UNIVERSIDAD
NACIONAL
DE COLOMBIA

MODELO DE RECOMENDACIÓN PARA INVERSIÓN EN ACCIONES COLOMBIANAS PERTENECIENTES AL ÍNDICE COLCAP BASADO EN ANÁLISIS TÉCNICO Y SENTIMIENTO DEL MERCADO LOCAL

Julian Palacio Roldan

Trabajo final presentado como requisito para optar al título de:

Magister en Ingeniería - Analítica

Director:

Prof. Fernán Alonso Villa Garzón, Ph.D.

Universidad Nacional de Colombia

Facultad de Minas

Área Curricular de Sistemas

Medellín, Colombia

2022

Resumen

Modelo de recomendación para inversión en acciones colombianas pertenecientes al índice COLCAP basado en análisis técnico y sentimiento del mercado local

En el año 2018 se creó la Segunda Misión del Mercado de Capitales en Colombia, con el objetivo de estudiar el estado actual de los mercados financieros locales y sugerir soluciones para los problemas encontrados. Entre los hallazgos se encontró que el mercado de acciones ha tenido un decrecimiento notable en los años recientes, por lo que se deben tomar acciones que fomenten la participación en este mercado. Como aporte a la solución, en este trabajo se desarrolla un modelo de recomendación de acciones pertenecientes al índice COLCAP, cuyo objetivo es el de ser incluido en una herramienta que, además de mostrar los resultados del modelo, le permita ver al inversionista cifras de riesgo y los insumos que utilizó el modelo para llegar al resultado: indicadores basados en precios históricos de las acciones, noticias e indicadores financieros de cada emisor. Para ello, se proponen dos modelos de redes neuronales recurrentes que predicen los precios de las acciones, y luego estas predicciones son utilizadas para clasificar las acciones de mejor a peor según la rentabilidad calculada con dichos pronósticos. Para entrenar los modelos, se obtienen datos de precios, noticias y estados financieros, utilizando técnicas de *web scraping* para los primeros dos, luego se realiza un análisis de sentimientos con redes neuronales recurrentes para clasificar las noticias en positivas, negativas o neutras; y se genera un *dataset* final añadiendo variables calculadas. Una vez entrenados los modelos, se utiliza la técnica de validación cruzada para obtener medidas de pronóstico, y luego, después de comparar, se escoge el modelo *gated recurrent unit* como el mejor ya que es más preciso en los pronósticos.

Palabras clave: Acciones, *long short-term memory, gated recurrent unit*, análisis de sentimientos, mercados financieros.

Abstract

Recommendation model for Colombian COLCAP index stocks based on technical analysis and local market sentiment

In 2018 the second capital markets mission was created in Colombia, with the purpose of studying the current state of the local financial markets and to suggest solutions for the problems that were found. Among the findings, it was found that the stock market has had a notable decrease in recent years, so actions should be taken to encourage participation in this market. As a contribution to the solution, in this work, a COLCAP index stock recommendation model is developed, whose objective is to be included in an application that, not only displays the results of the model, but also allows the investor to see risk metrics and the inputs used by the model to get to the result: indicators based on historical stock prices, news and financial indicators for each issuer. To achieve this, two recurrent neural network models are proposed to predict stock prices, and then these predictions are used to classify the stocks from best to worst according to the returns calculated with those same predictions. Prices, news, and financial statements are obtained for the model training, using web scraping techniques for the first two, afterwards, a sentiment analysis is performed with recurrent neural networks to classify the news as positive, negative or neutral; and a final dataset is created adding calculated variables. After the models were trained, a cross-validation technique is used to calculate forecast metrics, and then, after comparing, the gated recurrent unit model is chosen as the best because of its more accurate forecasts.

Keywords: Stocks, long short-term memory, gated recurrent unit, sentiment analysis, financial markets.

Contenido

Pág.

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

Los mercados de capitales son un elemento fundamental para el desarrollo económico de un país, ya que permiten la distribución eficiente de los recursos en una economía a través del direccionamiento de flujos de capital entre sus participantes, donde aquellos con requerimientos de capital, con tal de obtenerlo, están dispuestos a pagar un interés a aquellos que tienen excesos y lo que buscan es rentabilidad.

El mercado financiero colombiano, comenzó a operar a finales de la década de los ochenta, y en sus primeros años era un mercado poco profundo[1], volátil, ilíquido y concentrado; esto llevó a que, en 1995, se realizara la Primera Misión del Mercado de Capitales, cuyo objetivo fue el de identificar posibles cambios en la normativa que permitieran solucionar los problemas del mercado y ampliar su tamaño; como resultado, la misión permitió trazar una hoja de ruta que orientó el crecimiento del mercado en los años subsiguientes (Asociación Bancaria y de Entidades Financieras de Colombia, 2019).

En la historia reciente de los mercados de capitales en Colombia, existen evidencias de que el crecimiento del mercado accionario se estancó en 2012, e incluso algunos indicadores como la liquidez han venido en declive, además de que se han venido reduciendo el número de emisores de valores (Córdoba Garcés & Molina Ungar, 2017). Esta es una de las situaciones que condujo a que, en el año 2019, se lanzara la Segunda Misión del Mercado de Capitales en Colombia, en la cual se confirmó el problema existente en el mercado accionario, y se encontró que, para solucionarlo, se debe promover el mismo, a través de capacitaciones al público, liberalizando la oferta de productos, permitiendo la asesoría de agentes independientes, simplificando la regulación para aumentar la vinculación, entre otros (Asociación Bancaria y de Entidades Financieras de Colombia, 2019).

[1] Mercado que posee un gran número de operaciones, tal que permiten la formación de precios que sean representativos del total de la oferta y la demanda existente (Quiñonez Avendaño, 2010)

El presente trabajo, busca generar un aporte para para la solución de algunos de los problemas encontrados por la Segunda Misión del Mercado de Capitales en Colombia, particularmente en el mercado de acciones colombiano, que necesita nuevos participantes. Para ello se desarrolla un modelo de recomendación de acciones, para luego incluirse dentro de una herramienta para inversionistas, donde se muestre cuáles acciones podrían ser más atractivas en el mercado y las principales variables que explicarían este resultado; como las noticias recientes, indicadores basados en movimientos del precio de la acción y algunos indicadores financieros de la empresa emisora, todos utilizados por el mismo modelo.

Dicha herramienta, promovería el aprendizaje y facilitaría el análisis financiero para las decisiones de inversión a personas naturales, que podrían, en tiempo real, tener acceso a las recomendaciones del modelo y a las variables explicativas. Este tipo de herramientas se hacen necesarias, además, porque recientemente, en el año 2021, se creó en Colombia la primera aplicación móvil para permitir a personas naturales invertir en acciones locales de forma independiente y con bajos costos de comisión (Cajamarca, La República, 2021), lo cual facilitará la entrada de personas naturales al mercado, que además tomarán decisiones de inversión de forma independiente, por lo que van a necesitar herramientas de apoyo en análisis financiero.

Además, en este trabajo se desarrollan y evalúan dos modelos de recomendación de acciones utilizando técnicas de aprendizaje de máquinas, específicamente, redes neuronales recurrentes *long short-term memory* y *gated recurrent unit*. Para el entrenamiento de dichos modelos, se acota el universo de acciones a aquellas incluidas en el índice COLCAP, que representa a las acciones más líquidas del mercado colombiano (Bolsa de Valores de Colombia, 2021), y se extrae, para dichas acciones, los precios históricos, noticias relacionadas con el emisor y estados financieros del emisor; utilizando técnicas de *web scraping* para obtener y actualizar los precios y las noticias de forma automática. Antes de ingresar al modelo, las noticias deben ser categorizadas como positivas, negativas o neutras, por lo que, se desarrolló también otro modelo *long short-term memory* utilizando una capa pre entrenada de *word embeddings*, que se encarga de realizar dicha categorización de forma automática.

La variable de salida para ambos modelos es el precio de cada acción para 1, 10 y 20 días hábiles; una vez el modelo predice dichos precios para todas las acciones pertenecientes al índice COLCAP, la herramienta se encarga de hallar las rentabilidades entre estos precios pronosticados y la fecha actual, para luego, utilizando estas rentabilidades, generar una lista ordenada de acciones de mayor a menor rentabilidad. Así, la herramienta genera 3 listas de recomendación de acciones, una para un plazo de 1 día, otra para un plazo de 2 semanas y otra para un plazo de un mes.

Para este trabajo, se sigue la metodología CRISP-DM, donde se comienza por el entendimiento del negocio, luego el entendimiento y preparación de los datos, seguido de la modelación, la evaluación de modelos y finalmente el despliegue. Todas las fases de la metodología se presentan en los diferentes capítulos del trabajo, de esta manera: en los Capítulos 1 y 2 se presenta el contexto de los mercados financieros y la problemática sobre el mercado de acciones en Colombia, a la que se trata de realizar un aporte con la herramienta de recomendación. En los Capítulos 3 y 4, se plantea el objetivo general y específico de investigación, además de la metodología, donde se describe las actividades a realizar en cada paso. En el capítulo 5 se definen los alcances que tendrá el trabajo. En el Capítulo 6 se realiza una contextualización teórica de conceptos de inversión y de modelos de redes neuronales recurrentes. En los Capítulos 9 y 10 se realiza un análisis descriptivo de los datos, se detalla cómo se realizó la obtención de los datos, su transformación, y se muestra la estructura de los datos en una base de datos SQLite generada para este trabajo. En el Capítulo 10 se define la arquitectura de los modelos, tanto para el modelo de sentimientos para noticias como para el de recomendación de acciones, y se detallan los resultados obtenidos en la validación de modelos. Finalmente, en el capítulo 12 se exponen las recomendaciones finales sobre el trabajo y una serie de recomendaciones para aportar a futuros trabajos relacionados.

1. Planteamiento del problema

Los mercados financieros son un mecanismo donde se facilitan las negociaciones entre compradores y vendedores de instrumentos financieros. El objetivo principal de estos mercados es el de distribuir la liquidez entre los participantes que tienen necesidades de financiación y aquellos que tienen excesos de fondos y buscan rentabilidad (Autorregulador del mercado de valores de Colombia, 2019).

En los mercados financieros alrededor del mundo se negocian diferentes tipos de valores, de acuerdo con Martín (2011) se pueden clasificar en 3:

1. Valores representativos de deuda: instrumentos de renta fija, como Bonos del tesoro, Bonos de empresas privadas, CDT, etc.
2. Valores representativos de propiedad o patrimonio: instrumentos de renta variable, es decir, acciones.
3. Valores representativos de derechos: productos derivados, como futuros, forward, opciones y swaps.

Estos valores, entre otros, son negociados diariamente tanto por entidades financieras como por empresas y personas naturales a través de intermediados del mercado, y sirven para cubrir diferentes necesidades de liquidez o de inversión que requiera cada tipo de participante en su situación particular.

Según Levine (1996) en su investigación para el Banco Mundial, los mercados financieros tienen una influencia positiva sobre la economía, y su desarrollo es un buen predictor de crecimiento económico, acumulación de capital y cambio tecnológico. Esos impactos son explicados por diferentes funciones de los mercados:

1. Facilitar el intercambio, coberturas, diversificación y asignación de riesgos.
2. Asignación de recursos.
3. Ejercer control corporativo.
4. Movilizar el ahorro.
5. Facilitar el intercambio de bienes y servicios.

Otras investigaciones por parte de miembros del Fondo Monetario Internacional han encontrado evidencia de que los mercados financieros desarrollados pueden prolongar las expansiones económicas, traer beneficios para los miembros más pobres de la sociedad y reducir la desigualdad (Chami, Fullenkamp, & Sharma, 2009).

1.1 Mercado de acciones en Colombia

Uno de los valores negociados dentro de los mercados financieros son las acciones, la Bolsa de Valores de Colombia (2021) las define como:

> *"Es un título que le permite a cualquier persona (natural o jurídica), ser propietario de una parte de la empresa emisora del título, convirtiéndolo en accionista de la misma y dándole participación en las utilidades que la compañía genere. Además se obtienen beneficios por la valorización del precio de la acción en las Bolsas y le otorga derechos políticos y económicos en las asambleas de accionistas."*

Siguiendo con esta definición, las acciones son instrumentos de deuda patrimonial que emiten las empresas para financiarse, y que le permiten a personas naturales o jurídicas obtener rentabilidad de dos maneras; primero a través de los dividendos que decrete periódicamente la empresa donde invierte, y segundo, a través de valorizaciones en el precio de la acción, pues podrá venderlas a un precio mayor a aquel por el que las compró.

En Colombia, el mercado de capitales tuvo sus inicios a finales de los años ochenta, con un desarrollo lento debido a deficiencias en cuanto a un marco regulatorio bancario, financiero y de valores (Rojas & Gonzalez, 2008). Más tarde en el año 1995, se realizó la primera misión del mercado de capitales por parte del Ministerio de Hacienda, el Banco Mundial y Fedesarrollo; con el objetivo de asesorar al ministerio en temas de mercado de capitales, como obstáculos a su

desarrollo, aplicabilidad de normas regulatorias internacionalmente y posibles reformas a la estructura institucional (Ministerio de Hacienda, Banco Mundial & Fedesarrollo, 1996). Gracias a los datos obtenidos en ese momento se trazó una hoja de ruta para la evolución del mercado de capitales, que permitió un desarrollo normativo y tecnológico que sigue utilizando. Durante la primera década del siglo XXI, el mercado accionario creció, entraron nuevos inversionistas y ese mercado tenía muy buenas perspectivas, sin embargo, en la segunda década se limitó su dinamismo debido a secuelas de la crisis financiera de 2008, ralentización de la actividad productiva y las caídas en el precio del petróleo; esta situación reveló la necesidad de una segunda misión del mercado de capitales que permitiera generar estrategias para fortalecer este mercado (Asociación Bancaria y de Entidades Financieras de Colombia, 2019).

En el año 2019, se publicó el informe final de la segunda misión del mercado de capitales, donde se encontraron los puntos a trabajar para desarrollar los mercados financieros. En el resumen publicado por la Asociación Bancaria y de Entidades Financieras de Colombia (2019), destacan las siguientes necesidades encontradas:

1. Revisar el régimen de inversiones.
2. Crear programas de educación financiera que permitan capacitar al público y facilitar la vinculación electrónica de clientes.
3. Liberalizar la oferta de productos, permitir la asesoría de agentes independientes y simplificar la regulación para aumentar la vinculación.
4. Simplificar el proceso de emisión, actualizar el régimen jurídico de las sociedades y ofrecer beneficios tributarios a nuevos emisores e inversores.
5. Eliminar barreras regulatorias que incentiven la entrada de administradores de portafolio y resuelvan los conflictos de interés de los conglomerados financieros.
6. Unificar la infraestructura, simplificar el marco normativo y centralizar la información, entre otros.

En la actualidad, analistas profesionales del sector financiero como Andrés Moreno, han expresado su preocupación en el tema, ya que desde el año 2011 salen en promedio alrededor de 65 mil accionistas del mercado accionario colombiano. Ante esta situación, resaltan que una de las principales soluciones a este problema es utilizar estrategias que incentiven a las personas naturales a invertir, como lo es aumentar la educación financiera (Moreno & Humberto, 2019),

una solución que va de la mano con la recomendación número 2 de la segunda misión del mercado de capitales.

Ante la situación actual del mercado accionario colombiano, surge la pregunta de investigación, ¿Es posible desarrollar un modelo de recomendación de inversión en acciones basado en sentimientos en el mercado accionario de Colombia que facilite el proceso de análisis al invertir y permita atraer nuevos inversionistas?

2. Justificación

En el año 2018, se realizó la segunda misión del mercado de capitales por el Ministerio de Hacienda, Banco Mundial y el Programa de Cooperación Económica y Desarrollo, con el objetivo de encontrar elementos que permitan desarrollar los mercados de capitales en Colombia, y de esa manera influir en el crecimiento económico del país. En el informe final, presentado en el 2019, se presentan cifras que demuestran una necesidad de políticas y herramientas que permitan reactivar el mercado accionario en Colombia.

“El mercado de renta variable, de forma similar al mercado de deuda privada, exhibió un rápido crecimiento hasta 2010, para luego presentar un relativo estancamiento hasta 2013 y una notable caída en los años recientes” (Ministerio de Hacienda, Banco Mundial & Programa de Cooperación Económica y Desarrollo, 2019, p.38).

