actual y futura.

En el presente capítulo se introducirán los conceptos clave y las teorías que marcan el campo dentro de cual se enmarca el estudio aquí realizado.

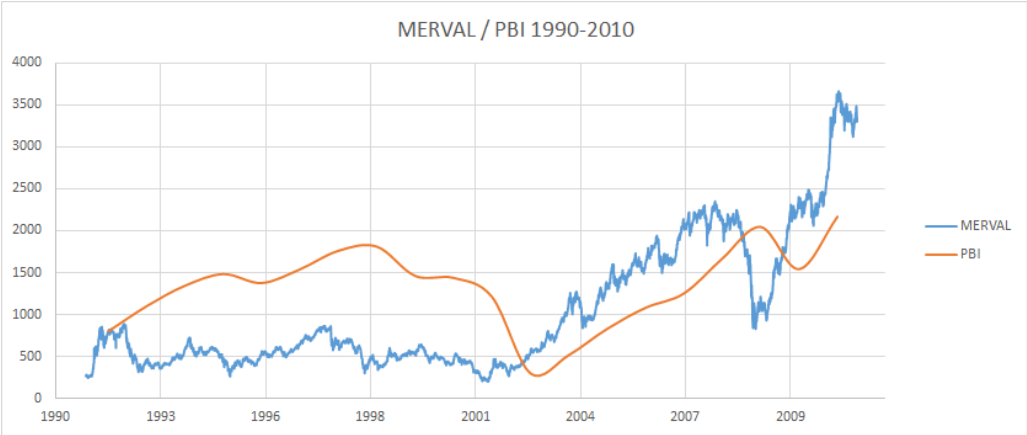
Los Ciclos Económicos

En general la economía suele moverse en ciclos, períodos de bonanza y crecimiento constante, y tiempos que se caracterizan por crisis, desempleo y contracción económica, pudiendo del estudio estadístico, inferir e identificar los ciclos. En los precios de los valores bursátiles, suele verse reflejado el impacto de estos periodos, afectando positiva o

negativamente en las cotizaciones que derivan de entre muchos factores, de las expectativas de la sociedad del momento, como se aprecia en la Figura 1.

Figura 1

PBI de Argentina versus cotizaciones al cierre del Índice Merval entre 1990 y 2010



Fuente: Elaboración Propia a base de datos del Banco Mundial y Yahoo Finance.

En la Figura 1 se denota claramente una similitud en los ciclos de la economía del país y los del mercado argentino, en donde las variaciones del Merval se adelantan ligeramente a las del PBI, esto explicado quizá en que el mercado toma como variable las expectativas de lo que ocurrirá. Se observa un periodo en forma de meseta desde 1990 a 2001, luego una tendencia al alza hasta 2008-09, ambas figuras terminadas en crisis financieras históricas, terminando otra vez en alza.

Los Ciclos del Mercado

Los ciclos del mercado bursátil son objeto de estudio, desde hace más de 100 años, empezando con la tan conocida “Teoría de Dow” (Charles Dow, 1902), un incursor en lo que se refiere al análisis técnico de gráficos bursátiles, en la que describe las tendencias de los precios del mercado y detalla la existencia de varios niveles para las mismas. Charles Dow en sus notas, define 3 horizontes de tiempo, el largo, mediano y corto plazo e identifica las tendencias que siguen los precios, quedando en evidencia la existencia de ciclos de diferentes horizontes temporales en las tendencias.

Para identificar una tendencia en una gráfica basta con observar un despliegue gráfico de precios históricos como la Figura 2, e identificar la dirección que “en general” tienden a moverse los precios, identificando ciclos alcistas con precios máximos y mínimos cada vez mayores y, ciclos bajistas donde el precio toca seguidamente mínimos y máximos menores.

Figura 2

Gráfico de velas del Índice Merval diario en pesos de 2013 a 2020

Fuente: Adaptado de Tradingview.com, https://es.tradingview.com/chart/X9YRbapu/

En la Figura 2, se muestra la tendencia alcista uniendo los mínimos y obteniendo una línea recta con pendiente positiva, además en azul la media móvil simple (otro indicador utilizado por Dow) de 200 sesiones también ascendente, señalando igualmente la tendencia creciente de los precios.

Existen otros métodos que intentan reconocer y predecir los ciclos del mercado de la mano del análisis técnico como por ejemplo las ondas Eliot, osciladores como el MACD y el RSI descriptos en la obra “Análisis técnico de los mercados financieros” (John J. Murphy,

1999), entre muchos otros. La identificación certera de los ciclos permite reducir el riesgo en la toma de decisiones financieras, además de mejorar la comprensión del fenómeno del mercado.

Explicando los Movimientos del mercado

El ser humano, desde sus inicios en las actividades económicas ha tratado con la incertidumbre y la dificultad para prever los hechos futuros, acudiendo a métodos esotéricos de toda clase para tratar de traer tranquilidad y certeza en sus decisiones.

Se puede decir lo mismo con el estudio del mercado, porque incluso a día de hoy, seguimos conviviendo con la incertidumbre. Es parte de la naturaleza y de nuestra vida como humanos, tener que tomar decisiones sin tener certezas de lo que sucederá, habilidad que permitió nuestra supervivencia y supremacía en el planeta a lo largo de nuestra historia.

En la carrera contra la incerteza, como seres racionales y usuarios de la ciencia y la filosofía, buscamos explicar y descifrar el funcionamiento y el porqué de las cosas, lo que permite surgir a los avances del conocimiento que suceden día a día a un nivel cada vez mayor. Para el caso de las cotizaciones bursátiles, aún no contamos con un modelo que explique a la perfección y pueda predecir con exactitud su comportamiento, pero existen distintas hipótesis y teorías que intentan explicar los mecanismos que trabajan por detrás de los mercados.

Teoría del Paseo Aleatorio

Plasmada en la obra “Un paseo aleatorio por wall street” de Burton G. Malkiel (1973) basada en la hipótesis del mercado eficiente (Eugene Fama**,**1970) bajo este paradigma, los precios del mercado son el resultado de la interacción de partícipes desiguales y por medio del arbitraje, todas estas inequidades se disuelven.

En el intercambio de activos financiero intervienen expertos e inexpertos, inversores racionales e irracionales, personas con información privilegiada y totalmente desinformada, el mercado según su nivel de eficiencia, es más o menos capaz de asimilar esto en sus precios

más rápidamente, “Se dice que un mercado de capitales es eficiente si refleja completa y correctamente toda la información relevante para determinar los precios de los valores” (Efficient market hypothesis and forecasting, Allan Timmermann, Clive W.J. Granger, 2004). Las desigualdades se compensan de manera tal que no afectan al precio, este último refleja en sus cotizaciones su valor justo a la fecha, pudiendo llegar a la conclusión de que como el activo en cuestión ya vale lo que debería valer, sus precios no siguen ninguna tendencia ni se dirigen a algún objetivo predecible, sino que ya está correctamente valuado, por consiguiente se puede llegar en consecuencia a la hipótesis más controversial de esta teoría: “Un chimpancé con los ojos vendados tirando dardos sobre las páginas de cotizaciones bursátiles de un periódico, podría seleccionar una cartera de valores tan buena como la seleccionada con el mayor cuidado por los expertos.” (Burton G. Malkiel, 1973, p.20)

El autor no pretende decir que el análisis no pueda tener buenos resultados, pero indica que solo comprando y manteniendo se pueden tener resultados igual o incluso superiores y que, de cumplirse la forma débil de la hipótesis del paseo aleatorio, y replicando las palabras de su colega Richard Quandt: "el análisis técnico se asemeja a la astrología y tiene tanto como ella de científica". (Burton G. Malkiel, 1973

Otra conclusión a la que se llegaría si se cumple la hipótesis de esta teoría es que los precios anteriores no tienen ninguna incidencia en los actuales, otra razón por la que es detractora del análisis técnico, análisis que se sustenta en datos anteriores para poder inferir la tendencia de los precios.

Esta teoría ampliamente difundida creó una escuela de estudio con su propio punto de vista de las finanzas conductuales, entendiendo que las masas no pueden adoptar una tendencia ni tomar decisiones irracionales.

Puntos Débiles de la Teoría del Paseo Aleatorio

La recomendación del autor de la teoría es invertir en fondos cotizados o ETFs que repliquen el rendimiento del mercado y mientras más abarcativo mejor como los “Total

Markets”, y que hacer esto es mejor que una administración activa. En la práctica, el rendimiento de los índices se utiliza como benchmark o punto de comparación para determinar el rendimiento de expertos inversores y administradores de carteras de inversión, y en defensa del análisis técnico y fundamental, existen muchos inversores y administradores que tuvieron un rendimiento sostenido superior al del mercado en general, hecho que el autor reconoce como excepciones.

Otro punto que el modelo del paseo aleatorio no puede explicar son eventos imprevistos e irracionales en las cotizaciones, como los “días negros” en los que reina el miedo irracional, sin fundamentos claros, o la formación de burbujas en las que los precios suben sin tener un techo aparente, todos hechos que carecen de racionalidad y que siguen el sentimiento de miedo o euforia de las masas.