En la **Figura 1** se observa la caída notable del mercado de renta variable desde el año 2013 en Colombia que preocupa a los analistas colombianos y en la **Figura 2** se aprecia este dato comparado con otros países. Para el año 2017 el índice de capitalización bursátil como porcentaje del PIB está por debajo de otros países con mercados más desarrollados como el de Chile.

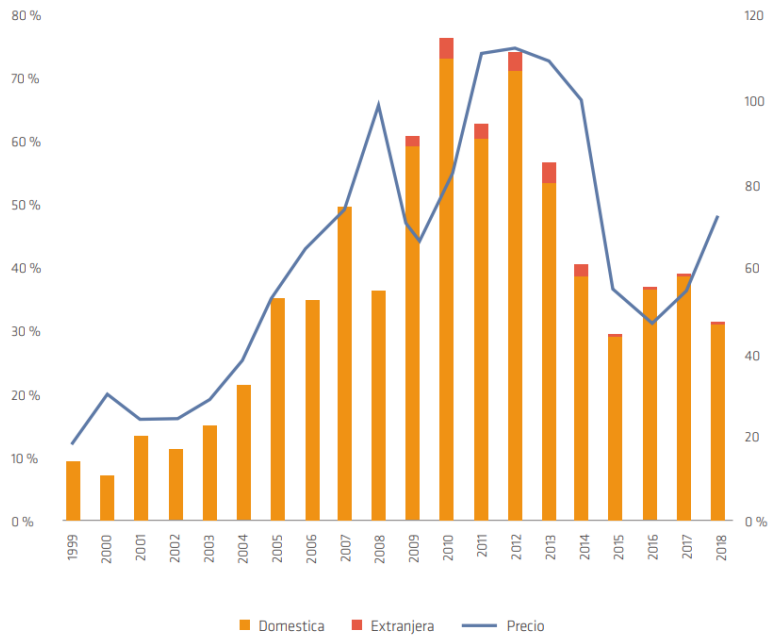

Figura 1. Capitalización bursátil (%PIB) y precio de petróleo (USD)**,** tomado de Ministerio de Hacienda, Banco Mundial & Programa de Cooperación Económica y Desarrollo, 2019.

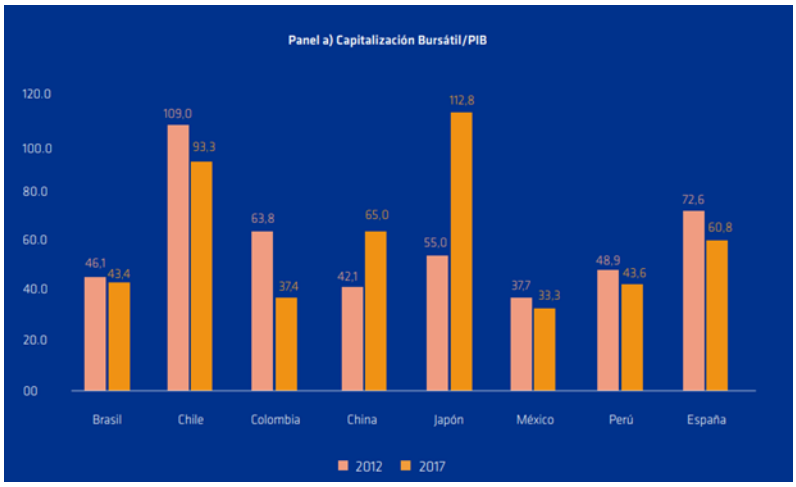

Figura 2. Capitalización bursátil (%PIB) en comparación con otros países, tomado de Ministerio de Hacienda, Banco Mundial & Programa de Cooperación Económica y Desarrollo, 2019.

También el número de empresas que se listan en bolsa es motivo de preocupación, en el año 2001 cuando comenzó la Bolsa de Valores de Colombia estaban listadas 110 acciones de compañías inscritas, mientras hoy se cotizan tan solo 66 (Portafolio, 2020). Como lo demuestra

la **Figura 3**, en comparación con la región, el número de empresas listadas en bolsa en Colombia está por debajo de Argentina, México, Perú, Chile y Brasil.

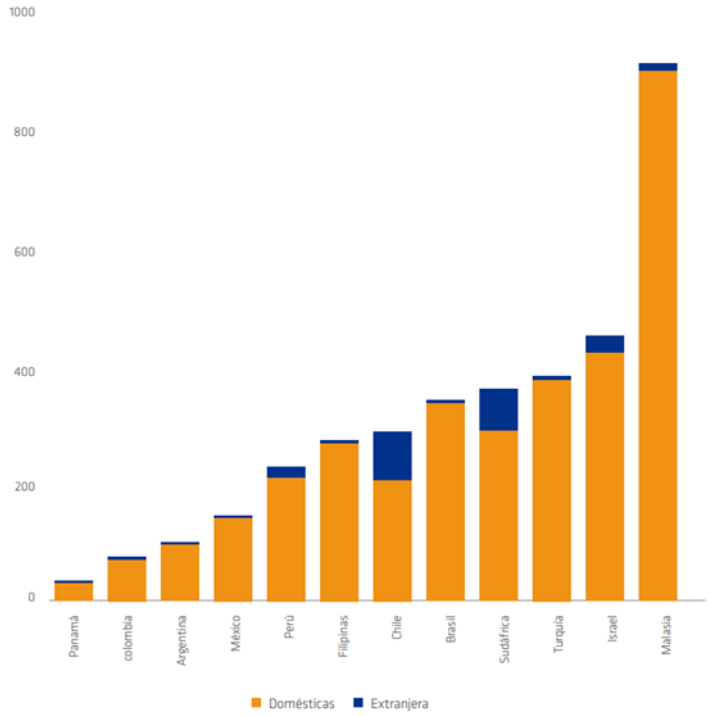

Figura 3. Número de compañías listadas en bolsa por país en 2018, tomado de Ministerio de Hacienda, Banco Mundial & Programa de Cooperación Económica y Desarrollo, 2019.

Uno de los fenómenos que han afectado en el desarrollo del mercado de acciones es que en Colombia el mercado de valores se ha convertido en una bolsa de renta fija (Lagos Cortés, 2013), es decir, Colombia tiene un mercado donde la inversión está muy concentrada en este tipo de títulos. Analizando los montos transados en los sistemas de negociación colombianos como el Mercado Electrónico Colombiano (MEC) de la Bolsa de Valores de Colombia y el Sistema electrónico de Negociación (SEN) del Banco de la República, se puede caracterizar el mercado colombiano por los diferentes productos de los segmentos de renta fija y renta variable.

La información de la **Tabla 1** muestra que los colombianos prefieren invertir en títulos de renta fija que hacerlo en renta variable, y entre estos, prefieren los CDT's. Esta concentración es evidencia del problema encontrado en las opiniones de expertos del mercado y la misión del mercado de valores, muestra que existe un desconocimiento sobre el mercado de acciones y la falta de herramientas que faciliten el acceso y análisis a estos.

Instrumentos de Renta Fija			Instrumentos de Renta Variable
Títulos de tesorería (TES)	**Certificados de depósito a término (CDT)**	**Bonos**	**Acciones**
265.4	18.5	6.17	3.5

Cifras en billones de pesos

Tabla 1. Montos mensuales operados en sistemas de negociación por producto en Colombia para enero del 2019, tomado de Superintendencia Financiera de Colombia, 2019.

Una de las conclusiones de la misión del mercado de valores (2019), es que deben existir políticas de promoción para el mercado de capitales, entendiendo promoción como “el conjunto de acciones que persigan impulsar su desarrollo sostenible pensando en el bienestar que puede transmitir al conjunto de agentes económicos que necesitan ahorrar, financiar e invertir”; entre las acciones que sugieren para promover la promoción del mercado de valores se encuentra la de “incrementar estándares de profesionalismos de sus participantes”, el cuál es uno de los fines del modelo propuesto de recomendación, brindar a personas naturales un análisis cuantitativo y cualitativo propio de un analista experto del mercado.

3. Objetivos de investigación

3.1 Objetivo general

1. Desarrollar un modelo de recomendación de acciones colombianas pertenecientes al índice COLCAP basado en análisis técnico y el sentimiento del mercado local.

3.2 Objetivos específicos

1. Caracterizar las fuentes de datos y las variables para realizar el análisis técnico y fundamental (de sentimientos) para una empresa emisora.
2. Implementar un modelo de procesamiento del lenguaje natural para el análisis de sentimientos sobre la empresa emisora.
3. Implementar un modelo de clasificación de acciones combinando análisis técnico y fundamental sobre la empresa emisora.
4. Validar un modelo de recomendación de acciones colombianas pertenecientes al índice COLCAP basado en análisis técnico y el sentimiento del mercado local.

4. Metodología a utilizar

La metodología que se utilizará es la CRISP-DM, que cuenta con las siguientes fases ilustradas en la **Figura 4** (Qaiser & Shafique, 2014):

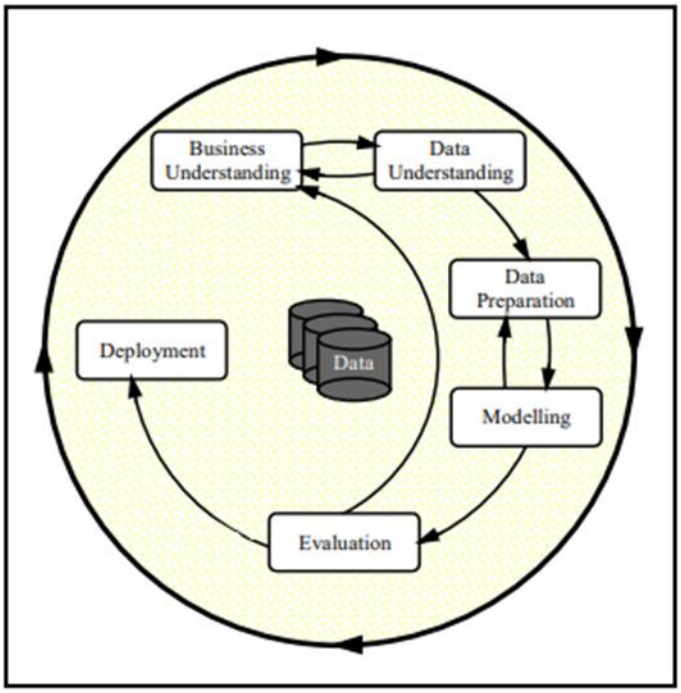

Figura 4. Metodología CRISP-DM, tomado de Wirth & Hipp, 2000.

4.1 Entendimiento del negocio

1. Una revisión de la literatura para generar un entendimiento del problema desde la perspectiva del mercado de acciones en Colombia, que desencadenará en el planteamiento del problema y su justificación.

2. Se identificarán los métodos de análisis financieros existentes y los modelos utilizados para el pronóstico, y se realizará un desarrollo teórico.

4.2 Comprensión de los datos

1. Se definirán las fuentes de datos y se obtendrán aquellos que apoyen el pronóstico tanto en la parte del análisis técnico como la parte del análisis fundamental (sentimientos).

2. Se realizará la revisión de calidad de los datos.

3. Se explorarán los datos para validar si aportan al análisis o si se deben considerar otros datos adicionales que no se tuvieron en cuenta inicialmente.

4.3 Preparación de los datos

1. Se seleccionarán los datos a utilizar en los modelos.

2. Se realizará la limpieza de los datos.

3. Se explorará la generación de variables adicionales con los datos existentes.

4. Se realizará la limpieza de datos y consolidación del dataset de entrenamiento para el modelo, con los formatos correctos y los requerimientos de entradas que tenga el modelo.

4.4 Modelado

1. Se implementan modelos para el análisis técnico y el análisis fundamental.

2. Se implementa un modelo de clasificación que combine los modelos mencionados en el numeral anterior.

4.5 Evaluación

1. Se definen métricas y se evalúa el modelo de clasificación, de acuerdo con los movimientos de precios de las acciones en un horizonte de tiempo determinado.

2. Se socializa el modelo con expertos.

4.6 Despliegue

1. Se implementa el despliegue para el beneficio de los usuarios.

2. Se produce el informe final.

5. Alcances del trabajo

- Tener un procedimiento de recolección de datos tanto para el análisis técnico como el fundamental (de sentimientos); y un procedimiento de limpieza de datos y consolidación de *dataset* para los modelos. Algunos medios de noticias pueden tener restricciones como un número limitado de noticias para usuarios gratuitos, por lo que, el procedimiento de extracción de datos debe estar validado para funcionar sin problemas de forma continua en los medios que lo permitan.
- Tener un modelo de redes neuronales para el análisis de sentimientos sobre noticias de los emisores de acciones pertenecientes al índice COLCAP.
- Tener un modelo de clasificación que combine los resultados del modelo de análisis de sentimientos con métricas relacionadas al análisis técnico.
- Tener los resultados de validación del modelo contra el comportamiento real del mercado y una comparación con otros trabajos relacionados en la literatura.

6. Productos desarrollados

En este trabajo se desarrolla un modelo de recomendación de acciones pertenecientes al índice COLCAP, cuyo objetivo es el de ser incluido en una herramienta que facilite el análisis de inversión en el mercado de acciones colombiano, especialmente para personas naturales. Los productos desarrollados en este trabajo se enuncian a continuación.

- Modelo de recomendación: Se entrena un modelo capaz de ordenar las acciones pertenecientes al índice COLCAP, de la más recomendada a la menos recomendada, para los horizontes de tiempo de 1, 10 y 20 días hábiles.
- Base de datos relacional: Se crea una base de datos relacional, donde se almacenan todos los datos de forma organizada y estructurada, de forma que la aplicación tenga mayor escalabilidad a medida que entren nuevos datos en el futuro.
- Modelo de sentimientos para noticias: Se entrena un modelo que categoriza las noticias en positivas, negativas o neutras respecto del impacto que pueden tener sobre el mercado accionario. Dentro de este modelo, se crea una herramienta de *web scraping* capaz de extraer de forma automática, a la base de datos, noticias relacionadas con emisores de acciones pertenecientes al índice COLCAP. Adicionalmente, se desarrolla una aplicación *web* para etiquetar fácilmente las noticias descargadas en la base de datos, en caso de que se quiera entrenar nuevamente el modelo con más noticias etiquetadas.
- Insumos para una herramienta de inversión: Además de la herramienta para la extracción automática de noticias, se desarrolla un procedimiento para la extracción de precios de acciones directamente de la página *web* de la Bolsa de Valores de Colombia, y otros desarrollos que permiten transformar y almacenar en la base de datos, los estados financieros y valor en libros por acción, actualizando los insumos de forma semestral. Como otros insumos, se crean procedimientos que calculan medidas relevantes para el inversionista, como volatilidad, medias móviles, indicadores de

análisis técnico e indicadores relacionados con la empresa emisora como el precio / valor en libros.

El código fuente y base de datos puede encontrarse accediendo a través del siguiente enlace.

https://drive.google.com/drive/folders/1Dvax_wG64Ttm3HdfpfKYfTSzpR-tTCLN?usp=sharing

Finalmente, se aclara que lo aquí desarrollado es académico, y que el modelo de recomendación de acciones no debe ser utilizado ciegamente como una recomendación de inversión, sino que el inversionista debe, además de ver el resultado, también analizar todas las variables que lo generaron y que aquí se dejan como insumo para una herramienta más completa.

7. Marco teórico

7.1 Inversión en acciones

En Colombia, existen múltiples plataformas para invertir en acciones, según Cajamarca (2021):

- Comisionistas de bolsa: Usualmente exigen montos mínimos que pueden ser una barrera a la entrada para personas de bajos ingresos, y cobran costos de comisión por administrar el portafolio. También existe la opción de e-trading a través de comisionistas de bolsa que permite tomar decisiones al inversionista de forma directa y por montos menores.
- Fondos de inversión colectiva: El inversionista no compra o vende acciones de forma independiente, sino que invierte su dinero en un fondo que se acomode a su perfil, donde las decisiones de inversión las toma el administrador del fondo.
- Aplicaciones móviles: En 2021 se creó la primera aplicación móvil que permite la compraventa de acciones en Colombia. Esta herramienta permite que personas naturales accedan más fácilmente al mercado accionario, ya que, además de la simpleza, permite realizar transacciones desde 15 mil pesos de forma independiente, con comisiones más bajas que las demás plataformas.

Aunque algunas de estas plataformas permitan un fácil ingreso a operar en el mercado accionario, existe una serie de riesgos financieros que los inversionistas deben reconocer y evaluar antes de comenzar a operar, entre ellos el riesgo de mercado y el riesgo de liquidez de mercado.

7.1.1 Riesgo de mercado

El riesgo de mercado de una inversión se define como el riesgo de incurrir en pérdidas debido a movimientos adversos en los precios de mercado (Adhikari, 2020). Este riesgo, es asumido por el inversionista en el momento en que entra al mercado, y para medirlo existen varias metodologías, entre ellas el valor en riesgo (VaR), definido como una medida estadística que refleja las posibles pérdidas dada una inversión, teniendo en cuenta los movimientos históricos del mercado. (Linsmeier & Pearson, 2000). De acuerdo con Jhonson (2001), esta medida puede ser calculada utilizando los rendimientos históricos de los activos financieros, como se expresa a continuación.

$$VaR = \alpha\sqrt{\sigma^2 * \Delta t}$$

Donde, α es el factor que define el área de pérdida de los retornos en una distribución normal estándar, σ^2 es la varianza de los retornos del activo y Δt es el horizonte de tiempo para el que se calcula el VaR.

7.1.2 Riesgo de liquidez de mercado

Se define como el riesgo de incurrir en pérdidas cuando un inversionista desea liquidar sus activos en el mercado y no encuentra con facilidad quién los compre, por lo que tiene que venderlos a un precio descontado (Tian, 2009). Una forma de cuantificar este riesgo sobre una acción es analizando los volúmenes históricos operados en el mercado, pues aquellas acciones con montos mayores poseen una mayor liquidez, y esto se traduce como un menor riesgo de liquidez de mercado. Para el caso del mercado accionario en Colombia, también existe un índice que refleja las acciones más líquidas operadas en la Bolsa de Valores, llamado el índice COLCAP (Bolsa de Valores de Colombia, 2021), este índice, podría servirle al inversionista como un indicativo sobre qué acciones tienen un menor riesgo de liquidez de mercado.

Según la teoría financiera, un inversionista requiere una compensación positiva por cualquier riesgo adicional que asuma en los mercados de acciones (Guo, 2002), por lo que, si el inversionista decide tomar mayores riesgos, también podrá esperar obtener a cambio una mayor rentabilidad en caso de que los movimientos del mercado lo favorezcan. Dependerá entonces de

cada individuo, según su perfil de riesgo y apoyado por cifras como el VaR y el volumen, el evaluar qué activos y en qué cantidad los desea incluir en su portafolio.

7.2 Pronósticos sobre acciones

Al invertir directamente en una empresa a través de la adquisición de acciones, el inversionista queda expuesto a las fluctuaciones diarias del mercado, que afectan el precio en que se valoran sus acciones causando posibles pérdidas o ganancias por valorización. En la literatura, se ha indagado sobre las variables que influyen en las variaciones del precio de las acciones, existen en dos grandes ramas de estudio: el análisis técnico, donde de forma cuantitativa se analiza la actividad del mercado y el comportamiento pasado para tratar de predecir el precio futuro; y el análisis fundamental, que tiene el mismo objetivo pero realiza un análisis cualitativo sobre variables relacionadas con la empresa emisora de las acciones, sus aspectos productivos, estados financieros, competencia dentro del sector y comportamiento bursátil (Martín, 2011).

Ambos tipos de análisis, tanto el técnico como el fundamental, son aplicados por los inversionistas para predecir e invertir en las acciones que podrían tener valorizaciones en el futuro; sin embargo, en la literatura, también existen teorías que tratan de desmeritar estos enfoques, como la hipótesis de los mercados eficientes, que sugiere que los precios de los activos en los mercados financieros absorben rápidamente la información nueva que se revela al mercado, generando cambios en los mismos (Fama, 1970). Esta teoría implica que de nada sirve realizar análisis técnicos o fundamentales para pronosticar los precios de los activos negociados, pues una nueva noticia terminaría cambiando completamente el pronóstico en cualquier momento; sin embargo, otros autores como Malkiel (2003) han desmentido esta afirmación, argumentando que existe un comportamiento que, en parte, se puede pronosticar por medio de análisis históricos y de mercado. A pesar de existe una componente que es susceptible de predicción, los modelos que utilizan únicamente información histórica tienen una gran debilidad; estos no tienen en cuenta la información nueva que afecta el mercado, por lo que los precios pueden predecirse con una precisión no mayor al 50%
(Sasank Pagolu, Reddy Challa, Panda, & Majhi, 2016).

Se concluye que, para predecir el precio de las acciones en los mercados financieros de una forma certera, se deben tener en cuenta variables cuantitativas y cualitativas de un análisis histórico, pero además es relevante contar con información en tiempo real, como las noticias e interacciones que puedan generar cambios en el sentimiento de los participantes del mercado.

7.3 Análisis de sentimientos

Algunos autores como Lasek & Lasek (2015) siguen el pensamiento de que la valoración de acciones más que un cálculo estrictamente matemático, está basado en psicología masiva, de forma que los inversionistas se guían más por las perspectivas del mercado que por los mismos números. Siguiendo esta teoría, analizar el sentimiento público sobre aquellos factores que influyen sobre la perspectiva de los inversionistas como las noticias financieras o comentarios de economistas es una variable importante para el pronóstico de los precios. Para realizar este tipo de análisis existen métodos de *Machine Learning* basados en análisis de sentimientos, que es el estudio computacional de las opiniones de las personas, sus actitudes y emociones respecto a una entidad (Medhat, Hassan, & Korashy, 2014).

El crecimiento de redes sociales como Twitter® ha facilitado a los investigadores una gran cantidad de datos para el planteamiento de modelos de análisis de sentimientos, y en el sector financiero ha sido de interés aplicar estos modelos para pronosticar movimientos en los mercados financieros de acciones. Ya se han realizado estudios que prueban que existen correlaciones entre índices financieros que imitan el crecimiento de la industria con el sentimiento público agregado encontrado en redes sociales como Twitter® (Sasank Pagolu, Reddy Challa, Panda, & Majhi, 2016).

No solo las interacciones en redes sociales han demostrado ser un indicador del sentimiento sobre los mercados financieros, sino que además, existe evidencia de que las noticias en medios públicos tienen una influencia sobre los mismos (Lasek & Lasek, 2015).

Según los autores Yadav y Vishwakarma (2020) en la actualidad, el análisis de sentimientos por medio de *Deep Learning* está volviéndose muy popular debido a sus buenos resultados recientes, por lo que recomiendan y explican, entre otros, los siguientes modelos:

7.3.1 Redes neuronales convolucionales

Las redes neuronales convolucionales han demostrado gran éxito en los campos de visión por computador y análisis de imagen; sin embargo, están siendo utilizadas para procesamiento del lenguaje natural a través de *Word Embedding*, una representación matricial de las frases de un documento. En la **Figura 5** se representa la arquitectura de esta red, en la cual se muestra el flujo desde la matriz de entrada hasta la salida. Dicha arquitectura contiene múltiples capas: una capa de entrada con la matriz, capas convolucionales que aplican filtros sobre la matriz, capas *pooling* que reducen la dimensionalidad y capas completamente conectadas.

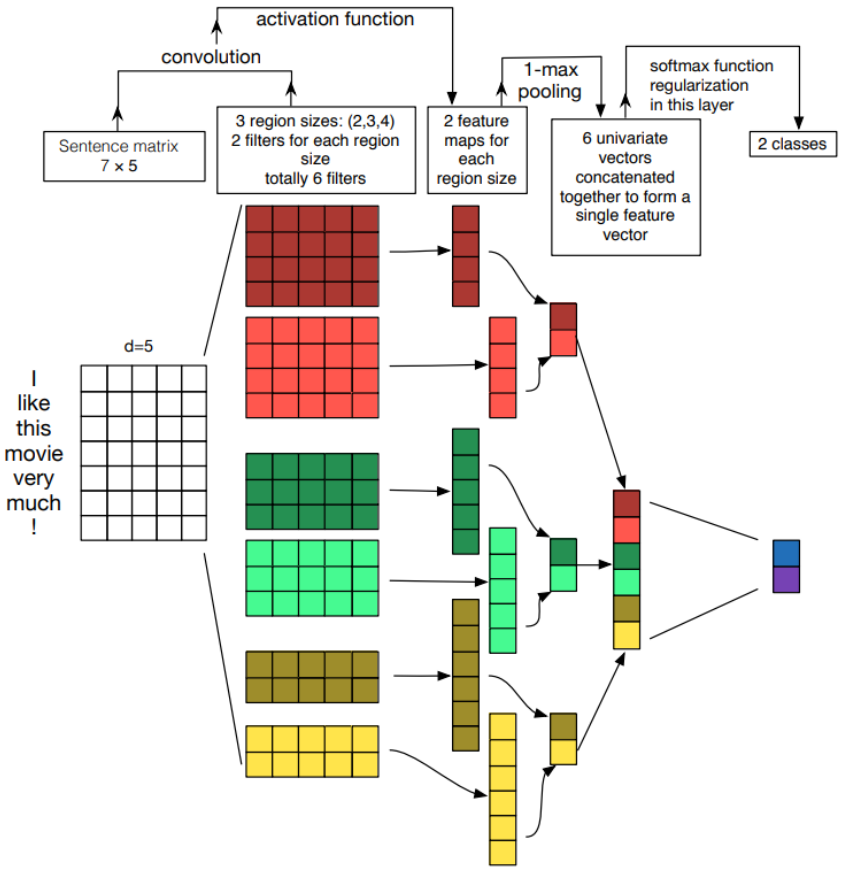

Figura 5. Arquitectura red neuronal convolucional, tomado de Zhang & Wallace, 2015.

7.3.2 Redes neuronales recurrentes

Las redes neuronales recurrentes son utilizadas para modelar datos que vienen en un orden o secuencia. Algunas aplicaciones de este tipo de redes, además del análisis de sentimientos, son la traducción de lenguajes, reconocimiento de voz y reconocimiento de actividades en video. Los dos tipos más populares de redes neuronales recurrentes son la variante *Long short-term memory* (LSTM) y *Gated recurrent units* (GRUs).

1. *Long short-term memory* (LSTM)

Esta es una de las variantes más populares de las redes neuronales recurrentes, fue creada para solucionar problemas de memoria de corto plazo.

Como se puede observar en la **Figura 6**, la arquitectura de estas redes posee un mecanismo basado en Gates (puertas), que se encargan de aprender cuál información en las celdas debe mantenerse y cuál puede olvidarse, de esa forma la red aprende y retiene información relevante para realizar predicciones.

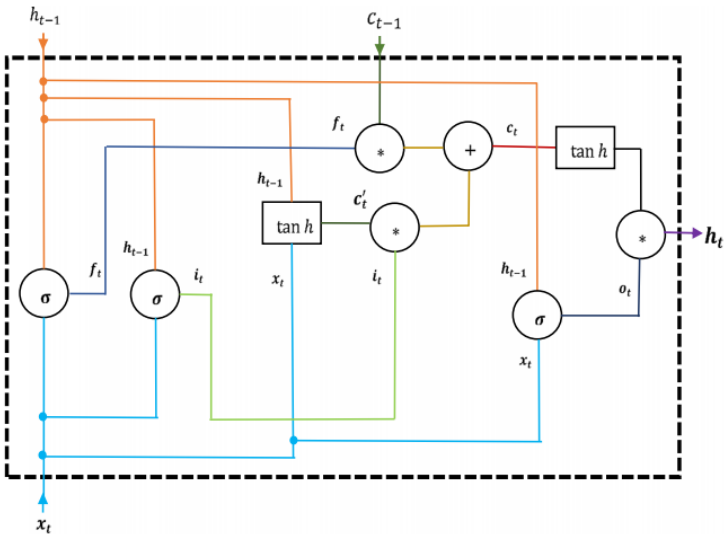

Figura 6. Arquitectura red neuronal Long short-term memory, tomado de Yadav & Vishmakarma, 2020.

2. *Gated recurrent units* (GRUs)

Son una variante de las redes LSTM, que no posee una Gate de salida. Estas redes tienen dos tipos de Gates, una de actualización y otra de reinicio. La Gate de actualización define cuánta memoria se debe conservar para el futuro, y la Gate de reinicio define cuánta información del pasado puede olvidar la red, el detalle de este flujo puede visualizarse en la **Figura 7.**

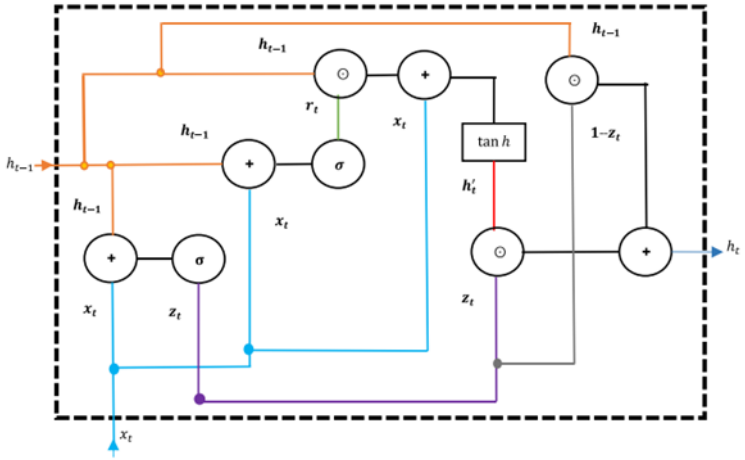

Figura 7. Arquitectura red neuronal gated recurrent units, tomado de Yadav & Vishmakarma, 2020.

7.3.3 Implementación

Estos tipos de redes neuronales pueden implementarse en Python utilizando la librería Keras que corre bajo Tensorflow, esta información con mayor detalle, junto con ejemplos de aplicación puede encontrarse en la documentación de Tensorflow (2021). Los pasos generales de implementación son los siguientes:

1. Se realiza la importación y limpieza de los datos con las librerías Numpy y Pandas.

2. Se representa el conjunto de palabras del dataset como números, para ello se utiliza la clase Tokenizer de keras:

$$tf.keras.preprocessing.text.Tokenizer(** kwargs)$$

Esta clase contiene métodos que permiten vectorizar en valores enteros las palabras del corpus.

$$fit\_on\_texts(text):$$
$$texts\_to\_sequences(text)$$

Estos métodos permiten representar el corpus en un lenguaje vectorial y numérico que podrá ingresarse a los modelos para el entrenamiento.

3. Se crea la arquitectura de capas del modelo utilizando Keras, para ello se utiliza la siguiente clase:

$$tf.keras.Sequential(layers = None, name = None)$$

Una vez se inicializa un objeto con esta clase (se denominará *model*) se comienza a agregar de forma secuencial las capas según la arquitectura deseada utilizando *add*.

a. Redes neuronales convolucionales (CNN):

$$model.add\big(Conv1D(** kwargs)\big)$$
$$model.add(MaxPooling1D(** kwargs))$$

De esta forma se pueden agregar capas de convolución y *max pooling* en una dimensión.

b. Redes neuronales recurrentes *long short-term memory*:

$$model.add(LSTM(** kwargs))$$

Gracias a Keras pueden agregarse fácilmente capas LSTM; sin embargo, se cuenta con un alto número de parámetros, que permiten personalizar dichas capas. El detalle los parámetros puede encontrarse en la documentación de Keras (TensorFlow, 2021).

c. Redes neuronales recurrentes *gated recurrent units* :

$$model.add(GRU(** kwargs))$$

Al igual que las redes neuronales LSTM, Keras incluye una capa especial para las redes GRU. El detalle de los parámetros puede encontrarse en la documentación de Keras (TensorFlow, 2021).