Finanzas Conductuales

Perteneciente al campo de estudio de la economía conductual, estudia cómo los sentimientos y tendencias cognitivas afectan a las decisiones financieras. La racionalidad, algo de lo que ya se habló anteriormente, es objeto de estudio de este campo en cuanto a su ausencia o presencia en la toma de decisiones.

Las finanzas conductuales son cada vez más estudiadas gracias a las redes sociales, y la masividad de la información en la actualidad, en la que las redes juegan un papel principal en la vida social moderna, drásticamente diferente a la de hace solo unos 10 o 20 años atrás.

Las observaciones realizadas en estudios que determinan su análisis teniendo en cuenta variables conductuales o sociales, a menudo llegan a conclusiones detractoras de la hipótesis del mercado eficiente, debido a que las decisiones irracionales o derivadas de tendencias sociales son la regla para este tipo de investigaciones.

La novedosa perspectiva de las finanzas conductuales ve al mercado como un fenómeno social en el que el estudio de las masas resulta clave, cualquier persona puede entrar en una red social desde su móvil y leer gratuita y cómodamente el análisis que hizo un

experto sobre un activo financiero, cosa que hace 10 o 15 años no tenía ni por asomo esta facilidad y accesibilidad.

Es indiscutible que todo mercado se rige por la Oferta y Demanda, ya que es el lugar natural en que se encuentran y se realizan transacciones, sin embargo, éstas al pertenecer a un conjunto de personas que pujan según sus objetivos, pueden actuar de una manera no esperada racionalmente, por ejemplo en un modelo económico básico de oferta y demanda, ante caídas en la demanda, lo esperado sería que los precios también caigan hasta el punto de equilibrio, pero puede pasar que por la aversión a las pérdidas, los tenedores, dueños de la oferta no liquidaron su tenencia a precios bajos y simplemente esperaron a que la demanda se recupere, por lo que el precio se movió contrariamente al razonamiento del modelo planteado.

Según un estudio académico titulado “Las Finanzas Conductuales, el Alfabetismo Financiero y su Impacto en la Toma de Decisiones Financieras, el Bienestar Económico y la Felicidad”, (Gonzalo Garay Anaya, 2015), el contexto social-economico-educativo pueden hacer a las personas tomar decisiones irracionales, y concluye en su análisis que existe una clara relación entre las emociones y su influencia positiva en el bienestar económico-financiero y la felicidad.

Al estudiar las finanzas conductuales, es común descubrir que existen sesgos de todo tipo en la mente del inversor que derivan de la conducta social, un sesgo al cual las redes sociales facilitan su existencia, es el conocido sesgo de confirmación, “El sesgo de confirmación es el fenómeno de respaldar nuestras propias opiniones con información selectiva. Los inversionistas buscan una confirmación de sus suposiciones. Evitan las opiniones e informes críticos, y leen únicamente aquellos artículos que ponen el producto bajo una luz positiva” (Thorsten Hens y Anna Meier, 2016, “Finanzas conductuales: La psicología de la inversión”), de este modo al encontrar en las redes comentarios e informes de todo tipo, se podría generar un sesgo de confirmación que concluya creando tendencia.

De esta forma observando el factor social y entendiendo al mercado, la oferta y demanda como herramientas sociales que nos acompañan desde el principio de nuestra historia como raza humana, se pude vislumbrar una nueva forma de entendimiento de las tendencias del mercado desde los sentimientos de las masas o market feelings.

Market Feelings

El análisis de “Market Feelings” (Sentimiento de mercado en español) es un estudio en el que se busca conocer los ciclos del mercado, usualmente detectando si detrás de las cotizaciones existen sentimientos bajistas o alcistas, por lo que resulta en otra forma de entender los movimientos del mercado.

Para conocer los sentimientos del mercado, es necesario partir de la idea de que estos se mueven en tendencias, y que la conducta y pensamientos de las personas son las que definen su dirección.

Existen muchas formas de conocer el sentimiento del mercado, entre ellas, las que buscan una fuente de información diferente a la de los precios, como la relacionada a opiniones, pensamientos y análisis de las personas y otras, que toman información directamente del mercado.

En su trabajo “Ciclos bursátiles e indicadores de sentimiento del mercado” Sergio Luis Olivo (2016) presenta distintas formas de realizar un análisis de sentimientos del mercado, entre las que encontramos métodos clásicos basados en encuestas a inversores o en datos de mercado, que vienen siendo probadas hace décadas y que fueron el objeto específico en ese trabajo, pero además nombra entre los métodos más novedosos, al del análisis de texto en redes sociales e información de internet. Olivo aclara en su textualización que el mercado puede tener sentimientos al igual que las personas, ya que según sus palabras “Los mercados son personas, detrás de los gráficos bursátiles hay personas que compran y que venden, que ganan y que pierden”. (Sergio Luis Olivo, 2016).

La visión de los mercados como un fenómeno humano y por lo tanto social forma parte de la teoría de las finanzas conductuales.

Antecedentes

Ciclos Bursátiles e Indicadores de Sentimiento del Mercado (Market feeling) (Sergio Luis Olivo 2016).

En este trabajo, el autor describe distintas formas de medir los sentimientos de mercado, y como estos se relacionan con sus ciclos. Detalla en profundidad el funcionamiento y resultados de distintas técnicas clásicas para medir los market feelings utilizando datos extraídos de la información que provee el mercado como precios, volumen, opciones etc., y utilizando información encuestas a inversores.

Análisis de Sentimiento en Twitter: El bueno, el Malo y el >:((Carlos Martín Becerra, 2016)

Este trabajo presenta un proceso de análisis de sentimientos en la red social twitter, estudiando como acontecimientos que produjeron tendencias afectaron la opinión de los usuarios. El análisis se realiza filtrando únicamente tweets que están relacionados a tópicos específicos y no de manera general, concluyendo en que hay palabras que están relacionadas a comentarios positivos y otras a comentarios negativos.

El Uso de Twitter en el Análisis Financiero: Aproximación Desde la Econofísica (Andrés García Medina, 2017)

Desde el análisis textual de tweets de un periodo de 7 meses en 2014, se construyeron en este trabajo, series de tiempo de polaridad mediante el análisis de sentimientos de twitter y New York Times, y se las relacionó con el rendimiento de 20 índices financieros de todo el mundo, incluido el MERVAL argentino, concluyendo y revelando que existen correlaciones positivas entre los índices financieros y la polaridad calculada.

Predicción de los Rendimientos de Acciones en Argentina en Base a Indicadores Técnicos y al Modelado de Tópicos en Foros Bursátiles (Ramiro h. Gálvez, 2016)

Mediante técnicas de procesamiento de lenguaje natural de información proveniente de un foro bursátil de Argentina, se intentó conocer, del análisis de los comentarios diarios, si se puede inferir el movimiento en las cotizaciones bursátiles a futuro, obteniendo resultados positivos. Se encaró el estudio utilizando modelos de aprendizaje automático basados en el uso de indicadores técnicos junto con la información del foro, concluyendo en que es posible predecir con cierta precisión los movimientos en los precios de acciones.

Capítulo 3

Herramientas Utilizadas

Procesamiento del Lenguaje Natural

Una de la Herramientas claves para cumplir con los objetivos planteados es la utilización del NLP (Natural Language Processing o Procesamiento del lenguaje Natural). El NLP es un procedimiento perteneciente al campo de las ciencias computacionales, con sus inicios desde la presentación de un artículo académico escrito por Alan Turing (1950) titulado "Computing Machinery and Intelligence", donde el autor estudió la posibilidad de que “las máquinas puedan pensar”. La NLP toma como variable de entrada el lenguaje natural del ser humano y lo procesa, obteniendo como resultado la interpretación por medio de un sistema computacional, pudiendo dotar a la máquina de una capacidad de interpretar y distinguir el lenguaje humano.

Hoy es una arista muy importante para la tecnología moderna, se utiliza en todo tipo de inteligencia artificial que nos facilita la vida, desde algoritmos de recomendaciones de contenido, respuestas automáticas, hasta interpretación del habla humana.

Puntualmente la herramienta de NLP a utilizar en este estudio, es una librería del lenguaje de programación python llamada “Textblob”.

Textblob

Es una librería Python para el procesamiento de datos textuales que provee una serie de herramientas y bases de datos para realizar tareas relacionadas con el NLP sencillamente. Es ampliamente utilizada para la traducción, interpretación, clasificación, tokenización y análisis de todo tipo de texto, además de ser una propuesta sólida para encarar el análisis de sentimiento en texto.

La ventaja de trabajar con textblob es que cuenta con una base de datos de una gran cantidad de palabras ya clasificadas, por lo que, al utilizar el método para la obtención de polaridad en texto, ésta realiza comparaciones del texto con su base de datos para determinar el resultado. Al utilizar textblob se pueden realizar estudios de gran cantidad de datos como es el caso de este trabajo, no siendo necesario el desarrollo de programas complejos desde cero, sino que se puede partir de una base ya probada y testeada por muchos expertos.