Además de estos casos específicos, se pueden añadir capas más genéricas completamente conectadas utilizando Dense()

$$model.add(Dense(** kwargs))$$

4. Se compila el modelo, configurando el optimizador, las funciones de pérdida y métricas a utilizar.

$$model.compile(optimizer, loss, metrics,** kwargs)$$

5. Se entrena el modelo

$$model.fit(x, y, pred ** kwargs)$$

6. Se evalúa el modelo

$$model.evaluate(x, y, ** kwargs)$$

7. Se utiliza el modelo para predecir

$$model.predict(x, ** kwargs)$$

8. Antecedentes

En la actualidad, existen de trabajos que intentan realizar este tipo de análisis en diferentes mercados del mundo.

Stock Prediction Using Twitter® Sentiment Analysis (Mittal & Goel, 2011):

Este es uno de los primeros artículos donde se utilizó análisis de sentimientos en redes sociales para la predicción de movimientos en los mercados financieros, en él ponen a prueba la teoría de que el sentimiento de los mercados está correlacionado con el sentimiento público. Para ello, utilizaron datos de precios del índice financiero *Dow Jones Industrial Average* (DJIA) entre junio de 2009 y diciembre de 2009, al igual que la información de 476 millones de *tweets* entre esas mismas fechas.

Para el conjunto de datos, hacen un filtrado de *tweets*, dejando solo aquellos que contienen palabras que expresan sentimientos, y generan un sistema de puntaje para cada palabra. Luego, prueban 4 modelos diferentes:

1. Regresión Lineal.
2. Regresión Logística.
3. *Support Vector Machine.*
4. *Self Organizing Fuzzy Neural Networks.*

Finalmente, el modelo que realizó el mejor pronóstico fue el de la red neuronal, es decir, el número 4, con una precisión de 75.56%.

Sentiment Analysis on Social Media for Stock Movement Prediction *(Nguyen, Shirai, & Velcin, 2015):*

En este trabajo el objetivo era pronosticar el precio de acciones individuales del mercado de Estados Unidos en vez de hacerlo para un índice que refleja el mercado en general, para ello utilizaron mensajes enviados en 18 foros de mensajes de *Yahoo Finance*® desde

julio de 2012 hasta julio de 2013, cada uno de ellos dedicado a una acción diferente. Como variable respuesta utilizaron los precios de las acciones para ese mismo periodo de tiempo.

En este trabajo probaron 5 modelos que combinan análisis históricos y análisis de sentimientos, todos utilizan como método un *Support Vector Machine* pero con diferentes parámetros de entrada, definidos así:

1. *Price Only:* solamente se utilizan precios históricos para predecir los nuevos precios.
2. *Human Sentiment:* Un 15.6% de los datos que utilizaron ya habían sido etiquetados con un sentimiento de compra por el usuario, así que para este modelo utilizaron esos sentimientos humanos junto con precios históricos como entrada.
3. *Sentiment classification:* Para el 84.4% de los datos que no estaban etiquetados, utilizaron *Natural Lenguage Processing* para extraer los sentimientos de los mensajes. Para ello utilizaron la librería de *Standford* CoreNLP para lematizar las palabras, y luego utilizaron un modelo de clasificación *Support Vector Machine*. Adicionalmente, utilizaron los precios históricos.
4. *LDA-Based method*: Utilizan los precios históricos y un modelo de *Latent Dirichlet Allocation,* que es un modelo probabilístico para el corpus. La idea de este modelo es que cada documento es una mezcla de temas, y cada tema es caracterizado por una distribución de palabras.
5. *JST-Based method*: Se considera cada mensaje como una mezcla de temas y sentimientos ocultos, que se extraen como tuplas por medio de un modelo *Joined Sentiment Topic* que las extrae de forma simultánea. Adicionalmente se utilizan los precios históricos.
6. *Aspect-Based sentiment*: Primero extraen una lista de los temas que tengan una ocurrencia mayor a 10 veces, y luego extraen el sentimiento para cada uno de esos temas, las palabras que se consideran opiniones las extraen de SentiWordNet, que es un recurso para minería de opiniones. Adicionalmente se utilizan los precios históricos.

Finalmente, de estos modelos el que logra una mayor precisión es el número 6, con 54.4% en promedio para las acciones analizadas.

Incorporating Stock Prices and News Sentiments for Stock Market Prediction: A case of Hong Kong *(Li, Wu, & Wang, 2020):*

En este artículo utilizan una red neuronal artificial para predecir los precios de acciones en el mercado de Hong Kong, para ello generaron 2 entradas para la red:

1. La primera representa el análisis técnico hallado con indicadores que normalmente utilizan los inversores, que representan con un dato numérico

2. La segunda es una representación de sentimientos en vectores, donde utilizan múltiples diccionarios de sentimientos aplicados a artículos de noticias.

Estas entradas se ingresan a una red neuronal *Long Short Term Memory* (LSTM) de dos capas, que se encarga de aprender la información secuencial presentada en las series de estos vectores, y luego se genera una predicción de precios a través de una función de activación Softmax.

Como conclusiones de este trabajo, se encontró que, utilizando 5 años de precios y 4 diferentes diccionarios de sentimientos, el modelo LSTM supera a modelos que tan sólo usan información técnica o análisis de sentimientos de noticias, utilizando como medidas la precisión (*accuracy*) y el F1 *score*.

Incorporating Expert-Based Investment Opinion Signals in Stock Prediction: A Deep Learning Framework *(Wang, Wang, & Li, 2020):*

En este trabajo va más allá de un análisis de sentimientos simple sobre opiniones en acciones del mercado de China, ya que además se desarrolló un modelo capaz de identificar detalles relevantes en cada opinión, como por ejemplo si la persona que la escribió es un experto, o si el comentario que hizo habla sobre el pasado, el presente o el futuro; pues todos estos detalles son relevantes para generar pronósticos de precios.

El modelo que proponen consta de 3 módulos:

1. *Stance detection*: Para este módulo desarrollaron un modelo que llamaron *Multi-view Fusion Network* (MFN), su función es identificar si una opinión es *bullish* (alcista), *bearish* (bajista) o neutral. Este es un modelo de redes neuronales convolucionales que utiliza *encoders*.
2. *Stance aggregation*: Este algoritmo se encarga de analizar los usuarios que escriben múltiples opiniones, y comparando con los datos reales de 90 días miden su capacidad de predicción. De esta manera extraen y agregan solo las opiniones que generan mayor valor para la predicción.
3. *Stock forecasting*: utilizan una serie de tiempo con las características extraídas de las opiniones y la ingresan a una red neuronal recurrente *Gated Recurrent Unit* (GRU) que aprende secuencialmente en orden cronológico.

Para probar los modelos seleccionaron y etiquetaron manualmente 20 mil opiniones sobre acciones del mercado chino. Para evaluar los resultados utilizaron la métrica F1 *Score*, y

compararon su modelo con otros métodos populares en la literatura, como lo son los modelos *Transformer*. Finalmente, el modelo MFN obtuvo un F1 *score* de 75.61, superior a las otras metodologías implementadas

9. Entendimiento de los datos

9.1 Descripción de los datos

Siguiendo los lineamientos definidos en el alcance del presente trabajo, se acota el universo de acciones a clasificar dentro de únicamente aquellas pertenecientes al índice COLCAP. El Banco de la República de Colombia (2021) define el COLCAP como aquel índice que refleja las variaciones de precio de las acciones más líquidas operadas en la Bolsa de Valores de Colombia, donde cada acción tiene un porcentaje de participación dentro de dicho índice, y este además debe estar compuesto por al menos 20 acciones de emisores diferentes. A la fecha de realización de este trabajo, se presentan las acciones que componen el índice COLCAP en la **Tabla 2.**

Nemotécnico	**Razón Social Emisor**
BCOLOMBIA	BANCOLOMBIA S.A.
BOGOTA	BANCO DE BOGOTA S.A.
BVC	BOLSA DE VALORES DE COLOMBIA S.A.
CELSIA	CELSIA S.A
CEMARGOS	CEMENTOS ARGOS S.A.
CNEC	CANACOL ENERGY LTD
CORFICOLCF	CORPORACION FINANCIERA COLOMBIANA S.A.
ECOPETROL	ECOPETROL S.A.
ETB	EMPRESA DE TELECOMUNICACIONES DE BOGOTA S.A. E.S.P.
GEB	GRUPO ENERGIA BOGOTA S.A. E.S.P.
GRUBOLIVAR	GRUPO BOLIVAR S.A.
GRUPOARGOS	GRUPO ARGOS S.A.
GRUPOSURA	GRUPO INVERSIONES SURAMERICANA
ISA	INTERCONEXION ELECTRICA S.A. E.S.P.
MINEROS	MINEROS S.A.
NUTRESA	GRUPO NUTRESA S.A
PFAVAL	GRUPO AVAL ACCIONES Y VALORES S.A.
PFBCOLOM	BANCOLOMBIA S.A.
PFCEMARGOS	CEMENTOS ARGOS S.A.
PFCORFICOL	CORPORACION FINANCIERA COLOMBIANA S.A.
PFDAVVNDA	BANCO DAVIVIENDA S.A
PFGRUPOARG	GRUPO ARGOS S.A.
PFGRUPSURA	GRUPO INVERSIONES SURAMERICANA
PROMIGAS	PROMIGAS S.A. E.S.P.
TERPEL	ORGANIZACION TERPEL S.A

Tabla 2. Acciones y Emisores pertenecientes al índice COLCAP, tomado de Bolsa de Valores de Colombia, 2021.

Siguiendo la definición del índice COLCAP, estas 24 acciones son las más negociadas en Colombia, por lo que, para este trabajo, se realizará una recolección de datos únicamente para estas especies y su respectivo emisor.

Los datos que serán utilizados para alimentar el modelo de recomendación de acciones están clasificados en dos grupos: los datos de precios para el análisis técnico, y los datos de noticias para el análisis fundamental.

9.1.1 Estructura de los datos

Con el fin de administrar los datos de forma eficiente y ordenada, se utilizará una base de datos relacional SQLite, creada a través del *Object Relational Mapping* (ORM) de Python SQLAlchemy. Dicha base de datos estará conectada con todos los scripts de recolección de datos y de los modelos para recomendación de acciones. A continuación, en la **Figura 8** se presenta el diagrama entidad – relación de la base de datos.

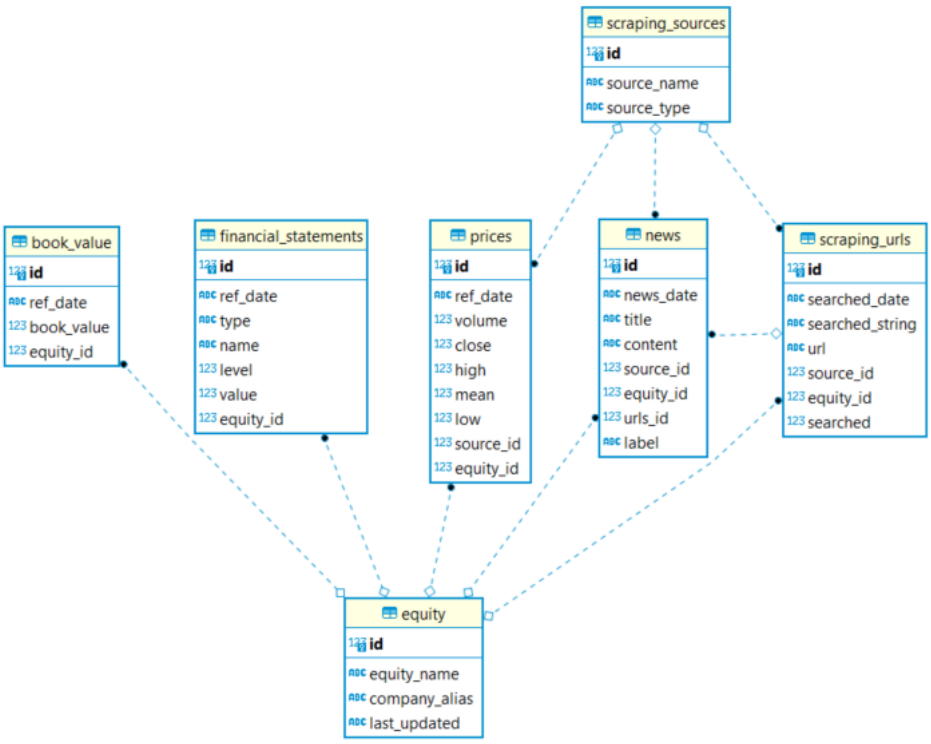

Figura 8. Diagrama entidad relación de la base de datos SQLite, elaboración propia.

La base de datos cuenta con 4 *dataset* almacenados en las tablas *news*, *prices, financial_statements y book_value*, además de otras 3 tablas auxiliares para la recolección de

datos, que contienen la lista de acciones del COLCAP, la lista de fuentes de datos y una lista de URLs de noticias sobre acciones.

9.1.2 *Dataset* de datos diarios de precios de valoración y volumen de negociación

Se cuenta con un *dataset* de precios con 51861 registros, que será utilizado como *input* del componente de análisis técnico para el modelo de recomendación de acciones. Se describen los campos del *dataset* en la **Tabla 3**.

Campo	Descripción
ref_date	Día de referencia
volume	Volumen negociado en el día
Close	Precio de cierre del día, valorado por el proveedor de precios PRECIA S.A
High	Precio más alto del día
Mean	Precio promedio del día
Low	Precio más bajo del día
source_id	id de la fuente donde se obtuvieron los datos
equity_id	id de la acción de referencia

Tabla 3. Campos del dataset de precios y volumen, elaboración propia.

Este *dataset* se obtuvo utilizando técnicas de *web scraping* con Python en la página web de la Bolsa de Valores de Colombia.

Dentro de la obtención de datos de precios se encontraron algunas dificultades:

- Restricción de consulta de máximo 6 meses de historia en la web de la Bolsa de Valores de Colombia, por lo que se tuvo que separar cada consulta en múltiples *request* de 6 meses cada uno.
- El nemotécnico de la acción del Grupo de Energía de Bogotá en el año 2019 cambió su nombre, por lo que se tuvieron que realizar los *request* previos a 2019 con el nombre “EEB” y unificarlo en la base de datos con el nemotécnico actual “GEB”.

9.1.3 *Dataset* de noticias de medios Colombianos relacionadas con los emisores de acciones

Se obtuvo un *dataset* con 4112 registros de noticias de 2 importantes medios de noticias Colombianos: la revista Portafolio y el periódico La República. Estos datos serán utilizados para

la componente de análisis fundamental del modelo de recomendación de acciones. Se describen los campos del *dataset* en la **Tabla 4**.

Campo	Descripción
news_date	Día de la noticia
title	Título de la noticia
content	Corpus de la noticia
source_id	id de la fuente donde se obtuvieron los datos
equity_id	id de la acción de referencia
urls_id	id de la URL de donde se obtuvo la noticia

Tabla 4. Campos del dataset de noticias, elaboración propia.

Para obtener este *dataset* se utilizaron técnicas de *web scraping* con Python en las páginas web de Portafolio y La República. Inicialmente, el *web scraping* se utilizó para encontrar una lista de URL de las noticias relacionadas con cada una de las empresas emisoras de acciones que pertenecen al COLCAP, esta lista de noticias se almacenó en la tabla *scraping_urls* de la base de datos, y se obtuvo a través de la utilización de motores de búsqueda, tanto de las mismas *web* de noticias como de Google®. Como criterio de búsqueda para las URL se utilizó el nombre de la empresa emisora de cada acción del COLCAP. En una fase posterior, se utilizó la lista de URL en otro *script* de Python para ir a cada noticia y extraer los corpus, títulos y fechas de cada una.

Al realizar ambos procesos de *web scraping* aparecieron ciertas dificultades:

- Cantidad limitada de *requests* por unidad de tiempo y bloqueos temporales de dirección IP, lo que obligó a generar pausas aleatorias de hasta 1 minuto en los scripts.
- Algunas noticias dentro del *corpus* contenían publicidad hacia otras noticias, por lo cual se tuvo que realizar una limpieza a través de clases CSS o expresiones regulares dependiendo del caso.

9.1.4 *Dataset* de estados financieros y valor en libros acciones

Con el objetivo de hallar indicadores sobre las empresas emisoras de acciones pertenecientes al índice COLCAP, se obtuvo un dataset de 12639 registros, con cada una de las cuentas contables del balance general y estado de resultados; además, de otro dataset con la información del valor en libros por acción.

Campo	Descripción
ref_date	Fecha de corte del informe financiero
type	Balance general o estado de resultados
name	Nombre de la cuenta contable
level	Nivel de jerarquía de la cuenta contable
value	id de la acción de referencia
Equity_id	id de la acción de referencia

Tabla 5. Campos del dataset de estados financieros, elaboración propia.

La información de balance general y estado de resultados fue obtenida a través del servicio de información financiera EMIS® en archivos con formato ".xlsx". Luego, estos archivos, son leídos por un desarrollo que los transforma en el formato de la **Tabla 6** y los almacena en la base de datos. Esta información a diferencia de las noticias y precios, no se descarga de manera automática, por lo que debe actualizarse manualmente cada 6 meses con la publicación de estados financieros de fin de año o semestrales.

Campo	Descripción
ref_date	Fecha de corte
book_value	Balance general o estado de resultados
equity_id	id de la acción de referencia

Tabla 6. Campos del dataset de estados financieros, elaboración propia.

Esta información contiene implícitamente el número de acciones en circulación, el cual es necesario para calcular indicadores importantes para el análisis fundamental. Estos datos se obtienen directamente de la página *web* de la Bolsa de Valores de Colombia, dentro del informe mensual de renta variable. Mensualmente se debe descargar el archivo, guardarlo en la carpeta correspondiente y ejecutar el desarrollo que lee, transforma y almacena la información en la base de datos.

9.2 Análisis exploratorio de los datos

9.2.1 Volumen diario de negociación

Como se mencionó previamente, en Colombia existen una serie de factores que han incentivado a un gran crecimiento de la renta fija, pero a un decaimiento en el mercado accionario. Esta situación se evidencia observando los volúmenes negociados de las acciones más líquidas del país, presentadas en la **Tabla 7**.

Acción	25%	50%	75%	Media	Desviación
ECOPETROL	11,730	18,803	29,329	23,273	17,096
PFBCOLOM	10,144	16,202	24,162	20,301	17,697
EXITO	491	3,024	7,575	10,739	194,650
GRUPOSURA	3,302	5,679	9,460	8,940	18,758
BCOLOMBIA	2,426	4,885	9,369	8,665	18,437
PFAVAL	2,688	4,267	7,294	6,559	12,541
CEMARGOS	1,450	2,710	4,991	5,798	21,973
ISA	2,015	3,753	6,653	5,537	7,158
GRUPOARGOS	1,776	3,051	5,220	5,258	12,844
PFDAVVNDA	1,529	2,986	5,939	5,010	8,423
PFGRUPSURA	1,007	2,146	3,985	4,043	9,313
NUTRESA	1,140	2,201	4,033	3,410	4,979
GEB	575	1,479	3,259	3,400	17,431
CORFICOLCF	872	1,676	3,389	2,995	7,577
PFGRUPOARG	255	651	1,529	1,924	7,372
PFCEMARGOS	155	481	1,461	1,517	4,151
CNEC	300	765	1,519	1,338	2,434
CELSIA	293	626	1,336	1,168	1,683
BOGOTA	196	448	925	867	2,848
PROMIGAS	0	19	125	418	3,051
ETB	33	105	276	348	1,287
TERPEL	0	13	102	274	1,559
MINEROS	18	86	251	234	662
PFCORFICOL	0	9	99	180	682

Tabla 7. Análisis de Volumen diario negociado en acciones pertenecientes al COLCAP, datos desde 2015, cifras en millones, elaboración propia.

De los datos presentados, se evidencian varios problemas; primero, una gran concentración del volumen en las primeras especies, y segundo, unos muy altos niveles de desviación estándar, debido a que es muy irregular la negociación, algunos días hay altos volúmenes de negociación, mientras que en otros pueden llegar a ser incluso cero. En el caso de la acción EXITO, la alta

desviación estándar se origina por un evento específico, la oferta pública de adquisición que tuvo en 2019, donde se transaron más de 7 billones de pesos.

Aquellas acciones con niveles casi nulos de liquidez deben tratarse con especial cuidado, ya que, al no transarse, los precios se quedan estáticos y pueden dar un falso indicio de baja volatilidad.

9.2.2 Precios diarios de valoración

Para el análisis de precios, uno de los componentes más relevantes para el inversionista se trata de la volatilidad de los activos, pues esto define qué tanto puede ganar o perder en un horizonte de tiempo, y finalmente define qué tanto riesgo está tomando. Como menciona Miskolczi (2017) en la literatura, la medición de la volatilidad no se realiza sobre los precios per se, sino que se realiza sobre los rendimientos diarios, que no son más que el cambio diario en el precio del activo; usualmente se elige utilizar retornos logarítmicos en vez de retornos simples, debido a las ventajas que tienen cuando se analizan múltiples periodos de tiempo.

$$R^*_{[0,T]} = ln\left(\frac{P_T}{P_0}\right)$$

$$donde:$$

$$R^*_{[0,T]}: Retorno\ logarítmico\ de\ un\ activo\ para\ un\ periodo$$

$$P_T: Precio\ del\ activo\ en\ el\ periodo\ T$$

$$P_0: Precio\ del\ activo\ en\ el\ periodo\ inicial$$

En la **Tabla 8** se muestran métricas sobre los retornos logarítmicos de las series de precios de las acciones pertenecientes al índice COLCAP.

equity_name	mean	std	min	max
CNEC	0.04%	2.41%	-13.67%	14.76%
ECOPETROL	0.02%	2.37%	-22.40%	12.26%
GRUPOARGOS	-0.04%	2.33%	-22.45%	28.72%
PFGRUPOARG	-0.06%	2.15%	-26.53%	17.26%
CEMARGOS	-0.04%	2.14%	-27.44%	20.07%
BCOLOMBIA	0.00%	2.05%	-19.64%	21.75%
ETB	-0.06%	2.01%	-20.74%	17.08%
EXITO	-0.05%	2.00%	-19.85%	13.63%
PFCEMARGOS	-0.06%	2.00%	-19.89%	16.25%
ISA	0.06%	1.91%	-27.56%	13.87%
PFGRUPSURA	-0.05%	1.88%	-25.82%	19.49%
TERPEL	-0.06%	1.85%	-17.44%	15.55%
GRUPOSURA	-0.05%	1.84%	-17.66%	22.31%
PFBCOLOM	0.00%	1.82%	-19.66%	14.00%
MINEROS	0.03%	1.80%	-10.54%	17.05%
PROMIGAS	0.01%	1.73%	-10.54%	13.82%
PFDAVVNDA	0.00%	1.68%	-15.69%	12.45%
CORFICOLCF	-0.02%	1.63%	-22.31%	11.46%
PFAVAL	-0.01%	1.61%	-13.71%	20.59%
CELSIA	-0.02%	1.61%	-15.25%	10.36%
PFCORFICOL	-0.03%	1.60%	-10.38%	9.36%
BOGOTA	0.00%	1.54%	-14.60%	10.73%
GEB	0.03%	1.39%	-14.27%	14.03%
NUTRESA	-0.02%	1.27%	-10.54%	7.22%

Tabla 8. Análisis de rendimientos logarítmicos sobre acciones pertenecientes al COLCAP, datos desde 2015, elaboración propia.

Se observa que los rendimientos promedio logarítmicos se aproximan a cero, lo cual es consistente con la teoría. Adicionalmente, se hace notar que, aunque el promedio se acerca a cero y las desviaciones estándar no superan en ningún caso el 3%, los rendimientos máximos y mínimos llegan a valores muy altos, incluso superiores al 20% en algunos casos. Este comportamiento es explicado por las grandes caídas y rebotes que han ocurrido en la historia debido a crisis económicas, como se muestra en la **Figura 9.**

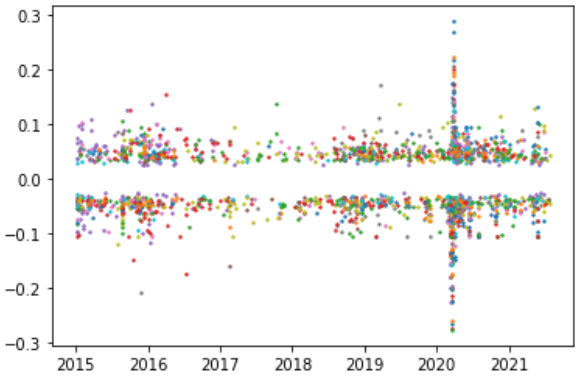

Figura 9. Rendimientos logarítmicos que superan dos desviaciones estándar, para todas las acciones pertenecientes al índice COLCAP, elaboración propia.

Gráficamente se comprueba que los altos valores encontrados en los máximos y mínimos de los rendimientos logarítmicos se explican principalmente por el evento ocurrido en marzo de 2020 a causa del COVID-19. También se puede observar como la alta volatilidad afecta a todo el mercado casi de manera homogénea en diferentes periodos de tiempo.

9.2.3 Noticias sobre los emisores de acciones pertenecientes al índice COLCAP

Después de analizar el comportamiento de los precios de valoración de las acciones, es claro que existen momentos de estrés de mercado, por lo que, entra en relevancia conocer la distribución temporal de los datos de noticias a utilizar en el modelo de recomendación.

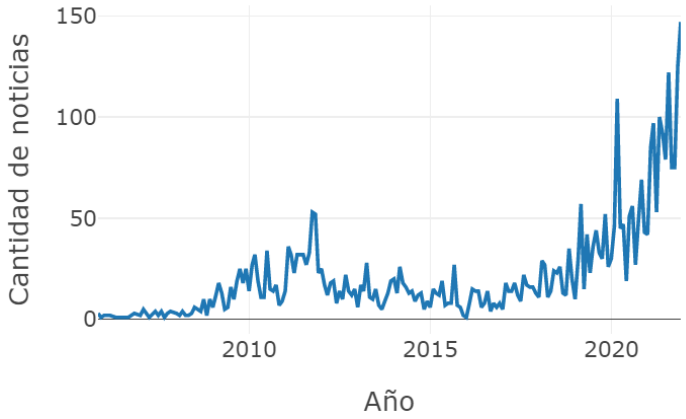

Figura 10. Cantidad de noticias obtenidas en total por cada mes, para los emisores de acciones pertenecientes al índice COLCAP, elaboración propia.

El mes con una mayor cantidad de noticias relacionadas con los emisores de acciones pertenecientes al índice COLCAP, es el de diciembre de 2021, seguido por el mes de marzo de 2020, que coincide con la crisis observada en el análisis de precios. También es claro que hay una mayor cantidad de noticias en los años más recientes, lo cual es positivo, pues el interés principal es modelar correctamente la dinámica del mercado en la época actual.

company_alias	News
banco de bogotá	186
bancolombia	273
canacol	154
celsia	143
cementos argos	345
corficolombiana	161
davivienda	207
ecopetrol	445
etb	307
exito	111
grupo aval	178
grupo de energía de bogotá	101
isa	246
mineros s.a.	59
nutresa	261
promigas	133
sura	562
terpel	240

Figura 9. Cantidad de noticias obtenidas en total por cada emisor de acciones pertenecientes al índice COLCAP, elaboración propia.

Otra dinámica de interés dentro de los datos, es que pueden existir variables que generen un desequilibrio en el número de noticias publicadas para cada empresa, ya que no todas tienen la misma dinámica con la prensa, ni el mismo tamaño o proyectos relevantes. En la **Tabla 10** se muestran las palabras más utilizadas en las noticias de cada emisor de acciones, donde se nota una prominencia de la palabra “millones” casi para todas las empresas, lo cual sugiere que la gran mayoría de estas noticias habla principalmente de temas monetarios, de ingresos, pérdidas, inversiones, crecimiento, entre otros; por este motivo, empresas involucradas en movimientos grandes de dinero, podrían tener una probabilidad mayor de captar la atención de la prensa.

Empresa	Top 10 Palabras
banco de bogotá	entidad, clientes, digital, digitales, país, través, mejor, crédito, colombia
bancolombia	año, millones, clientes, entidad, billones, personas, colombia, país, empresas
canacol	gas, producción, millones, colombia, año, exploración, barriles, compañía, natural
celsia	energía, millones, generación, año, compañía, colombia, solar, mw, operación
cementos argos	millones, cementos, ciento, cemento, año, compañía, pesos, colombia, empresa
corficolombiana	año, millones, odebrecht, aval, crecimiento, trimestre, colombia, corporación, inversiones
davivienda	millones, año, colombia, pesos, clientes, entidad, billones, mercado, parte
ecopetrol	millones, petrolera, producción, año, compañía, empresa, billones, us, país
etb	empresa, millones, pesos, compañía, año, telecomunicaciones, ciento, proceso, servicios
exito	colombia, tamales, empresa, usuarios, años, cuenta, personas, negocio, tener
grupo aval	aval, millones, pesos, año, billones, bancos, sarmiento, ciento, colombia
grupo de energía de bogotá	geb, energía, millones, gas, operación, además, presidente, transmisión, sistema
isa	energía, millones, año, compañía, colombia, empresa, transmisión, ciento, billones
mineros s.a.	millones, pesos, mineros, empresa, ciento, año, oro, compañía, producción
nutresa	millones, año, ventas, compañía, pesos, colombia, ciento, billones, crecimiento
promigas	gas, millones, natural, ciento, colombia, país, compañía, empresa, energía
sura	millones, seguros, compañía, colombia, año, ciento, inversiones, pesos, suramericana
terpel	colombia, combustibles, mercado, millones, estaciones, ciento, año, compañía, país

Tabla 10. Top 10 palabras más utilizadas en noticias por cada emisor de acciones pertenecientes al índice COLCAP, elaboración propia.

Para el presente trabajo, se consideraron las noticias de todos los emisores del índice COLCAP de los últimos 5 años, las cuales se etiquetaron para el modelo de recomendación en 4 categorías:

- Noticia positiva
- Noticia neutra
- Noticia negativa
- Omitir Noticia

Dentro de los datos utilizados, se encontró que la mayor proporción de noticias corresponde a aquellas que pueden influir de forma positiva en el precio de la acción, seguida por noticias neutras, que no tienen un efecto significativo en el precio de la acción, y finalmente, una minoría

de noticias negativas, que generan tendencias a la baja en el precio de la acción. Dentro de los datos también se encontraron algunas noticias que por su contenido no eran relevantes para el mercado de acciones o el emisor particular, por lo que fueron omitidas.

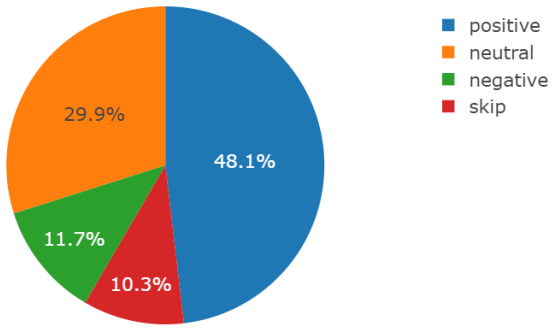

Figura 11. Proporción de etiquetas para las noticias de los últimos 5 años de emisores de acciones pertenecientes al índice COLCAP, elaboración propia.

En la **Figura 11** se presenta el total de noticias positivas, negativas y neutrales en el tiempo, en donde se podrían resaltar dos patrones importantes. El primero, es que hay una alta participación de noticias negativas durante marzo del 2020, cuando la economía y el mercado accionario se contrajeron a causa del COVID 19; y como segundo patrón, finalizando el año 2021 hay un crecimiento importante en el número de noticias positivas, lo cual podría estar relacionado con la acelerada recuperación económica del país tras el primero año de pandemia, medida por un valor de crecimiento del PIB que se estima superior al 9%, la cifra más alta en 115 años (Salazar, 2021).