El proceso que se realiza al utilizar textblob incluye en primer lugar, una tokenización, o sea la separación de la frase en palabras y frases más pequeñas, lo que permite la identificación de sustantivos, verbos y adjetivos para luego comparar su carga de polaridad implícita por separado, otorgando un valor promedio a la frase en su conjunto, para determinar la polaridad de las palabras, textblob posee una lista de palabras en la nube, que es consultada y con la que compara la frase estudiada. El resultado al hacer un requerimiento de polaridad es un valor numérico entre 1 y -1, donde 1 significa que la frase tiene una polaridad 100% positiva y -1 100% negativa, el resultado cero indica una polaridad indefinida o neutral.

La amplia lista de referencia de textblob se encuentra por el momento solo en idioma inglés, por lo que para el análisis en español no tiene ninguna utilidad y se debería traducir el texto o acudir a otras metodologías.

API de Twitter

Application Programming Interface o en español Interfaz de Programación de Aplicaciones, es un conjunto de protocolos de programación que permite la utilización de productos y servicios computacionales, simplificando la interacción y el desarrollo de programas.

Su utilización permite al programador aprovechar información y facilidades que ofrece una aplicación, por ejemplo, Twitter, que ofrece un servicio de API gratuito, así como también uno pago. El API de Twitter, es el protocolo estandarizado por el cual se puede hacer requerimientos con ciertas limitaciones, accediendo mediante la creación de una aplicación que

otorga permisos y credenciales para poder utilizar la información de la red social mediante la utilización de un lenguaje de programación, en este trabajo se utilizará el lenguaje de programación Python.

Python

Es un lenguaje de programación muy utilizado a nivel global para trabajos de Big data, Machine Learnig, y todo tipo de cálculo matemático y estadístico, siendo un lenguaje de programación orientado a objetos y de muy alto nivel. La programación en Python se caracteriza por buscar siempre la simpleza en el código, es un leguaje de rápida escritura y ejecución, resulta natural elegir entre Python y R como lenguajes principales para cumplir con objetivos como los de este trabajo. Usar Python permitirá el aprovechamiento de la API de twitter para la extracción de datos y cumplirá un papel muy importante en si procesamiento y en la obtención de resultados.

Capítulo 4

Metodología utilizada

Se eligió analizar el contenido de la red social twitter porque se caracteriza en que su principal contenido publicado por las personas es en su mayoría Texto, lo que resulta ser más accesible actualmente a la hora de un análisis e interpretación, a diferencia de otras redes utilizadas en Argentina como Facebook e Instagram donde el contenido es principalmente gráfico.

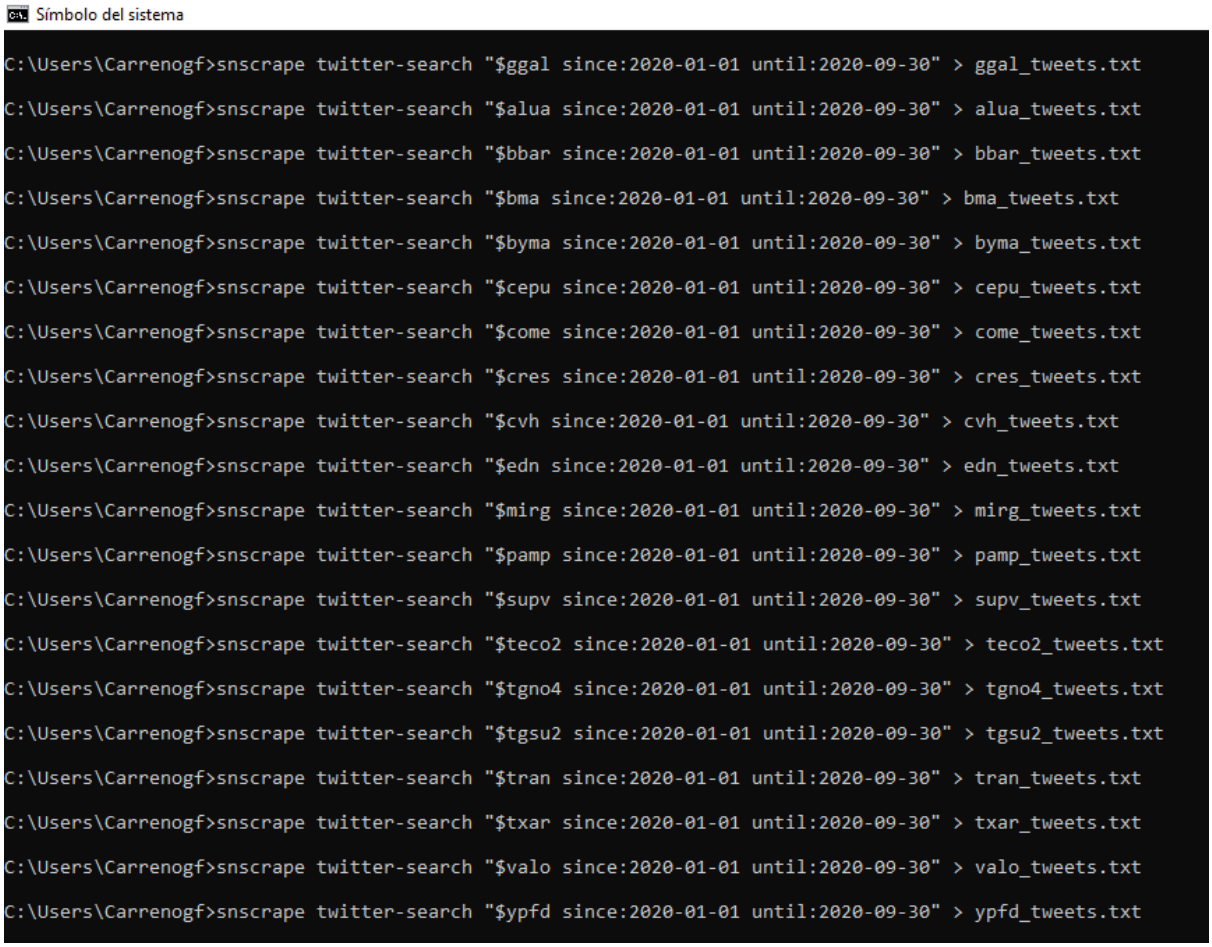
Extracción de datos

La extracción de los textos de los tweets se realizó mediante la combinación de un método de web-scraping (snscraper en python) para obtener el código ID de cada tweet para luego mediante consultas con el método Twitter Search de la API de twitter, obtener los tweets específicamente relacionados con cotizaciones bursátiles en el índice MERVAL, se utilizó como palabras clave para el criterio de búsqueda los tickers de cada acción que actualmente forma parte del índice MERVAL, añadiendo el símbolo “$” antes del ticker, simbología que en twitter se acostumbra a utilizar para referirse a un ticker bursátil. De modo que se extrajeron los Tweets de cada ticker correspondientes al periodo que va desde el 01/01/2020 al 30/09/2020. Luego de una primera limpieza y ajustes a la estructura de los datos para obtener una base de datos uniforme y sin espacios vacíos ni errores, se obtuvo un total de 50405 tweets, en una matriz que contiene la fecha y el texto de cada uno.

El texto de cada tweet contiene una cadena con caracteres alfanuméricos, además de caracteres especiales como @, $, %, #, &, etc., también pueden encontrarse elementos que no suman ningún aporte para el análisis como ser menciones a otros usuarios, hashtags y urls que deben ser extraídos como se aprecia en la Figura 3

Figura 3

Scripts utilizados para la extracción de los IDs de Tweets con criterio de búsqueda



Nota. Uso de snscraper en python para extraer los ID de los tweets.

Fuente: elaboración propia

Cada consulta se guardó en un archivo de texto plano de tipo TXT, que luego fue unificado en uno solo para poder simplificar el proceso del paso siguiente de la Figura 4.

Figura 4

Código Python aplicando el API de twitter para buscar el texto de cada ID extraído.

Script de extración de tweets de una lista de ID

Credenciales API twitter

```
import pandas as pd, tweepy

# Acceso credenciales API app Twitter

# Consume:
CONSUMER_KEY = '[redacted]'
CONSUMER_SECRET = '[redacted]'

#Access:
ACCESS_TOKEN = '[redacted]'
ACCESS_SECRET = '[redacted]'

auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
auth.set_access_token(ACCESS_TOKEN, ACCESS_SECRET)
api = tweepy.API(auth)
```

Importar archivo con IDs al dataframe

```
#Abre con pandas el TXT obtenido de la consulta con snscraper
tweet_url = pd.read_csv("imv_tweets.txt", index_col= None, header = None, names = ["links"])
```

Separar el texto de cada registro en ID y url

```
af = lambda x: x["links"].split("/")[-1]
tweet_url['id'] = tweet_url.apply(af, axis=1)
```

Lista de tweets y variables parametro

```
ids = tweet_url['id'].tolist() #crea una lista de los ID
total_count = len(ids) #determina cuantos ID hay en la lista
chunks = (total_count - 1) // 50 + 1  #variable que determina en cuantas partes (chunks) se hará el proceso para evitar errores
```

Definicion del metodo que ejecutará la busqueda

```
def busca_tw(ids):
    list_of_tw_status = api.statuses_lookup(ids, tweet_mode= "extended",truncated="false")
    empty_data = pd.DataFrame()
    for status in list_of_tw_status:
            tweet_elem = {"tweet":status.full_text,
                          "date":status.created_at}
            empty_data = empty_data.append(tweet_elem, ignore_index = True)
    empty_data.to_csv("imv_tweets.csv", mode="a")
```

Ejecucion del metodo de busqueda

```
for i in range(chunks):
        batch = ids[i*50:(i+1)*50]
        result = busca_tw(batch)
```

Fuente: Elaboración Propia

El Script python utilizado en la Figura 4 sirve para la búsqueda de los tweets en base a la lista de IDs obtenida en el paso anterior, como resultado crea un archivo de texto plano con el texto del tweet y la fecha.