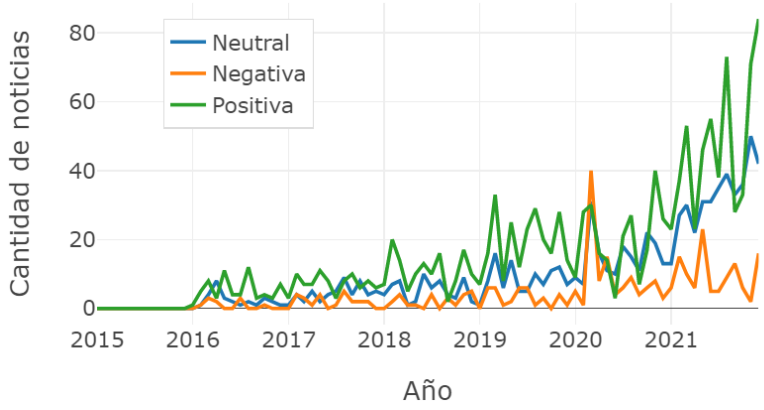

Figura 12. Cantidad de noticias positivas, negativas y neutras de los últimos 5 años de emisores de acciones pertenecientes al índice COLCAP, elaboración propia.

9.2.4 Estados financieros de los emisores de acciones pertenecientes al índice COLCAP

Uno de los principales indicadores de la rentabilidad de una empresa es el margen EBITDA, según LUKIC (2018), este indicador es clave para analistas que buscan evaluar el rendimiento de una empresa en el largo plazo, y es calculado como (Ingresos – gastos (excluyendo intereses, impuestos, depreciación y amortización) / ingresos totales). En la **Tabla 11** se presentan los márgenes EBITDA para cada uno de los emisores de acciones pertenecientes al índice COLCAP.

Acción	EBITDA	Margen EBITDA
ISA	2,930,900	56.38%
CNEC	73,190	50.69%
ECOPETROL	16,767,425	46.10%
GEB	1,174,965	44.75%
BCOLOMBIA	2,702,327	42.41%
PFBCOLOM	2,702,327	42.41%
CORFICOLCF	2,153,623	42.05%
PFCORFICOL	2,153,623	42.05%
BOGOTA	2,473,422	39.30%
PROMIGAS	862,805	35.41%
CELSIA	630,070	32.74%
PFAVAL	4,870,677	31.35%
GRUPOARGOS	2,254,069	28.16%
PFGRUPOARG	2,254,069	28.16%
ETB	193,549	27.40%
CEMARGOS	1,141,219	22.76%
PFCEMARGOS	1,141,219	22.76%
MINEROS	170,299	18.52%
PFDAVVNDA	881,506	17.17%
NUTRESA	749,186	12.97%
PFGRUPSURA	1,180,225	11.59%
GRUPOSURA	1,180,225	11.59%
EXITO	582,198	7.73%
TERPEL	493	4.93%

Tabla 11. Indicadores EBITDA y margen EBITDA correspondiente al emisor de cada acción, elaboración propia.

Al analizar los datos, se pueden destacar las empresas del sector energía, que poseen todos los primeros puestos con los mayores márgenes, seguido por el sector financiero, especialmente bancos e intermediarios. Casi al final de la tabla, se observa el Grupo Sura, que, a pesar de ser parte del sector financiero, pertenece a un negocio diferente, los seguros. Un punto importante, es que más allá de comparar los márgenes entre las diferentes empresas y sectores, el modelo de recomendación hará un seguimiento individual de la evolución de este indicador en cada empresa, de forma que, sirva como un *benchmark* para cada la acción.

Otro indicador importante que observan los inversionistas del mercado accionario, tal como lo plantea Drakopoulou (2015), es el indicador precio / valor en libros, que es utilizado para medir qué tanto valor le da el mercado a la empresa emisora. Aquellas empresas con valores pequeños de este indicador son más atractivas para los inversionistas. El cálculo de este indicador se realiza como se muestra a continuación.

$$Precio\ /\ Valor\ en\ libros\ = \frac{Precio\ de\ la\ acción}{Valor\ en\ libros\ por\ acción}$$

$$donde: Valor\ en\ libros\ por\ acción = \frac{Activo - Pasivo}{Acciones\ en\ circulación}$$

Acción	Precio/valor contable
CNEC	2.83
ISA	1.78
ECOPETROL	1.78
GEB	1.67
PROMIGAS	1.64
NUTRESA	1.23
PFAVAL	1.06
PFDAVVNDA	1.00
CORFICOLCF	0.97
BOGOTA	0.97
EXITO	0.92
PFBCOLOM	0.91
BCOLOMBIA	0.90
CELSIA	0.87
MINEROS	0.83
CEMARGOS	0.83
PFCORFICOL	0.77
PFCEMARGOS	0.58

TERPEL	0.55
GRUPOARGOS	0.52
GRUPOSURA	0.44
PFGRUPOARG	0.42
PFGRUPSURA	0.39
ETB	0.36

Tabla 12. Indicador precio/valor en libros correspondiente al emisor de cada acción, elaboración propia.

Para el indicador de precio/valor en libros calculado, se ve como, nuevamente las empresas del sector energético están a la cabeza, esto quiere decir que el mercado valora a estas empresas en particular por encima de las demás. Contrastando estas cifras, se deduce que podrían existir grandes oportunidades de inversión en el mercado de acciones colombiano, ya que hay muchas empresas valoradas por debajo de su valor en libros. Adicionalmente, como mencionan analistas del mercado, podría utilizarse el periodo entre 2005 y 2006 como punto de comparación para entender qué tan bajas son estas cifras; en este periodo en específico, acciones de bancos colombianos llegaron a negociarse hasta 3 veces su valor en libros en la Bolsa de Valores (Portafolio, 2021)

10. Preparación de los datos

10.1 Etiquetado de los datos de noticias

Antes de servir como insumo al modelo final de clasificación de acciones, las noticias sobre emisores de acciones del índice COLCAP primero deben ser clasificadas como positivas, negativas o neutras; dependiendo del impacto que puedan generar sobre el sentimiento de los inversionistas que participan en el mercado de acciones. Para lograr este objetivo, se entrenó un modelo LSTM de clasificación, el cual necesitó como insumo que todas las noticias obtenidas a través de *web scraping* fueran primero etiquetadas, por lo que se procedió a crear una aplicación *web* sencilla, que permitiera etiquetar de manera ágil la gran cantidad de noticias.

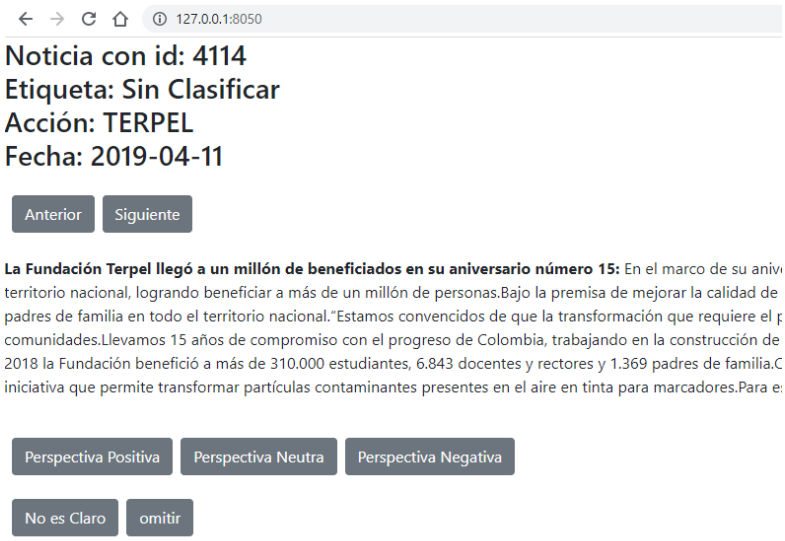

Figura 13. Herramienta de etiquetado de datos diseñada con la librería Dash de Python, elaboración propia.

Dentro de este proceso, al revisar las noticias una por una, se encontraron varias particularidades:

- **Noticias sin relación con el emisor**: A pesar de que en el proceso de *web scraping* se filtraron solo aquellas noticias donde el nombre de la empresa emisora de acciones se encontrara en el título, existen noticias que no tienen que ver con dicha empresa o que no tenían un corpus, pues su contenido era multimedia. Para estos casos se creó el botón "omitir" en la herramienta *web*, que les asigna la etiqueta "skip" a estos datos, lo que permite filtrar estas noticias fácilmente dentro de la base de datos.
- **Noticias con un alto nivel de ambigüedad:** El objetivo del etiquetado de datos es el de asignar un sentimiento a cada noticia, que represente la influencia que tiene dicha noticia para generar movimientos al alza o a la baja dentro del mercado accionario. Al revisar los datos a detalle, se encontró que, aunque muchas noticias podrían ser positivas o negativas desde diferentes perspectivas, esto no necesariamente significa que representen un sentimiento positivo o negativo en el mercado, especialmente aquellas noticias que no tienen mucho impacto sobre los flujos futuros de la empresa. En estos casos se presenta un alto nivel de subjetividad, pues depende del encargado del etiquetado el interpretar si la noticia influye en el mercado y cómo lo hace, ya que, aunque una noticia podría ser positiva para un individuo, se le debe asignar una etiqueta neutra si no tiene la capacidad de generar un movimiento en precios de mercado. Para aquellos casos donde el nivel de ambigüedad de las noticias era alto, se les asignó una etiqueta de "no es claro", lo cual permitió identificarlas en la base de datos, para ser consultadas con un experto.
- **Dificultad en los criterios de búsqueda:** Se encontró dificultad al realizar la búsqueda automática de noticias para algunos emisores, como es el caso de Éxito®, ya que, a pesar de filtrar solo aquellas donde el nombre del emisor se encuentre en el título, en estos casos el resultado contiene un gran número de noticias no relacionadas con la empresa. Para estos casos en particular, se realizaron cambios manuales en los criterios de búsqueda y se filtraron las noticias no relacionadas.
- **Cantidad de noticias según el emisor:** Los diferentes emisores de acciones tienen una visibilidad diferente en medios de comunicación, por lo que se encontró una diferencia importante en la cantidad de noticias encontrada por cada uno. Algunas de los emisores son grandes empresas, que constantemente están generando negocios en el territorio

nacional o en el exterior, por lo cual tienen una gran exposición en medios de comunicación, mientras que otras solo son el centro de atención en casos particulares.

10.2 Limpieza de los datos

10.2.1 Precios de acciones

Los datos de precios sobre las acciones del índice COLCAP provienen directamente desde una herramienta web de la Bolsa de Valores de Colombia® diseñada para este fin, por lo cual la información ya se encuentra estructurada y depurada.

Dado que la magnitud de los datos de precios varía de manera importante dependiendo de la especie, se realizó una normalización de estos datos para el entrenamiento del modelo de recomendación.

10.2.2 Noticias sobre emisores

En el proceso de etiquetado de datos se revisó cada una de las noticias, y se encontraron algunas que tenían errores en el proceso de extracción, sin embargo, estos errores fueron ajustados uno a uno en el código, de forma que al final, el desarrollo logró descargar de forma limpia y consistente las noticias.

10.2.3 Estados financieros

En los estados financieros descargados, se encontró el problema de que los formatos no venían completamente estandarizados, primero, porque algunos archivos tenían una o dos cuentas más que los demás, lo que afectaba el procedimiento de cargue masivo a la base de datos; y segundo, porque los estados de resultados consolidados vienen con cuentas diferentes a los de empresas individuales, por lo que se tuvo que almacenar manualmente 4 de los estados de resultados tomando la información del Registro Nacional de Valores y Emisores en la página *web* de la Superintendencia Financiera de Colombia.

10.3 Estructura del *dataset* de entrenamiento

Para entrenar el modelo de recomendación de acciones del índice COLCAP se creó primero una base de fechas tomando todos los días hábiles desde el año 2016 para cada una de las 24 acciones del índice; luego, sobre esta base, se añadieron como columnas los precios de los activos, las noticias y variables calculadas como parte del análisis técnico. Como variable respuesta se utilizó el precio de la acción a 1, 10 y 20 días.

10.3.1 Datos de precios y noticias

Como datos de precios se utilizó el precio de cierre de mercado de cada acción por cada fecha de la base, mientras que para los datos de noticias se realizó un conteo de cuántas noticias positivas, negativas o neutras ocurrieron por cada fecha y por cada acción.

10.3.2 Variables calculadas

Como variables complementarias para el análisis técnico se utilizaron los siguientes indicadores, definidos y recomendados por la Bolsa de Valores de Colombia (2021):

- Medias móviles simples entre 5 y 20 días (corto plazo): Promedio de precios de cierre en un periodo determinado para cada día de cotización.
- % Williams: Mide qué tan cercanos han estado los precios de cierra al máximo o mínimo de un periodo determinado. Por lo general se elige un periodo de tiempo N de 14 días.

$$\%R = \frac{Máximo\ N\ días - Cierre\ Hoy}{Máximo\ N\ días - Mínimo\ N\ días} * -100$$

- *Moving average convergence divergence* (MACD): Uno de los indicadores más utilizados por los analistas, sirve para medir la convergencia o divergencia de los promedios móviles respecto del precio de los activos. Se compone de 3 elementos.
    1. MACD = PME(12) – PME(26), donde PME: Promedio Móvil Exponencial
    2. Señal = PME(9,MACD)
    3. Histograma = Señal – MACD

Como parte del análisis fundamental, se utilizaron los estados financieros y el valor en libros por acción, para calcular los indicadores margen EBITDA y precio / valor en libros, mencionados en el análisis exploratorio de datos.

$$Margen\ EBITDA = \frac{EBITDA}{Ingresos\ totales}$$

$$Precio\ /\ Valor\ en\ libros\ = \frac{Precio\ de\ la\ acción}{Valor\ en\ libros\ por\ acción}$$

$$donde: Valor\ en\ libros\ por\ acción = \frac{Activo - Pasivo}{Acciones\ en\ circulación}$$

De esta forma, se presenta en la **Figura 13** la estructura de donde parte el *dataset* final de entrenamiento.

Columna	Descripción
date	Fecha de referencia
equity_id	Identificador de la acción
close	Precio de cierre
MM5	Media móvil 5 días
MM20	Media móvil 20 días
W	% Williams
MACD	MACD
signal	Señal MACD
histogram	Histograma MACD
positive	# Noticias positivas
neutral	# Noticias neutrales
negative	# Noticias negativas
ebitda_margin	Margen EBITDA
price/book_value	Precio / valor en libros
price	Variable respuesta, precio

Tabla 13. Estructura de datos para entrenamiento del modelo, elaboración propia.

11. Modelamiento

11.1 Arquitectura modelo de recomendación para inversión en acciones colombianas pertenecientes al índice COLCAP

Para el modelo de recomendación final se requieren 3 categorías de insumos: una serie de precios, que ingresa al modelo y sirve para crear variables calculadas para representar el análisis técnico, una serie con la cantidad de noticias positivas, negativas y neutras por cada fecha y cada acción, y una serie de medidas relacionadas con el emisor, calculadas con los estados financieros; estas dos últimas representan el análisis fundamental. Todos estos insumos se concatenan al final, formando una serie que funciona como insumo al modelo final para predecir el precio de la acción en un horizonte de tiempo.

En cada una de las iteraciones y para las predicciones, el modelo recibe una serie de datos, variables calculadas y noticias; y como salida, el modelo devuelve una rentabilidad que será calculada para 1, 10 y 20 días hábiles; de forma que la herramienta de recomendación pueda, no sólo recomendar una acción, sino además hacerlo por horizontes temporales hasta 1 mes, teniendo en cuenta que a mayor plazo existe una mayor probabilidad de error de pronóstico.