Para poder extraer información de twitter es necesario en primer lugar crear un perfil de desarrollador, explicando los motivos para los cuales se desea acceder a la información pública, luego se debe crear dentro de la interfaz de la página de apps de twitter una aplicación, que otorga credenciales de acceso compuesta por un código alfanumérico único para cada aplicación, en el script se las identifica tachadas en gris por motivos de privacidad.

La API de twitter tiene una restricción para extraer tweets gratuitamente con criterios de búsqueda generales hasta 30 días atrás como máximo, por este motivo fue que se utilizó un medio de web-scrapping para obtener los ID de los tweets y luego por medio de la API hacer una consulta específica de cada ID, cosa que no tiene limitaciones temporales.

Si bien el código es corto, es el resultado de muchas pruebas y errores, y de haber pasado por varias versiones hasta poder obtener el resultado esperado.

El proceso de extracción final demoró alrededor de 1 hora para completar con todos los requerimientos hechos a la API de twitter, debido a la gran cantidad de información histórica solicitada, sin embargo, si se buscara analizar información en tiempo real este proceso sería seguramente instantáneo.

La información obtenida por este método resultó tener muchos defectos que debían ser corregidos antes de pasar al análisis, debido a que la matriz resultado tenía falencias en su estructura ya que existían espacios vacíos, o tweets que estaban divididos en partes por ser muy extensos, por lo que se realizaron tareas de corrección en la estructura de la matriz eliminando espacios vacíos y combinando los tweets que estaban separados, tarea que se completó con el lenguaje de programación VBA (Visual Basic Applications) en Microsoft Excel, aplicando el código de la Figura 5.

Figura 5

Código VBA utilizado para la corrección de la estructura de la matriz de datos.

```
Sub Arreglo_tweets()
Range("B2").Select
Do While ActiveCell.Offset(0, -1) <> ""
    If IsNumeric(Left(ActiveCell.Offset(0, -1), 1)) Then
        ActiveCell = ActiveCell.Offset(0, -1)
    For i = 1 To 15
        If Not IsNumeric(Left(ActiveCell.Offset(i, -1), 1)) Then
            ActiveCell = ActiveCell & ActiveCell.Offset(i, -1)
        Else: Exit For
        End If
    Next i
    ActiveCell.Offset(i, 0).Select
    Else:
    ActiveCell.Offset(1, 0).Select
    End If
Loop
End Sub
```

Fuente: Elaboración Propia.

En el código de la Figura 5 se observa un método que recorre toda la lista de tweets compuesta por 2 columnas, Fecha y Tweet y detecta cuando hay una fecha que es un dato numérico, en caso contrario (no hay fecha) concatena el texto de esa fila con el de la anterior, uniendo los tweets que estaban cortados en un solo texto y relacionado a una sola fecha. Eliminando las filas vacías se obtuvo al fin una matriz con una estructura correcta con 50405 registros.

Preparación y Limpieza de los Datos

Si bien hasta este punto ya se realizó una primera limpieza de datos, ésta sólo se orientó a corregir la estructura de la matriz y a quitar tweets Re twitteados y repetidos ya que algunos hacían referencia a más de un ticker, por lo que entraron dentro del criterio de búsqueda en más de una acción, lo que seguía, es limpiar cada uno de las cadenas de texto de los elementos que resultan en ruido para el estudio, como ya se dijo, hashtags, menciones, urls, símbolos, emojis, y además de la mención de cada ticker etc.

Luego de la limpieza, la matriz de 50405 registro se vio reducida a 46126 registros, perdiendo 4279 tweets, esto debido a que algunos tweets sólo contenían la mención del ticker por la que entraron dentro del criterio de búsqueda, o solo una url o una mención a otro usuario, y que no aportan nada para el análisis, quedando completamente vacíos luego de la limpieza. Adicionalmente se quitaron los tweets que contenían menos de 3 caracteres o que solo quedaron con espacios en su contenido luego de la limpieza.

Traducción y clasificación de polaridad

El método elegido para el análisis de sentimientos es el de la muy difundida y utilizada librería de procesamiento del lenguaje natural de python conocida como “TextBlob”, el inconveniente es que esta librería solo tiene la capacidad de analizar texto en el idioma inglés, lo que resulta en una dificultad por que la gran mayoría de tweets del universo analizado están escritos en lenguaje español, por lo que se requería traducir cada uno de los 46126 tweets, de los cuales en algunos casos alcanzaban a superar los 800 caracteres. Para proceder con la traducción se trabajó con el servicio ofrecido gratuitamente de googletranslate usando la librería de python “googletrans”, servicio conocido mundialmente por su gran capacidad de traducir correctamente gracias a su inteligencia artificial.

Se podría haber optado por otro método que contemple el idioma español, pero esto requeriría un esfuerzo mayor además de muchísimo más tiempo y datos, una forma para realizar un análisis en español sería elaborando un Clasificador Bayesiano, es un sistema cuya principal característica es la ingenuidad, porque, al principio no tiene ninguna información, es ingenuo, por lo que para que funcione debe ser alimentado con información ya correctamente clasificada, como el análisis del lenguaje natural es tan amplio se optó por Textblob, ya que para el clasificador bayesiano se requería un gran caudal de información.

Mediante un script en python se realizó en pasos separados la limpieza, traducción y clasificación de polaridad de todos los tweets.

Como ya se indicó en el funcionamiento de textblob, la polaridad obtenida se clasifica en un valor que oscila entre -1 y 1, indicando negatividad en valores menores a cero y positividad en valores mayores a 0, y neutralidad en valores iguales a cero.

Figura 6

Código Python para la Limpieza traducción y clasificación de los tweets.

Script de limpieza, traduccion y clasificiación

Librerias a utilizar

```
import pandas as pd
from textblob import TextBlob
import csv
import re
import time
import string
from googletrans import Translator
```

Importar datos al dataframe

```
df = pd.read_csv('Data.txt', sep=';')
```

Definición del metodo de limpieza

```
def clean_text(text):
    text = re.sub(r'^RT[\s]+', '', text)
    text = re.sub(r'https?:\/\/.*[\r\n]*', '', text)
    text = re.sub(r'#', '', text)
    text = re.sub(r'@[A-Za-z0-9]+', '', text)
    return text
```

Ejecución del metodo de limpieza

```
df['Clean_tweet'] = df['Clean_tweet'].apply(str)
```

Definición del metodo de traducción

```
def traduct(texto):
    translator = Translator()
    try:
        result= str(translator.translate(texto,dest='en').text)
    except:
        result= texto
    return result
```

Ejecución del metodo de traducción

```
df['En_tweet'] = df['Clean_tweet'].apply(traduct)
```

Definición del metodo de Clasificación por polaridad

```
def get_polarity(texto):
    analysis=TextBlob(texto)
    result = analysis.sentiment.polarity
    return result
```

Ejecución del metodo de Clasificación por polaridad

```
df['polarity'] = df['En_tweet'].apply(get_polarity)
```

Guardar el dataframe resultado en un archivo CSV

```
df.to_csv('resultado.csv',sep=';')
```

Fuente: Elaboración Propia.

Del Script utilizado en la Figura 6 para limpiar, traducir y clasificar según polaridad, como resultado se obtuvo un archivo de texto plano de tipo csv, que servirá como materia prima del análisis y comprobación o no de la hipótesis planteada. La ejecución del mismo demoró alrededor de 8 hs. en terminar.

Una vez terminado con estos procesos pudo obtenerse una base de datos que estaba compuesta por los siguientes datos:

Tabla 1

Descripción de la matriz de datos

Nombre:	Etiqueta	Fecha	Texto Original	Texto limpio	Texto traducido	Polaridad
Tipo de Dato:	Número Entero	Fecha	Cadena de Texto	Cadena de Texto	Cadena de Texto	Número decimal
Observ.:	N° orden	Fecha tweet	-	-	-	Valor entre -1 y 1

Fuente Elaboración propia.

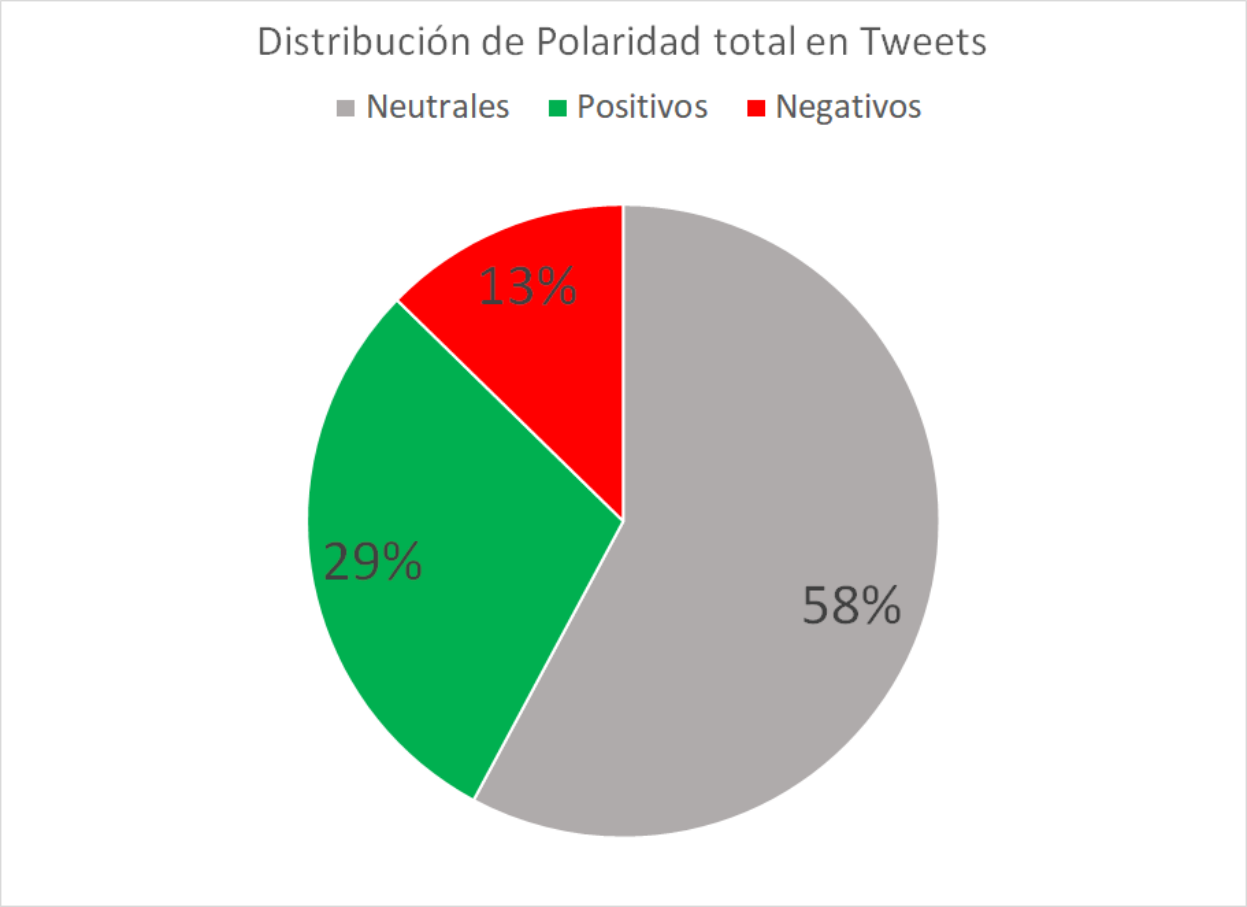
Capítulo 5

Resultados obtenidos

Los 46126 tweets analizados se clasificaron en positivos, negativos y neutros, obteniendo el resultado mostrado en la Figura 7.

Figura 7

Gráfico de Torta con la distribución de polaridad en tweets extraídos.



Fuente: Elaboración propia.

Del total analizado 26.682 tweets resultaron neutrales representado el 58% de los datos, 13.593 Positivos con el 29% y 5.851 Negativos con el restante 13%, ponderando con

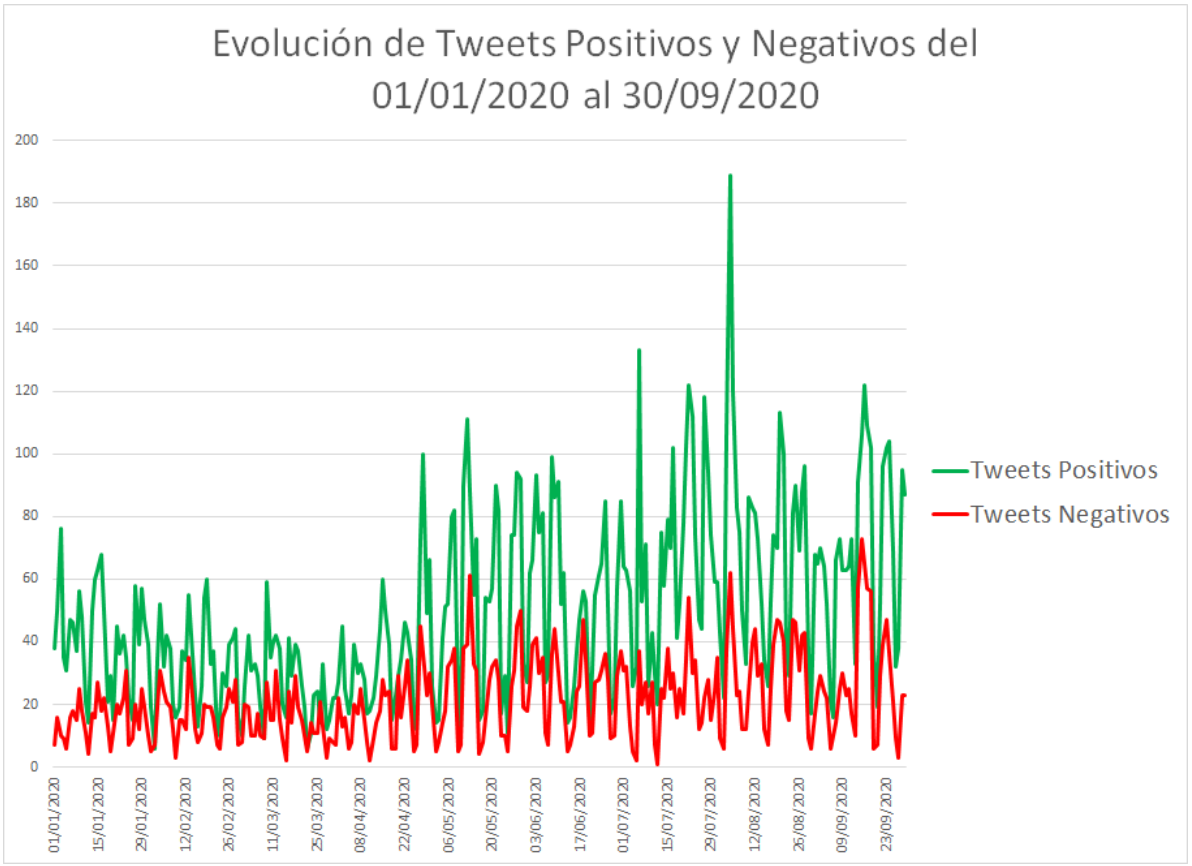
gran mayoría los neutrales debido a que el análisis textual realizado no logra identificar una polaridad definida. Sin tener en cuenta los neutrales, que no suman datos relevantes al objetivo de estudio, tenemos que la mayoría de tweets con carga en su polaridad son positivos, indicando que el sentimiento de periodo analizado en su mayoría tiende a la positividad.

Tweets Positivos y Negativos

Al ordenar cronológicamente y acumular en días la información, se obtiene las siguientes series (Figura 8) de la evolución en la cantidad de tweets positivos y negativos a lo largo del periodo bajo estudio.

Figura 8

Series de la evolución en la cantidad de tweets positivos vs negativos.



Fuente: Elaboración propia.

A simple vista se nota que la cantidad de tweets positivo se incrementó notablemente a partir del mes de abril, teniendo un pico máximo de 189 tweets positivos en un día el 04/08/2020 (Figura 9), fecha en que el índice MERVAL tocó máximos históricos en 56.114,04 puntos, sin haber sido superados hasta el momento de esta redacción, lo que a priori podría sugerir que es posible una relación en algún grado entre los precios del mercado y lo plasmado en redes sociales.