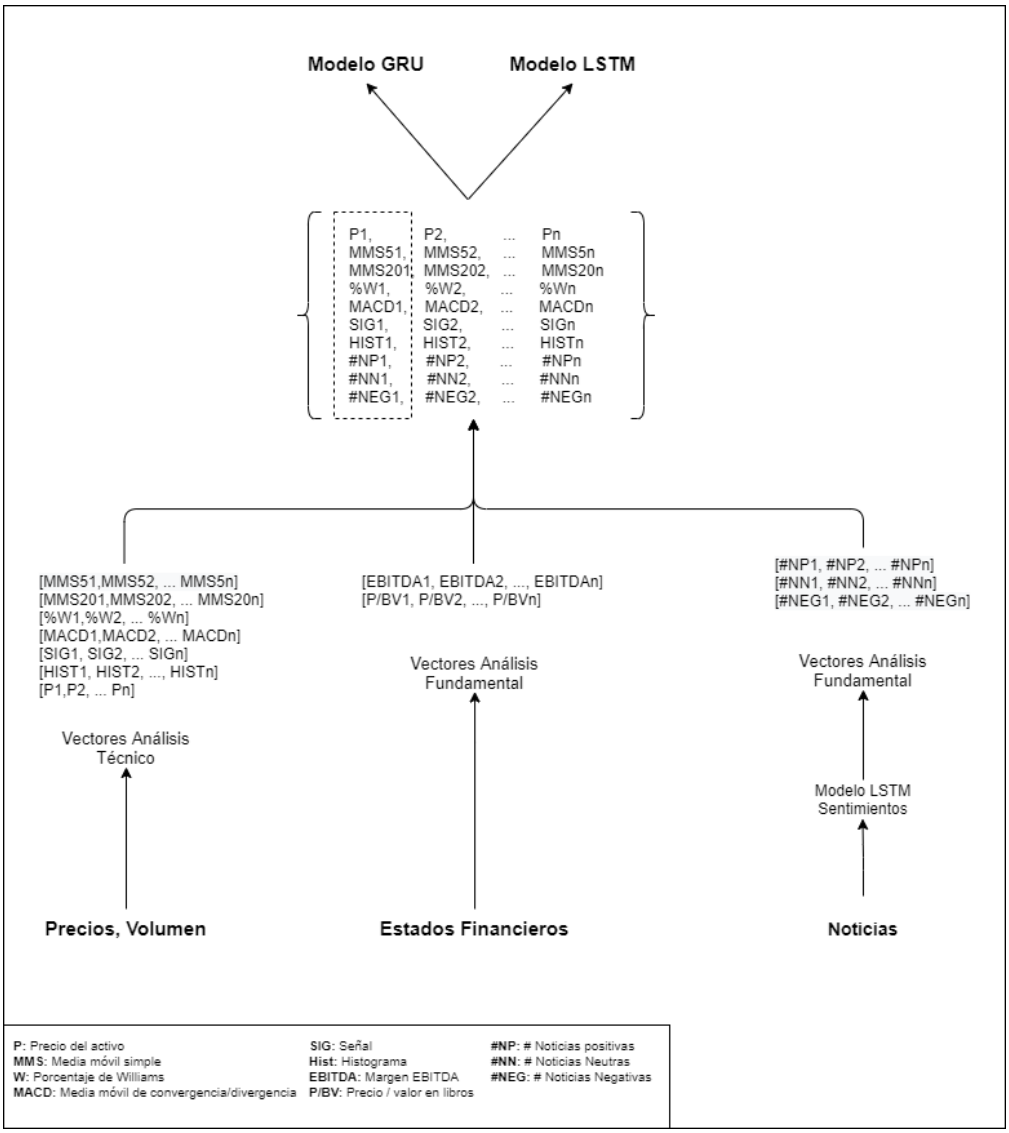

Figura 14. Estructura de datos para entrenamiento del modelo, elaboración propia.

Para el modelo LSTM, se diseñó una estructura secuencial de capas que se enumera a continuación.

1. Capas *Long Short-Term Memory* (LSTM) con 32 neuronas y función de activación relu.
2. Capas *Long Short-Term Memory* (LSTM) con 8 neuronas y función de activación relu.
3. Capa densa completamente conectada con una salida de 1 elemento.

El modelo GRU, fue diseñado de forma similar al LSTM.

1. Capas *Gated Recurrent Unit* (GRU) con 32 neuronas y función de activación relu.
2. Capas *Gated Recurrent Unit* (GRU) con 8 neuronas y función de activación relu.
3. Capa densa completamente conectada con una salida de 1 elemento.

Para el entrenamiento se empleó como función de pérdida el error cuadrático medio (MSE) y para la optimización el algoritmo *adam*.

Antes de ingresar al *dataset* final, la serie de noticias debe pasar por un modelo que sea capaz de clasificar cada una en positiva, negativa o neutra. La arquitectura de dicho modelo es mostrada en el siguiente numeral.

11.1.1 Modelo de Sentimientos para noticias sobre emisores de acciones del índice COLCAP

Como modelo de sentimientos se plantea un modelo con capas LSTM con activación *Softmax* haciendo uso de un vector con un millón de *word embeddings* preentrenado con un corpus en español utilizando la metodología word2vec, y publicado por Tatman (2017) en el sitio oficial de kaggle®.

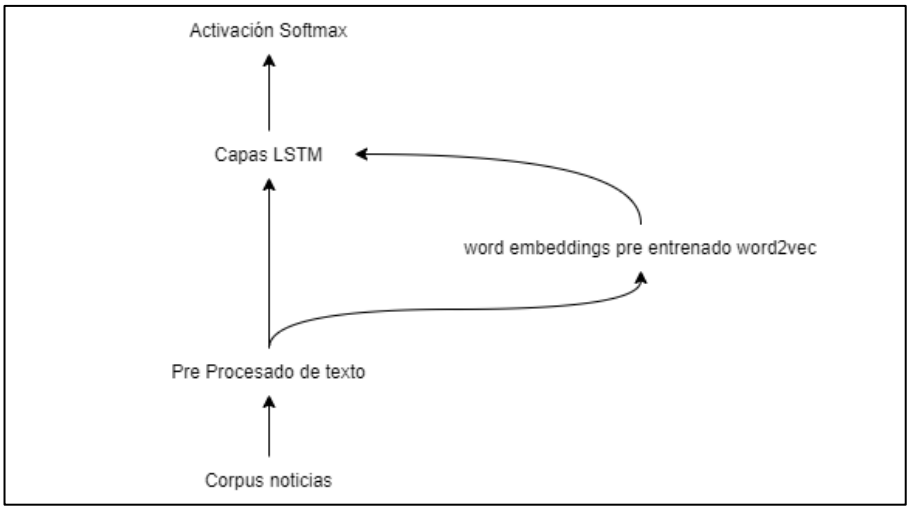

Figura 15. Modelo de sentimientos para noticias relacionadas con emisores del índice COLCAP, elaboración propia.

Luego de probar diferentes combinaciones de hiper parámetros, se llega a una arquitectura de modelo secuencial que cuenta con:

1. Una capa de *Word Embeddings* pre entrenada, donde cada palabra es representada por un vector de 300 elementos.
2. Una serie de capas *Long Short-Term Memory* (LSTM) de 190 neuronas con un *dropout* del 20%.
3. Una capa densa totalmente conectada, con una salida de 3 elementos (categorías) y una función de activación *softmax.*

En adición, para el entrenamiento del modelo se utilizó la función de pérdida *categorical crossentropy,* el algoritmo optimizador *Adam* y como métrica para validar los resultados se utilizó *accuracy.*

11.2 Resultados obtenidos

11.2.1 Modelo de Sentimientos para noticias sobre emisores de acciones del índice LSTM

Como métrica para la validación del modelo encargado de clasificación de noticias se utilizó *Categorical Accuracy,* una medida diseñada para problemas de clasificación donde existen más de dos categorías. Esta métrica expresa de forma general qué proporción de las predicciones del modelo son correctas (Grandini, Bagli, & Visani, 2020).

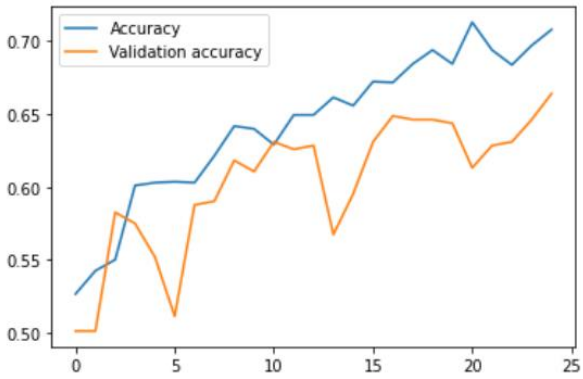

Figura 16. Evolución de la métrica Accuracy por cada época en el entrenamiento del modelo, elaboración propia.

En la **Tabla 14** se detallan los resultados del modelo para los datos de entrenamiento y validación, donde finalmente se obtiene un *accuracy* del 66.41%. Esta medida, puede contrastarse con el trabajo realizado por Bouazizi & Ohtsuki (2019), donde se muestra que existe una reducción muy relevante en medidas de validación para aquellos modelos que categorizan los datos en más de 2 categorias; llegando a encontrar, incluso, disminuciones de alrededor de 30% en *accuracy* al pasar de 2 a 3 categorías. Finalmente, en su trabajo, ejemplifican dicho efecto realizando un análisis de sentimientos, donde al agregar una tercera categoría, los modelos, en promedio, obtienen un *accuracy* de 72.5%.

	Entrenamiento	**Validación**
Accuracy	70.80%	66.41%

Tabla 14. Resultado final de la métrica Accuracy para el set de entrenamiento y el set de validación, elaboración propia.

Un mejor resultado en medidas de validación podría obtenerse encontrando métodos que permitan reducir el nivel de ambigüedad en las noticias, y aumentando aún más el número de datos para el entrenamiento, posiblemente incluyendo noticias en inglés traducidas de medios internacionales como Bloomberg®.

11.2.2 Modelo de recomendación para inversión en acciones colombianas pertenecientes al índice COLCAP

Para el modelo de recomendación final se postulan dos modelos de redes neuronales recurrentes, un modelo *gated recurrent unit* (GRE) y un modelo *long short-term memory* (LSTM). Ambos modelos son alimentados por todas las variables del dataset final, y como salida de cada uno se plantea una regresión que realice un pronóstico para el precio de cada acción del índice COLCAP para 1, 10 y 20 días hábiles. Una vez obtenidos los pronósticos del precio, la herramienta calcula la rentabilidad respecto del precio actual para cada acción; y finalmente, realiza una jerarquización de mayor a menor rentabilidad, mostrándole al usuario cuáles acciones podrían ser más atractivas según las variables tenidas en cuenta y

según el plazo deseado, teniendo en cuenta que a mayor plazo se tendrá una probabilidad mayor de error por parte del modelo.

Tras ejecutar ambos modelos para las acciones pertenecientes al índice COLCAP, se obtuvieron los resultados mostrados en la **Tabla 15** para el conjunto de validación, donde, para hacer más sencillo el análisis, se calculó un promedio de cada medida realizada a los pronósticos de las 24 acciones. En esta sección se resalta que ambos modelos poseen una calidad de pronóstico similar, lo cual podría atribuirse a que ambos son modelos recurrentes, que utilizan la memoria como su característica destacada. Para el caso de los pronósticos a 1 día, el modelo LSTM hace un trabajo marginalmente mejor que el GRU, sin embargo, este último tiene un error considerablemente menor en los pronósticos de 10 y 20 días para todas las medidas, por lo que se prefiere este modelo sobre el LSTM.

Modelo	Promedio MAE	Promedio MAPE	Promedio RMSE
GRU 1 día	402	3.38	477
LSTM 1 día	397	3.12	473
GRU 10 días	721	4.40	865
LSTM 10 días	836	5.23	978
GRU 20 días	1149	6.51	1332
LSTM 20 días	1209	6.77	1370

Tabla 15. Resultado final de las métricas de pronóstico para la validación cruzada, elaboración propia.

Una particularidad esperada en los datos resultantes es que a medida que se incrementa el horizonte del pronóstico, así también incrementa el error promedio al pronosticar, sin embargo, se destaca que, aunque efectivamente incrementan estas medidas, todavía se podrían encontrar en un rango aceptable para determinar cuáles acciones serían las más atractivas.

Al analizar los resultados detallados, se encontró que la acción que en promedio tuvo las mejores medidas de calidad de pronóstico fue CORFICOLCF, cuyos pronósticos se muestran en la **Figura 17**. Las gráficas de esta figura Uno de los motivos que podría explicar este hallazgo es que, como se mostró en el análisis exploratorio de datos, esta acción es la que menor volumen promedio diario de operación posee en la Bolsa de Valores de Colombia, lo cual lleva a esta acción a tener una volatilidad baja en comparación a las demás, como se comprobó previamente en la **Tabla 8**,

donde esta acción está en la parte inferior de las volatilidades de los rendimientos logarítmicos de las acciones pertenecientes al índice COLCAP.

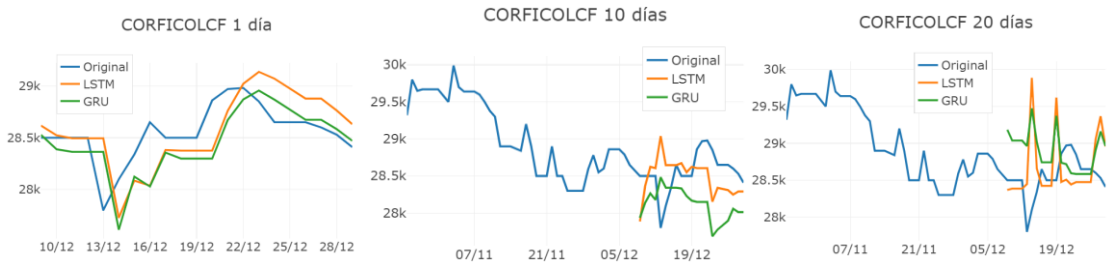

Figura 17. 20 pronósticos de CORFICOLCF utilizando datos de 1, 10 y 20 días atrás, elaboración propia.

Por otra parte, se encontró que GRUPOSURA fue la acción con peor calidad de pronóstico, lo cual podría ser explicado por un incremento inesperado cercano al 15%, generado por un apetito mayor hacia el emisor a causa de la oferta pública de adquisición sobre las acciones ordinarias de sura (Cajamarca, La República, 2022). Este incremento se puede observar en la **Figura 18**, y fue tan paulatino, que fue difícil para el modelo adaptarse rápidamente, incluso teniendo en cuenta que es precisamente este emisor aquel que tiene un mayor número de noticias recolectadas en la base de datos, como se mostró en el análisis exploratorio de datos.

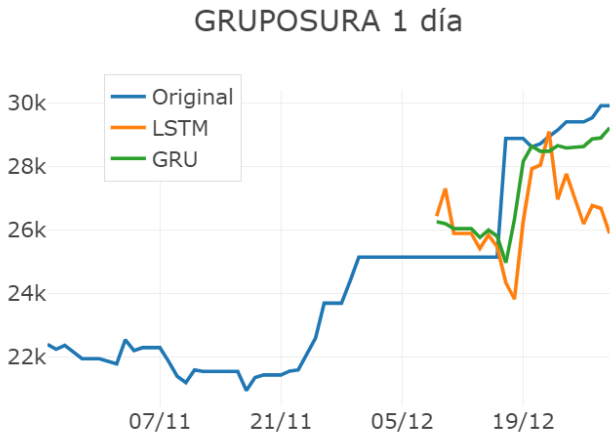

Figura 18. 20 pronósticos de GRUPOSURA utilizando datos de 1 día atrás, elaboración propia.

En una revisión de la literatura realizada por Rouf et al (2021) se contrastan metodologías y nuevos desarrollos utilizados para la predicción de acciones en mercados financieros utilizando *machine learning*. En esta revisión exponen otros trabajos similares utilizando redes neuronales recurrentes LSTM; los cuales obtienen medidas de pronóstico con una precisión entre el 56% y el 72%, lo cual, a pesar de ser difícil de comparar, podría considerarse consistente con los resultados aquí obtenidos, y es una muestra de que todavía queda un largo camino en la tarea de predecir con exactitud movimientos en los activos de los mercados financieros.

12. Conclusiones y recomendaciones

12.1 Conclusiones

Objetivo 1: Caracterizar las fuentes de datos y las variables para realizar el análisis técnico y fundamental (de sentimientos) para una empresa emisora.

Para el desarrollo del modelo de recomendación, se utilizaron 3 tipos de datos relacionados con el emisor de cada acción: noticias, precios y estados financieros.

- Para el caso de las noticias, se desarrolló una herramienta de *web scraping* capaz de buscar las noticias de cada emisor, guardar las url de dichas noticias y luego extraer el corpus, título y fecha de cada una. Como fuente de datos para la extracción de noticias se utilizaron los diarios Dinero® y La República®.
- Los precios de valoración y volumen de negociación de cada acción fueron obtenidos también a través de *web scraping,* al desarrollar una herramienta que va directamente a la *web* de la Bolsa de Valores de Colombia® y extrae esta información por cada acción y para cada fecha definida.
- Los estados financieros de los emisores de acciones del índice COLCAP son publicados por la Superintendencia Financiera de Colombia®, sin embargo, esta información no es estandarizada y es publicada en formato pdf. Como solución a este problema, se utilizó un servicio pago llamado EMIS®, que ofrece información estandarizada en formato de Microsoft Excel®; a través de este servicio se obtuvo la información histórica de estados financieros, y luego, se desarrolló una herramienta que lee los datos y los tabula en un formato adecuado para bases de datos.