Figura 9

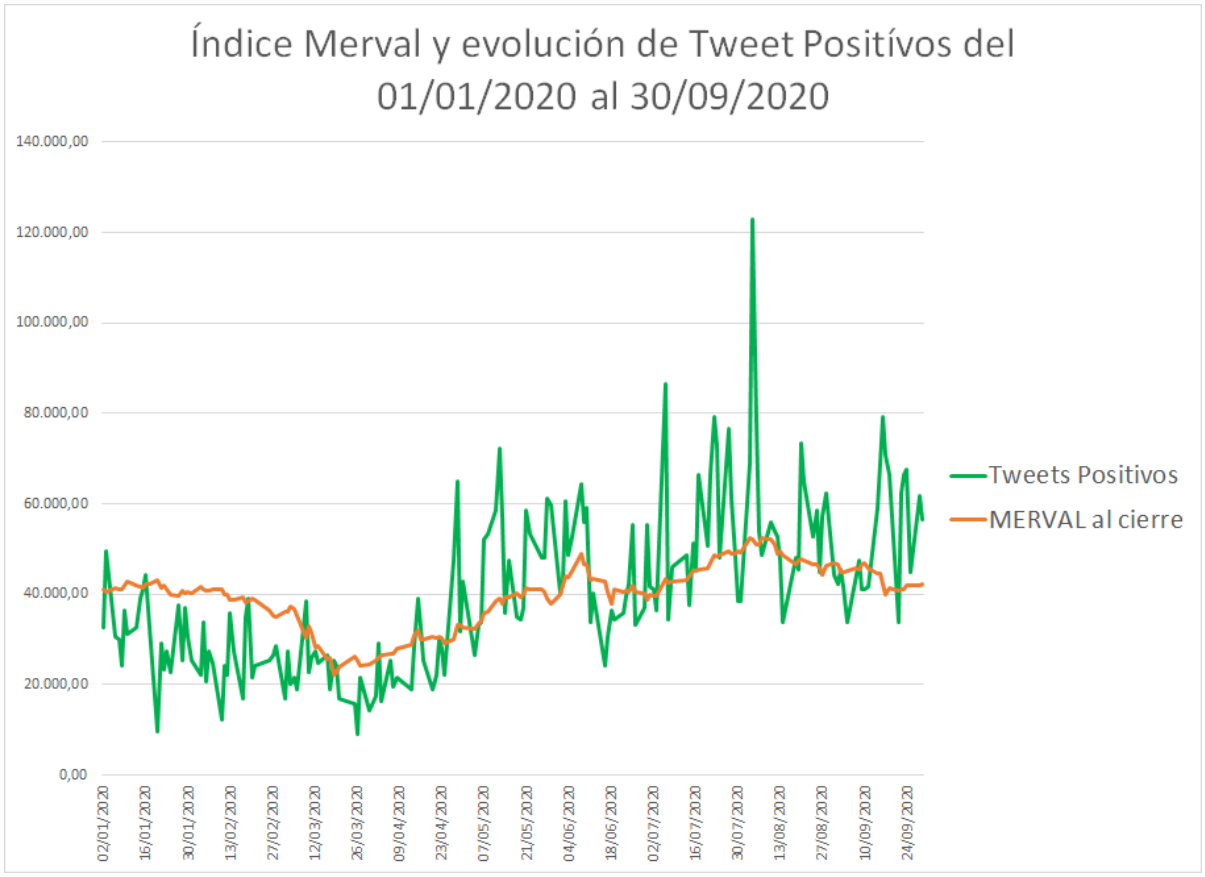
Gráfico de velas de la cotización del Índice Merval desde Mayo a Octubre de 2020.

Fuente. Adaptado de tradingview.com , https://es.tradingview.com/chart/X9YRbapu/

La evolución de tweets positivos sugiere una tendencia alcista en el sentimiento, tendencia que también repite el índice MERVAL en el mismo período, con altos y bajos similares en ambas series, como puede apreciarse en la Figura 10:

Figura 10

Índice MERVAL al cierre en pesos vs evolución de Tweets positivos.



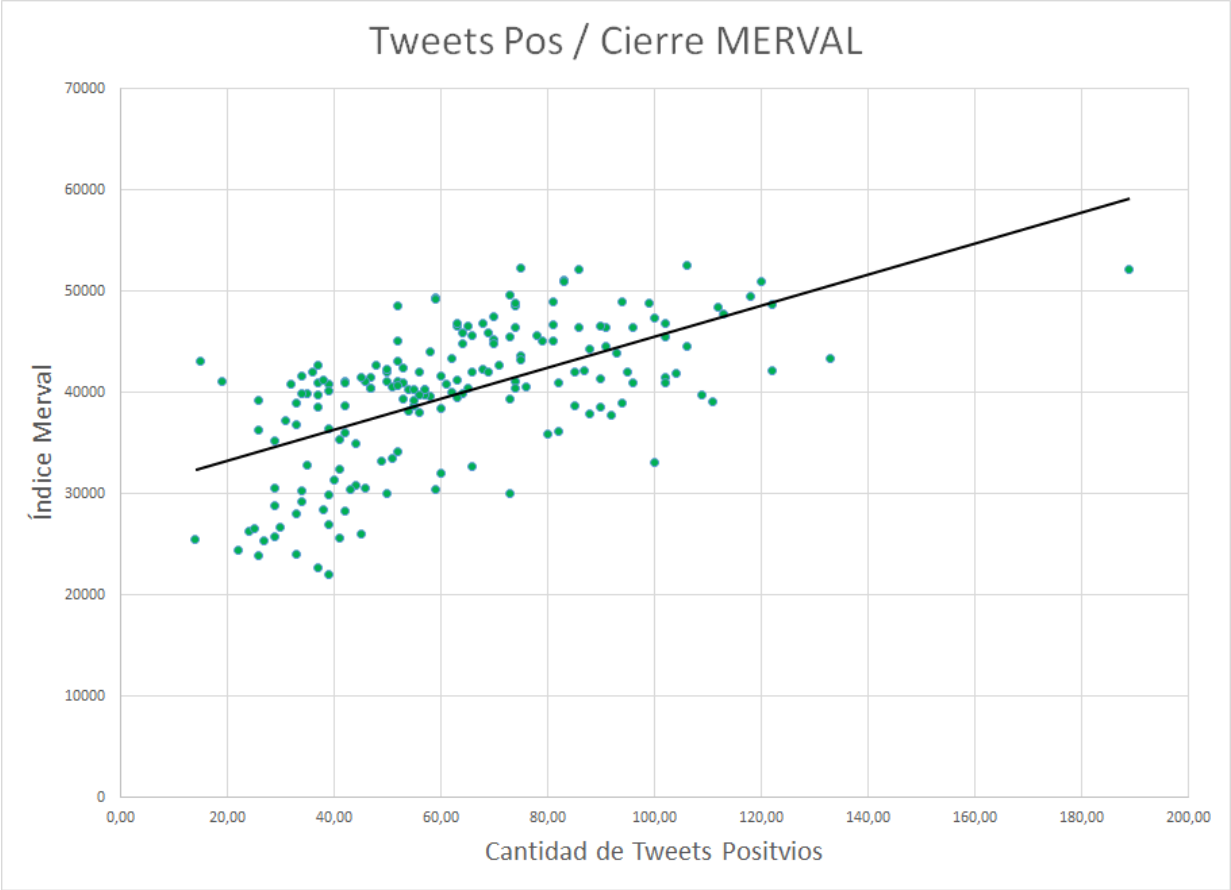
Fuente: Elaboración propia.

En el gráfico de la figura 10, a serie de la evolución de tweets positivos, fue adecuada a una escala de 650 a 1 para aproximar su gráfico al del índice MERVAL.

A este punto resulta por lo menos interesante conocer si existe al menos una correlación entre estas series, ya que a simple vista parecen moverse de una manera muy similar, salvando que la cantidad de tweets positivos muestra una variación mucho más violenta con respecto al cierre del índice Merval. Se realizó una regresión lineal entre las variables recién mencionadas (Figura 11).

Figura 11

Gráfico de regresión lineal entre la evolución de tweets positivos y el índice MERVAL al cierre.



Fuente: Elaboración propia.

Según el modelo de regresión lineal se puede observar una relación positiva en el comportamiento del cierre del índice MERVAL y la variación en el número de tweets positivos, arrojando la relación entre éstos, un coeficiente de correlación del valor de 0.60, sugiriendo una correlación positiva de una intensidad mediana.

Elaboración del Índice de sentimientos

Para cumplir con el objetivo de elaborar un índice, es necesario tener en cuenta que éste debería poder replicarse en el futuro, por lo que debería poder adaptarse a los cambios

que puedan darse en las costumbres de las redes sociales, al ser las redes tan cambiantes y encontrarse en constante crecimiento, es casi seguro que el número de comentarios no hará más que seguir aumentando cada vez más, situación que debe contemplarse en la elaboración del índice.

Para poder adaptarse al futuro número, no sería recomendable basar el índice directamente en la suma de tweets positivos o negativos, como se hizo hasta el momento, ya que lo que podría ocurrir es que, por el incremento de partícipes en las redes, el índice no haga más que subir y pierda toda correlación con el mercado, por lo que el resultado que arroje debería ser relativo a la cantidad de tweets y tener en cuenta el balance entre tweets de polaridad positiva y negativa.

Un índice basado únicamente en la suma de la cantidad de tweets podría funcionar, pero solo para periodos cortos donde las variaciones en la masa que expresa su opinión no sean tan significativas.

Para solventar esto, se pensó en la razón entre tweets positivos y negativos, pero lo que sucedía era que el resultado era igual para un día que tiene 200 tweet positivos y 100 negativos, en comparación con otro día que solo tenía 2 positivos y un negativo. La razón Positivos/negativos es la misma, 2 a 1, añadiendo un simple factor de corrección que contempla la cantidad de tweets relativa al primer día medido, dando mayor significatividad a resultados de días con más cantidad de tweets y menor ponderación a días con menor cantidad, este factor de corrección puede solucionar este defecto. La función para el cálculo del Indicador de sentimientos quedaría de la siguiente forma:

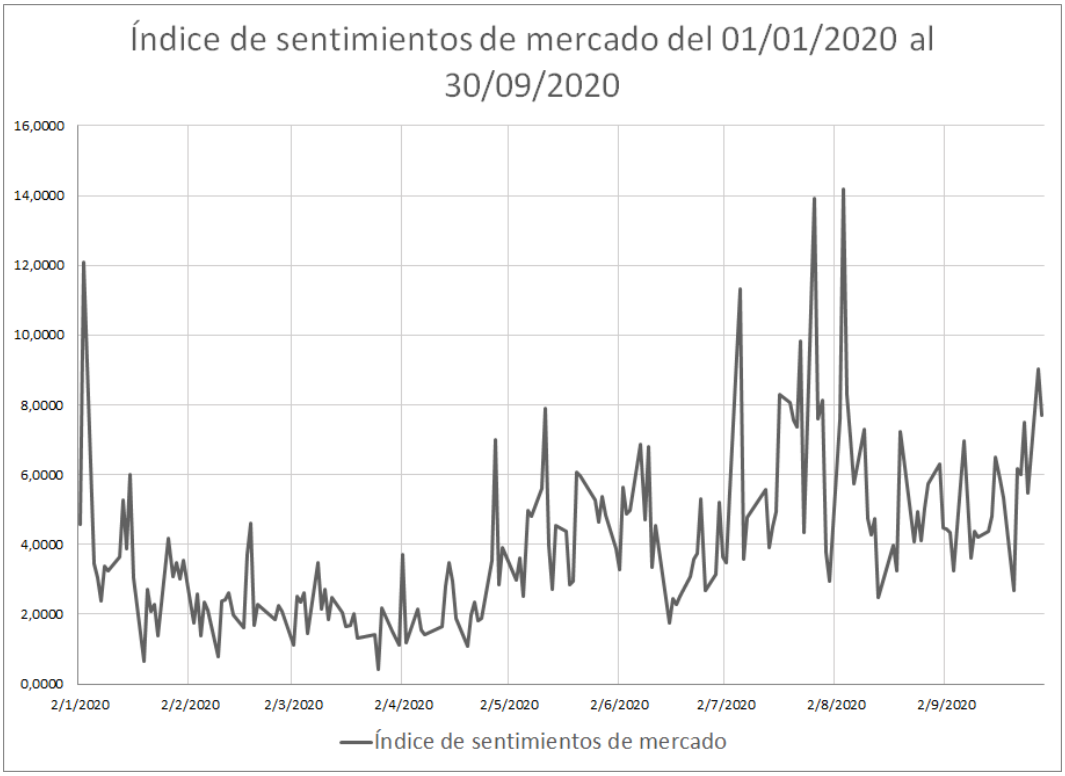
$$\boldsymbol{ISM_d = \frac{\sum Tpos_d}{\sum Tneg_d} * \frac{\sum Tpos_d + \sum Tneg_d}{\sum Tpos_{d0} + \sum Tneg_{d0}}}$$

En la fórmula, d indica el día que se pretende calcular, ISM es el índice de sentimientos de mercado, Tpos y Tneg son los tweets positivos y negativos respectivamente, d0 hace referencia al primer día de medición que sirve como parámetro de ponderación.

Aplicando esta función a los datos de los tweets obtenidos y procesados se obtiene la serie graficada en la Figura 12.

Figura 12

Índice de sentimientos de mercado del 01/01/2020 al 30/09/2020.



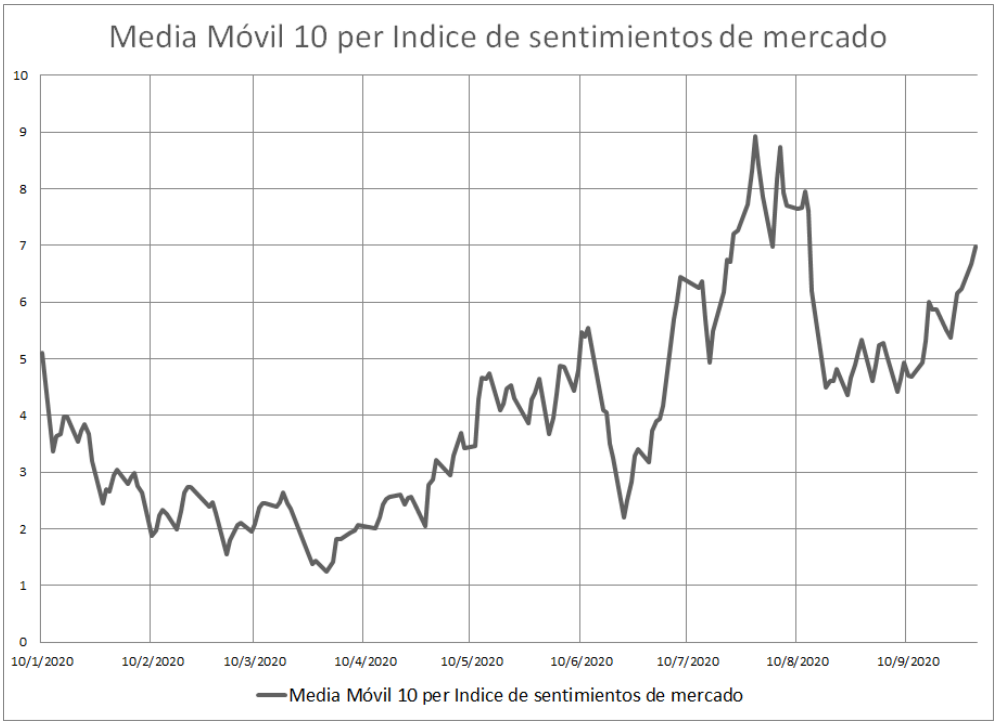
Fuente: Elaboración propia.

Para poder entender mejor el sentimiento o humor del mercado, es necesario reconocer su tendencia, una medida comúnmente utilizada para este fin es la utilización de medias móviles que dan lugar a una serie de promedios, lo que se busca es suavizar las fluctuaciones de corto plazo para hacer más evidentes las tendencias en el gráfico y en los valores del índice, a continuación se aplicó el cálculo de la media móvil de 10 periodos, para disminuir la dispersión excesiva, perdiendo poca información en el proceso y sin quitar

demasiada ponderación al dato de cada día, además la utilización de medias móviles permite la incidencia de los tweets publicados en fines de semana, pudiendo estos incidir dentro del promedio móvil de los días siguientes, ya que para comparar las series solo se utilizan los datos de días en los que hay cotización, resultando en la siguiente serie en la Figura 13:

Figura 13

Media móvil simple de 10 periodos aplicada al Índice de sentimientos.



Fuente: Elaboración propia.

Al aplicar el promedio móvil se hacen más claras las tendencias y se redujo en gran medida la dispersión de la serie.

Para probar la utilidad o no, de este índice elaborado se lo comparará con las cotizaciones al cierre del índice MERVAL (Figura 14) como ya se hizo con la serie de tweets positivos.

Figura 14

Comparación grafica entre el Índice MERVAL y el índice de Sentimientos MM10.