Como un primer resultado de este objetivo, se obtiene una base de datos *SQLite,* donde se almacenó toda la información aquí mencionada, de forma organizada en 7 tablas bajo un esquema de entidad relación.

Una vez obtenidos los datos, se desarrolló una herramienta que los extrae de la base de datos y los consolida en un *dataset* final que luego será utilizado para el entrenamiento, este sería otro resultado de este objetivo.

Objetivo 2: Implementar un modelo de procesamiento del lenguaje natural para el análisis de sentimientos sobre la empresa emisora.

Un insumo importante del modelo de recomendación son las noticias relacionadas con emisores de acciones pertenecientes al índice COLCAP, sin embargo, estas noticias primero deben ser clasificadas en positivas, negativas o neutras; para ello, se desarrolló un modelo de procesamiento del lenguaje natural utilizando redes neuronales recurrentes del tipo *long short-term memory,* además de una capa pre entrenada de 1 millón de palabras de *word embeddings* utilizando la metodología word2vec. Este modelo obtuvo una precisión final de 66.41% para el *dataset* de validación, lo cual es acorde a otros trabajos encontrados en la literatura, que muestran que los modelos de clasificación múltiple tienden a tener una precisión más baja incluso desde 3 categorías.

Objetivo 3: Implementar un modelo de clasificación de acciones combinando análisis técnico y fundamental sobre la empresa emisora.

Para la implementación del modelo de clasificación de acciones se prepararon dos modelos de redes neuronales recurrentes, un modelo *gated recurrent unit* (GRU) y uno *long short-term memory* (LSTM). En el primer modelo, se utilizaron 3 capas GRU con 64, 32 y 8 neuronas respectivamente, donde para esta última capa se utilizó una activación de tipo *relu*; y finalmente una capa densa de una neurona. Para el modelo LSTM se utilizaron 3 capas LSTM de 64, 32 y 8 neuronas, cada una con función de activación *relu,* y luego una capa densa de salida de una neurona. Para ambos modelos se utilizaron 10 épocas por cada una de las 24 acciones, una tasa de aprendizaje de 0.01, como método de optimización la función *Adam* y como función de pérdida el *mse.*

Las variables de entrada para ambos modelos son las mismas, y fueron seleccionadas siguiendo las dos grandes categorías de análisis mencionadas literatura acerca de inversión en acciones; el análisis técnico, que trata de predecir el precio futuro basado en su comportamiento pasado, y el análisis fundamental, que trata de valorar de forma objetiva la empresa emisora utilizando toda su información disponible (Castro Alfaro & Anturi Santos, 2015).

- Variables para el análisis técnico: Se utilizaron los precios de cada una de las acciones del índice COLCAP, y con ellos, se calcularon otras variables que también sirven de insumo:

  I. Medias móviles simples entre 5 y 20
  II. Porcentaje de Williams
  III. *Moving average convergence divergence*, señal e histograma

- Variables para el análisis fundamental: Las variables utilizadas para este tipo de análisis son las relacionadas con cada empresa emisora de acciones del índice COLCAP. Las primeras variables son las noticias positivas, negativas y neutras de cada emisor; por otro lado, utilizando los estados financieros de cada emisor se utilizaron las variables:

  I. Margen EBITDA
  II. Precio / valor en libros

Como resultado del objetivo, se obtienen dos modelos, ambos entrenados para cada una de las 24 acciones pertenecientes al índice COLCAP y utilizando la librería *Keras.*

Objetivo 4: Validar un modelo de recomendación de acciones colombianas pertenecientes al índice COLCAP basado en análisis técnico y el sentimiento del mercado local.

Para validar los modelos propuestos se utilizó la metodología de validación cruzada, que es aquella donde se parten los datos en dos, un subconjunto de datos para entrenar los modelos y otro subconjunto que sirve para probar los modelos entrenados con el primer subconjunto (Yang, 2007). Para el caso de este trabajo, se retiraron los últimos 20 datos de cada serie de precios

para cada acción, de forma que después los modelos entrenados intentaran predecir estos 20 datos que no fueron incluidos en el entrenamiento. Utilizando estos pronósticos para cada una de las acciones del índice COLCAP, se hallaron las medidas de calidad de pronóstico RMSE, MAE y MAPE; que sirvieron para realizar una comparación de modelos.

Después de validar las medidas de calidad de pronóstico, se encontró que el modelo *long short-term memory* tuvo unos resultados ligeramente superiores en las predicciones a 1 día, sin embargo, el modelo *gated recurrent unit* tuvo mejores medidas de pronóstico en los demás horizontes de tiempo, por lo que se elige este último como mejor modelo. También se encontró que la acción con mejor pronóstico fue CORFICOLCF, posiblemente explicado por ser la de menor volumen de negociación y de una volatilidad relativamente baja frente a las demás; y que la acción que peor fue su pronóstico fue GRUPOSURA, que tuvo un incremento repentino en su precio explicado por una oferta pública de adquisición, que generó ruido en el modelo para las predicciones, especialmente después del día del incremento.

Finalmente, se mencionan otros trabajos similares encontrados en la literatura sobre predicción de precios de acciones, donde se muestra que los resultados obtenidos en este trabajo son consistentes con otros publicados, y que además todavía hay mucho campo para la investigación en este tipo de pronósticos en particular.

12.2 Recomendaciones

Como una forma de aportar a futuros trabajos relacionados, se enumeran a continuación las recomendaciones encontradas.

- Modelo LSTM para noticias relacionadas con emisores del COLCAP: al analizar los datos extraídos de noticias, se encontró que existe un problema de desbalanceo en los datos en dos sentidos
    I. Existen empresas que tienen una mayor visibilidad ante la prensa, por lo que, en el mismo periodo de tiempo analizado, algunas empresas emisoras pueden tener hasta 9 veces más noticias que otras. Dentro de las posibles causas de este desbalanceo, se encuentran, entre otras:

- Empresas que poseen negocios en el exterior, pueden tener un alto número de noticias relacionadas con los mismos.
- Empresas que son muy grandes tienden a mover montos muy altos de dinero, realizar muchas inversiones y ganar reconocimientos, lo que luego se convierte en noticia.
- Existen empresas que son más visibles ante el público general, bien sea porque producen algún bien o servicio que va directo hacia las personas naturales, o porque tienen estrategias de mercadeo exitosas. Todo esto genera que sea más atractivo para la prensa publicar sobre estas empresas por encima de las menos reconocidas.

Como mejora a este desbalanceo, se propone incluir insumos que complementen el modelo de análisis de sentimientos, como opiniones de analistas en redes sociales o medios en inglés como Bloomberg®.

II. Existe un número reducido de noticias negativas: en este trabajo, se encontró que, dentro de las noticias de cada emisor, hay una proporción mucho mayor de noticias positivas y neutras. Este desbalanceo puede variar según el horizonte temporal, y sería tema de estudio analizar si los medios donde se publican estas noticias tienen un sesgo hacia publicar noticias positivas por encima de las negativas. Adicionalmente, se podría trabajar en cómo mejorar estos modelos con métodos que reduzcan el sesgo en clases desbalanceadas.

Adicionalmente, se encuentra que existe ambigüedad al etiquetar las noticias, pues, en algunos casos, es difícil definir con certeza si una noticia será positiva, negativa o neutra para el mercado de acciones. Algún negociador del mercado podría interpretar que una noticia tendrá un efecto positivo en el precio de la acción, mientras que, para otro negociador, es posible que la misma noticia, aunque la considere positiva, no tenga el impacto suficiente para afectar el precio de la acción y termine siendo neutra. Se recomienda como mejora para este tipo de modelos, que se investigue en métodos para reducir la ambigüedad en el etiquetado de datos.

- Modelo de recomendación de acciones pertenecientes al índice COLCAP: Como hallazgo, se encontró que, el modelo puede tener dificultades para predecir los precios

justo después de una oferta pública de adquisición, ya que esta puede generar un movimiento brusco muy repentino en el precio. Se recomienda estudiar el efecto de las ofertas públicas de adquisición en el precio y volatilidad de las acciones, y qué métodos podrían utilizarse para generar predicciones más acertadas bajo este escenario.

Bibliografía

Adhikari, U. R. (2020). The Market Risk Framework. *ResearchGate*.

Asociación Bancaria y de Entidades Financieras de Colombia. (2019). Caracterización del mercado de capitales colombiano: una perspectiva integral.

Asociación Bancaria y de Entidades Financieras de Colombia. (2019). *Segunda Misión del Mercado de Capitales: reflexiones desde el sector financiero.* Bogotá: Semana Económica.

Autorregulador del mercado de valores de Colombia. (2019). Guía de estudio: Análisis Económico y Financiero. 144.

Banco de la República de Colombia. (05 de 09 de 2021). *Mercado accionario*. Obtenido de https://www.banrep.gov.co/es/estadisticas/mercado-accionario#:~:text=El%20COLCAP%20es%20un%20%C3%ADndice,ajustada%20(flotante%20de%20la%20compa%C3%B1%C3%ADa

Bolsa de Valores de Colombia. (2021). *Análisis técnico BVC*. Obtenido de https://www.bvc.com.co/pps/tibco/portalbvc/Home/Mercados/Analisis_Tecnico?action=dummy

Bolsa de Valores de Colombia. (16 de Enero de 2021). *Bolsa de Valores de Colombia*. Obtenido de https://www.bvc.com.co/pps/tibco/portalbvc/Home/Glosario

Bouazizi, M., & Ohtsuki, T. (2019). Multi-Class Sentiment Analysis on Twitter: Classification Performance. *BIG DATA MINING AND ANALYTICS*, 14.

Cajamarca, I. (25 de 10 de 2021). *La República*. Obtenido de https://www.larepublica.co/finanzas/conozca-tres-opciones-para-comprar-acciones-en-la-bolsa-de-valores-colombiana-3251818

Cajamarca, I. (26 de 01 de 2022). *La República.* Obtenido de https://www.larepublica.co/especiales/opa-por-nutresa/durante-el-periodo-de-aceptacion-de-opa-el-grupo-sura-transo-227679-millones-3291967

Castro Alfaro, A., & Anturi Santos, R. (2015). EL ANÁLISIS TECNICO Y FUNDAMENTAL EN UN CONTEXTO DE GLOBALIZACIÓN: BANCOLOMBIA. *AGLALA*.

Chami, R., Fullenkamp, C., & Sharma, S. (2009). A Framework for Financial Market Development. *International Monetary Fund* , 60.

Córdoba Garcés, J., & Molina Ungar, E. (2017). Elementos para alcanzar el mercado de capitales que Colombia necesita: hoja de ruta y desafios estructurales. *Centro de Estudios Sobre Desarrollo Económico, Universidad de los Andes*, 37.

Drakopoulou, V. (2015). A Review of Fundamental and Technical Stock Analysis Techniques. *Journal of Stock & Forex Trading*.

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*.

Grandini, M., Bagli, E., & Visani, G. (2020). METRICS FOR MULTI-CLASS CLASSIFICATION: AN OVERVIEW.

Guo, H. (2002). Understanding the Risk-Return Tradeoff in the Stock Market. *FEDERAL RESERVE BANK OF ST. LOUIS*.

Johnson, C. A. (2001). Value at Risk: Teoría y aplicaciones. *Estudios de Economía vol. 28*.

Khairi, T., Zaki, R., & Mahmood, W. (2019). Stock Price Prediction using Technical, Fundamental and News based Approach. *2nd Scientific Conference of Computer Sciences (SCCS), University of Technology - Iraq*.

Lagos Cortés, D. (2013). Análisis de las prácticas de Gobierno Corporativo en la Bolsa de Valores de Colombia. *AD-minister*, 19.

Lasek, M., & Lasek, J. (2015). Are Stock Markets Driven More by Sentiments than Efficiemcy? *Journal of Engineering, Project, and Production Management*.

Levine, R. (1996). Financial Development and Economic Growth. *Policy Research Department, World Bank*, 84.

Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. *Information Processing & Management*, 19.

Linsmeier, T. J., & Pearson, N. D. (2000). Value at Risk. *Association for Investment Management and Research*.

LUKIC, R. (2018). The Analysis of the Operative Profit Margin of Trade Companies in Serbia. *Review of International Comparative Management*.

Malkiel, B. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives*, 60.

Martín, M. Á. (2011). *Mercado de capitales: Una perspectiva global.* Buenos Aires: CENGAGE Learning.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications:A survey. *Ain Shams Engineering Journal*, 21.

Ministerio de Hacienda, Banco Mundial & Fedesarrollo. (1996). *Misión de Estudios del Mercado de Capitales.* Bogotá: Fundación para la educación superior y el desarrollo.

Ministerio de Hacienda, Banco Mundial & Programa de Cooperación Económica y Desarrollo. (2019). *Misión del mercado de capitales.*

Miskolczi, P. (2017). NOTE ON SIMPLE AND LOGARITHMIC RETURN. *APSTRACT*, 127-136.

Mittal, A., & Goel, A. (2011). Stock Prediction Using Twitter Sentiment Analysis. *Standford University*.

Moreno, A., & Humberto, J. (23 de Noviembre de 2019). Unos 65.600 accionistas salen de la Bolsa de Valores de Colombia cada año. *La República*, pág. 1.

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*.

Pernagallo, G., & Torrisi, B. (2020). Blindfolded monkeys or financial analysts: who is worth your money? New evidence on informational inefficiencies in the US stock market. *Physica A: Statistical Mechanics and its Applications*.

Portafolio. (13 de Octubre de 2020). Bolsa de Colombia: cada vez menos emisores de acciones. *Portafolio*.

Portafolio. (2021). Valor en libros de las acciones muestra que están baratas. *Portafolio*.

Qaiser, H., & Shafique, U. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*.

Quiñonez Avendaño, C. (2010). INTERVENCIÓN DEL ESTADO COLOMBIANO EN EL MERCADO DE VALORES. *Iustitia*.

Rojas, C. I., & Gonzalez, A. (2008). Mercado de Capitales en Colombia: Diagnóstico y Perspectivas de su. *Asociación Nacional de Instituciones Financieras*.

Rouf, N., Bashir Malik, M., Arif, T., Sharma, S., Singh, S., Aich, S., & Kim, H.-C. (2021). Stock Market Prediction Using Machine Learning Techniques: A Decade Survey on Methodologies, Recent Developments, and Future Directions. *Electronics*.

Salazar Sierra, C. (17 de 12 de 2021). *Editorial La República S.A.S.* Obtenido de https://www.larepublica.co/economia/colombia-seria-lider-en-la-region-en-crecimiento-cerrando-2021-con-un-pib-de-97-3278348

Sasank Pagolu, V., Reddy Challa, K. N., Panda, G., & Majhi, B. (2016). Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. *International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016*, 6.

Superintendencia Financiera de Colombia. (2019). *Actualidad del Sistema Financiero Colombiano.* Dirección de Investigación y Desarrollo.

Tatman, R. (2017). *Pre-trained Word Vectors for Spanish*. Obtenido de https://www.kaggle.com/rtatman/pretrained-word-vectors-for-spanish

TensorFlow. (22 de 01 de 2021). *TensorFlow*. Obtenido de https://www.tensorflow.org/api_docs/python/tf/keras

Tian, Y. (2009). Market Liquidity Risk and Market Risk Measurement. *Delft University of Technology*.

Wang, H., Wang, T., & Li, Y. (2020). Incorporating Expert-Based Investment Opinion Signals in Stock Prediction: A Deep Learning Framework. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.* .

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. *Artif Intell Rev 53*.

Yang, Y. (2007). CONSISTENCY OF CROSS VALIDATION FOR COMPARING REGRESSION PROCEDURES. *The Annals of Statistics*.

Zhang, Y., & Wallace, B. (2015). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *arXiv preprint arXiv:1510.03820*.