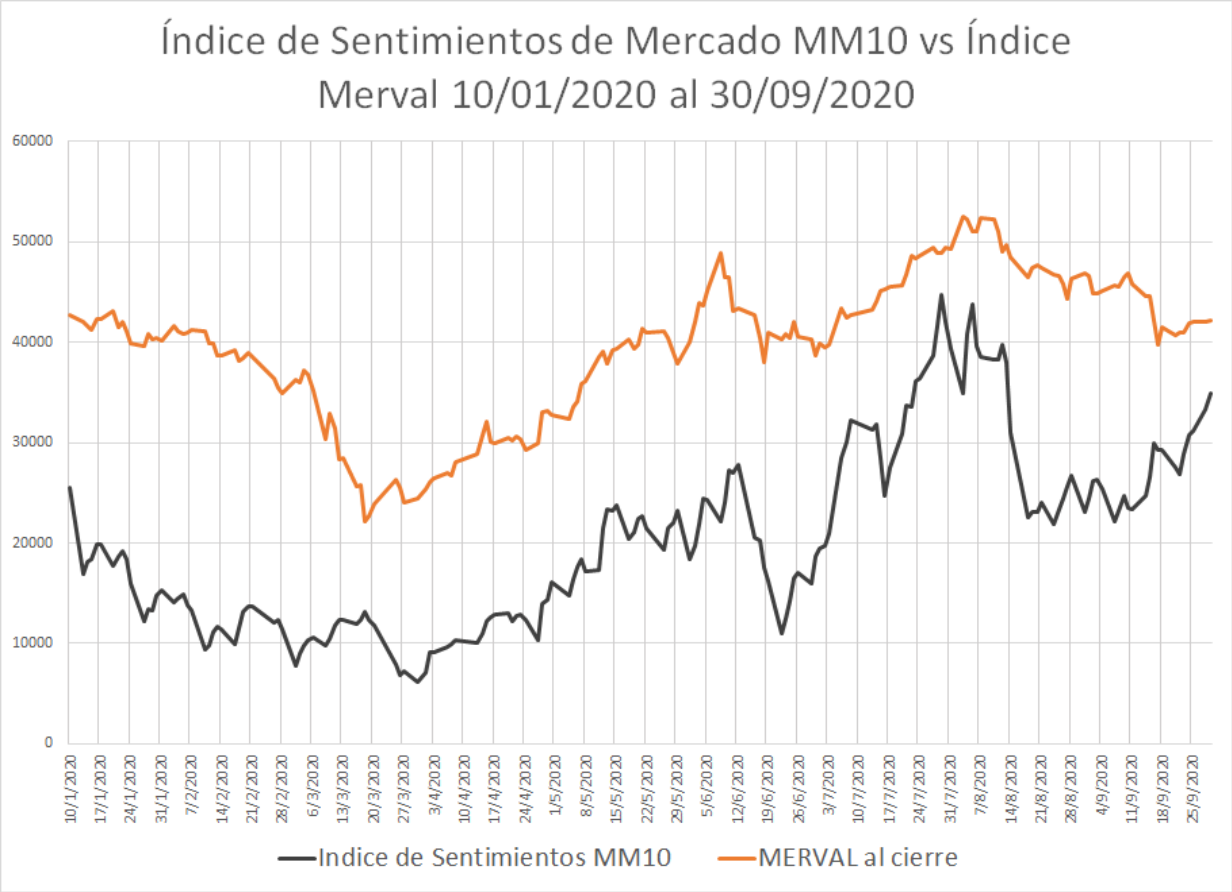


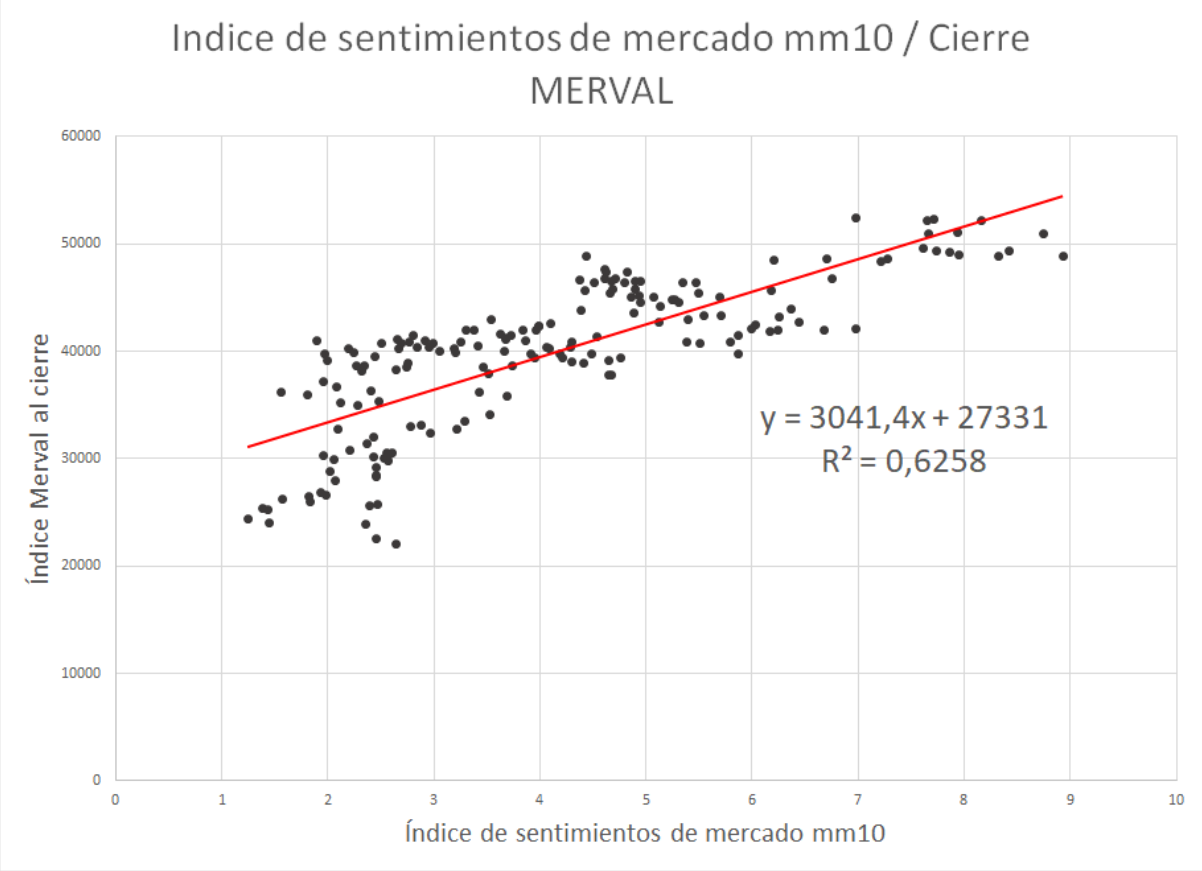
Gráfico comparativo Índice de sentimientos mm10 con Índice Merval en el periodo estudiado, se adaptó el Índice de sentimientos a una escala de 5000 a 1 para facilitar la comparación en un mismo gráfico.

Del estudio de la relación entre ambas variables para el periodo elegido, surge que poseen un coeficiente de correlación de 0.7910, y un coeficiente de correlación de Pearson de 0.6258 indicando lo que podría ser una fuerte correlación positiva.

Para poder determinar una relación causa efecto, haría falta abarcar información correspondiente a un periodo mayor, sin embargo, es suficiente para dar conclusión a los objetivos de este trabajo conocer si existe una relación, como se muestra en la Figura 15.

Figura 15

Regresión lineal entre el índice de Sentimientos con el índice MERVAL.



Fuente: Elaboración Propia.

En el gráfico de dispersión de la Figura 15, se relacionó los valores del Índice de Sentimientos mm10 y el Merval, trazando una regresión lineal, se observa la clara correlación positiva.

Capítulo 6

Conclusiones

De la interpretación de los procesos aquí detallados, se puede llegar a la conclusión de que existe una correlación positiva para el periodo enero-septiembre de 2020, entre el movimiento del mercado argentino representado por el índice MERVAL y el índice elaborado de sentimientos de mercado en twitter, lo que permite pensar en la posibilidad de que se puede cumplir la premisa defendida por la teoría de las finanzas conductuales, por consiguiente, la existencia de tendencias, tanto en los precios como en sentimientos de inversores, sugiriendo que los precios se mueven en consonancia y se ven afectados por los market feelings, dando lugar a la posibilidad de incoherencias y sucesos inesperados en los precios. Quedando así cumplido el objetivo general planteado para este trabajo.

En cuanto a los objetivos específicos, se pudo obtener datos suficientes y aptos para el análisis pretendido, para posteriormente clasificarlos con inteligencia artificial según su polaridad. Lo que permitió finalmente la construcción del índice de sentimientos de mercado y probar su relación o no, con los precios del índice MERVAL.

De este resultado se puede reconocer la capacidad de las redes sociales para crear tendencias que pueden afectar las expectativas de los inversores expuestos al contenido compartido en ellas. Evidenciando que, el mercado argentino a pesar de su actual poca profundidad y baja participación de la población en la negociación de renta variable, tiene la suficiente eficiencia para reflejar en sus precios el sentimiento actual de los inversores que comparten su opinión y crean tendencias.

Referencias Bibliográficas

Anaya Garay Gonzalo, (2015), Las Finanzas Conductuales, el Alfabetismo Financiero y su Impacto en la Toma de Decisiones Financieras, el Bienestar Económico y la Felicidad, Universidad Católica Boliviana "San Pablo ".

Becerra Carlos Martín, (2016), Análisis de Sentimiento en Twitter: El bueno, el Malo y el >:(, Universidad Nacional de Córdoba.

Dow Charles, (1902), Wall Street Journal.

Fama Eugene, (1970), Mercado de Capitales Eficiente: Una Revisión del Trabajo teórico y Práctico.

Gálvez Ramiro H., (2016), Predicción de los Rendimientos de Acciones en Argentina en Base a Indicadores Técnicos y al Modelado de Tópicos en Foros Bursátiles, UBA.

Hens Thorsten,Meier Anna , (2016), Finanzas conductuales: La psicología de la inversión, Behavioral Finance Solutions.

Malkiel Burton G., (1973), Un paseo aleatorio por wall street, Alianza Editorial.

Medina Andrés García, (2017), El Uso de Twitter en el Análisis Financiero, UNAM.

Murphy Jhon J., (1999), Análisis técnico de los mercados financieros, NY Institute of finance.

Olivo Sergio Luis, (2019), Ciclos bursátiles e indicadores de sentimiento del mercado, XXXVI Jornadas Nacionales de Administración Financiera septiembre 2016.

Turing Alan, (1950), Computing Machinery and Intelligence, The observer.

Timmermann Allan, Granger Clive W.J., (2004), Efficient market hypothesis and forecasting, University of California San Diego.